Towards Effective and Scalable Vision-Language Models for Knowledge-intensive Visual Tasks

A PhD Thesis submitted by

Abhirama Subramanyam Penamakuri

in partial fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

Indian Institute of Technology Jodhpur
Computer Science and Engineering
May, 2026

Declaration

I hereby declare that the work presented in this thesis titled "*Towards Effective and Scalable Vision-Language Models for Knowledge-intensive Visual Tasks*", submitted to the Indian Institute of Technology Jodhpur in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out under the supervision of Dr. Anand Mishra (Indian Institute of Technology Jodhpur). The contents of this thesis, in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Abhirama Subramanyam Penamakuri
P19CSE001
Computer Science and Engineering
Indian Institute of Technology Jodhpur

Certificate

This is to certify that the thesis titled "*Towards Effective and Scalable Vision-Language Models for Knowledge-intensive Visual Tasks*", submitted by Abhirama Subramanyam Penamakuri (P19CSE001) to the Indian Institute of Technology Jodhpur for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work done by him under my supervision. To the best of our knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Anand Mishra
Ph.D. Thesis Supervisor
Computer Science and Engineering
Indian Institute of Technology Jodhpur

Abstract

The ability to interpret images in conjunction with language is central to how humans understand the world. Vision-Language Models (VLMs) aim to replicate this capability, and recent advances have enabled impressive performance on many vision and language tasks, including text-to-image retrieval, visual question answering, and image captioning. However, real-world scenarios often demand a deeper level of understanding, where interpreting an image also requires access to background knowledge, associated long-tail facts, or domain-specific context. For instance, answering a question about a brand, a historical landmark, or a fashion product appearing in a natural scene often goes beyond what is directly visible, requiring models to connect visual content with external knowledge sources. Such knowledge-intensive visual tasks remain a significant challenge, highlighting the limitations of current VLMs, which primarily rely on parametric knowledge acquired during pretraining. This thesis addresses this gap through two core research directions: enhancing the effectiveness of vision-language models through retrieval-based grounding and improving efficiency through scalable learning techniques.

On the effectiveness front, we develop retrieval-augmented frameworks that enable VLMs to incorporate factual, structured, or cross-modal knowledge at inference time. We begin with Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT), a model for knowledge-aware image retrieval that links visual entities to Wikipedia using a basic similarity-based visual entity linker. We then extend this vanilla visual entity linker into VisTEL, a VLM-based module that uses visual cues and surrounding text to disambiguate scene-text entities. We demonstrate the utility of VisTEL in improving knowledge-aware text-based visual question answering when integrated within Knowledge-aware Large Multimodal Assistant (KaLMA). While these methods are effective for single-image understanding, they are limited when a question requires knowledge mining across multiple images. To address this limitation, we introduce Retrieval-augmented Visual Question Answering (RetVQA), a benchmark and method that retrieve relevant visual evidence across multiple images to answer questions and thereby support knowledge mining. Finally, for image captioning in specialized domains such as fashion, we propose Retrieval-augmented Chain of Attributes (RA-CoA), a training-free framework that grounds descriptions in retrieved attribute–value pairs from exemplar images, ensuring detail and consistency. We perform extensive evaluations on public benchmarks and show consistent superiority over competitive baselines and state-of-the-art approaches.

On the efficiency front, we focus on improving the performance of smaller Vision-Language Models without incurring the heavy computational and data costs of large-scale training. To achieve this, we introduce the Model Parity Aligner (MPA), a simple yet powerful training framework designed to transfer knowledge from large, high-performing VLMs to their smaller counterparts. Rather than relying on expensive human annotations, MPA identifies knowledge gaps between model pairs and generates targeted, high-quality pseudo-labels that specifically address these gaps. This selective guidance enables small models to better mimic the reasoning patterns and factual grounding of larger models. In doing so, MPA not only improves the accuracy and robustness of compact VLMs but also makes them better suited for deployment in resource-constrained and low-supervision environments.

In summary, this thesis contributes both methodological and infrastructural advances to the development of knowledge-aware, scalable VLMs. Our proposed approaches significantly advance the state of the art. Together, these contributions aim to push the frontier of vision-language understanding toward systems that are not only more capable but also more practical to deploy in real-world, knowledge-intensive environments. All resources developed in this thesis are publicly open-sourced to support reproducibility and future research.

Acknowledgements

This thesis would not have been possible without the guidance, support, and encouragement of many people, to whom I am deeply grateful. Foremost, I would like to thank my PhD advisor, Dr. Anand Mishra, for his immense patience and guidance, especially during my initial days when I was still settling into research. His support and constant encouragement made me the researcher that I am today. His clarity of thought, scientific rigor, and faith in me have been invaluable throughout this journey. I am deeply thankful to him for his understanding and support during the personal challenges I faced in the course of my PhD, which meant a lot to me. I am also grateful to him for providing me the opportunities to attend national and international conferences and for his generous support in making them possible. Equally invaluable has been his sense of time and meticulous planning, qualities that I continue to aspire to, even as I still struggle to find the same balance in my own life.

I am also grateful to my Student Research Committee (SRC) members, Dr. Mayank Vatsa, Dr. Anirban Chakraborty, and Dr. Chiranjoy Chattopadhyay, for their valuable suggestions and constructive feedback at various stages of my PhD. I am thankful to Dr. Manish Gupta and Dr. Mithun Das Gupta for their mentorship during our work on RetVQA, which was an immensely valuable experience. I also thank Dr. Balaji Vasan Srinivasan, whom I first met during IJCAI and who later gave me the opportunity to intern at Adobe, an experience I truly cherish. I also thank the faculty members Dr. Lipika Dey, Dr. Richa Singh, Dr. Gaurav Harit, and Dr. Romi Banerjee, who have supported me in various ways, including through recommendations and academic advice, during this journey.

My next thanks go to my sponsors: Prime Minister's Research Fellowship (PMRF), Ministry of Education (MoE), and Bhashini (MeitY), for supporting my doctoral studies. I also thank MoE and Microsoft Research for providing travel grants, which allowed me to present and discuss my work with the research community. I am grateful for the dedicated assistance I received from the technical and administrative staff at IIT Jodhpur. I especially thank Mr. Rimpesh for his constant behind-the-scenes help in managing servers and promptly resolving technical issues, often even at odd hours. I thank Mr. Himmat and Mr. Mahindra for smoothly handling the paperwork and administrative processes. I also thank Mr. Tarun and Mr. Ravindra for their help with data annotation and for the many local food outings that made my stay in Jodhpur memorable.

I am grateful to have worked with several excellent collaborators, including research associates Revant, Prajwal, Anik, Lokesh, and Uday; undergraduate students Vaibhav, Mayank, Nakul, Shreya, Navlika, Piyush, and Akshat; intern Hitesh; and colleagues Kiran, Abhishek, and Somraj. Working with them was an enriching and joyful experience, and I thank them all for enduring my classic last-minute push mode!

This journey would not have been possible without the support of amazing friends. Vishnu, Yogesh, Rishabh, Rekhansh, and Narendra have been my constants, standing with me through my ups and downs and patiently listening to my PhD frustrations. Thanks, folks. I thank Yasmeena for all the time we spent together, especially our aligned interests in stress shopping. I thank Soumya, Kiran, Abhilash, Chetan, and Akash, with whom I shared many wonderful moments during my internship at Adobe Bengaluru. Special thanks to Atul, my gym guru and bro, who constantly motivated me to push harder. I also thank Sandeep, Shubham, Deepanshu, Lokesh, and Narendra for all the badminton games we played. I am thankful to Sudheer, Ankur, Himani, Kartik, Niveditta, Bharti, Kanika, Surbhi, Chiranjeev, Muskan, Shubham, Nandini, Himanshu and many other companions who, at different times in my PhD, supported me in meaningful ways. I also thank my junior Swapnil and the entire organizing team, for giving me the opportunity to perform the Ganesh Chaturthi pooja for four consecutive years, an unfor-

gettable experience of my PhD life. I also thank other members of the VL2G lab for the many informal discussions and the lively environment that made my PhD journey more enjoyable.

Above all, I owe everything to my family, who have been my strongest pillar of support throughout. My parents made countless sacrifices, supported my choices, and motivated me through their strength and resilience. I would like to take this moment to remember my late father Mr. PV Rangaraja, whose patience and calmness I still aspire to learn from. I know I may never reach his level, but his values continue to guide me, and I hope this achievement makes him proud. My mother, Adilakshmi Syamala, has been my greatest strength. Her constant prayers, unwavering faith, and blessings have always carried me forward in a positive way. My younger brother, Sarath Chandra, has been a steady source of support and motivation, and I deeply cherish his presence in my life. Finally, I bow with gratitude to Lord Venkateswara Swamy of Tirumala, our ancestral deity, and to Vārāhī Mātā, for Their blessings and guidance throughout this journey.

To my parents.

Contents

List of Figures

List of Tables

1
Introduction

"Vision modules interacting with other modules (like a knowledge graph in your case, or a memory module) will be one of the more important next steps for the field."

— Prof. Andrew Zisserman, in an email communication with my supervisor on KVQA [186], (an insight that became the starting point of this thesis).

A key emerging paradigm in artificial intelligence is to integrate perception systems with external modules, such as knowledge graphs, retrieval systems, or structured memory. Human behavior reflects this paradigm. With over 13.7 billion queries submitted daily, search engines have become our default mechanism for bridging knowledge gaps [43]. Rather than relying solely on memory, people increasingly depend on the ability to search, recognizing what they do not know, and knowing how to find it. For intelligent systems to reason similarly, they require to retrieve missing context before attempting to interpret or act.

Vision-Language Models (VLMs) [122, 133, 138, 271] have made significant strides in aligning visual and textual modalities over the past decade. Yet, most contemporary VLMs, particularly generative ones, are not inherently retrieval-aware. They rely primarily on parametric knowledge encoded during training and lack mechanisms for dynamic access to external context. As a result, they often struggle to generalize in knowledge-intensive visual tasks that require reasoning over unseen or context-specific information. This disconnect between what current VLMs[1] offer and what knowledge-intensive tasks demand reveals a critical gap: while many models excel at perception, they lack the ability to retrieve and integrate external knowledge as part of their core reasoning process. Beyond improving the effectiveness of VLMs, a growing concern is also the need to boost their efficiency. Scaling VLMs to billions of parameters has improved capabilities, however, at the cost of significant computational and memory demands, making deployment expensive and often impractical. Addressing this requires methods that retain the reasoning ability of the VLMs while reducing inference cost and resource footprint.

This thesis addresses these two challenges: effectiveness and efficiency of VLMs, by developing a suite of retrieval-augmented vision-language models that (i) explicitly integrate external

[1] Existing VLMs at the start of this thesis.

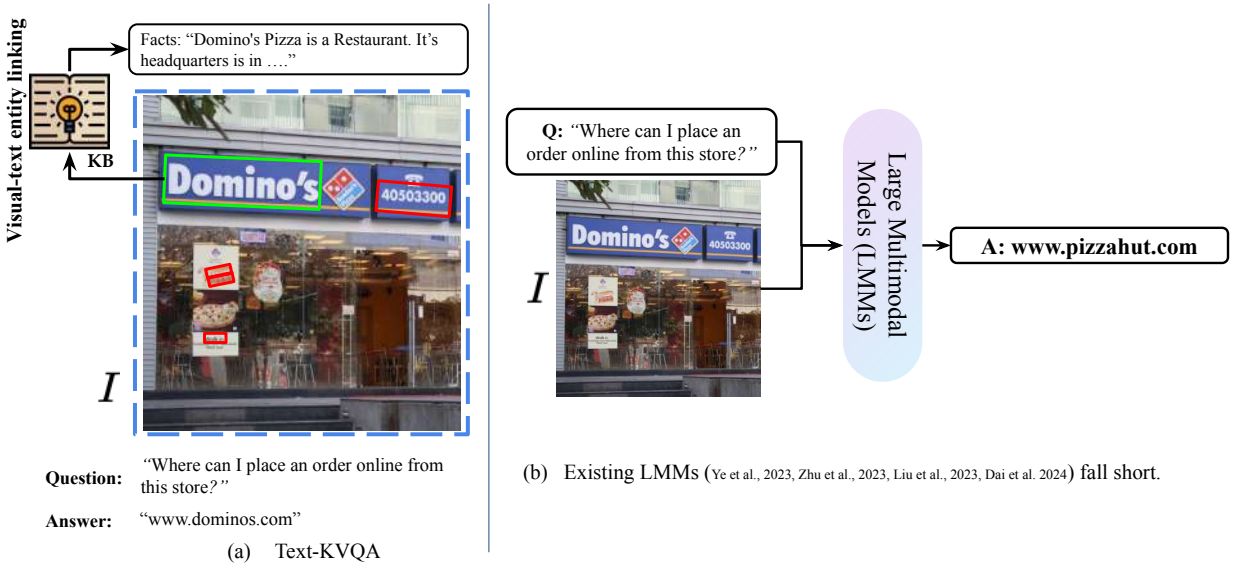


Figure 1.1: (a) In Text-KVQA [195], the goal is to answer questions grounded in named entities appearing as scene text, such as "Domino's." This requires linking the text to external knowledge. (b) Modern large vision-language models attempt such tasks without explicit grounding, often hallucinating answers based on superficial visual cues rather than the actual scene text.

knowledge for improved reasoning for knowledge-intensive visual tasks, and (ii) leverage small VLMs and knowledge distillation techniques to enable efficient deployment without sacrificing performance.

1.1 KNOWLEDGE-INTENSIVE VISUAL TASKS

Vision–language tasks empower artificial intelligence systems to jointly perceive, understand, and reason over visual and textual information. Core tasks such as image captioning [110, 129, 255], text-to-image retrieval [129, 255], and visual question answering (VQA)[7, 62] primarily align perception with language, relying on directly visible content. However, real-world applications demand more than object recognition; they require linking what is seen to background knowledge, associated facts, or broader context. This has given rise to a growing class of knowledge-intensive visual tasks, where the challenge shifts from "what do you see?" to "what do you know about what you see?" Examples include extending VQA to incorporate external knowledge[30, 146, 185, 186, 223], reasoning jointly across multiple images [15, 24, 164], and generating knowledge-grounded captions through retrieval [79, 164]. These tasks require models not only to recognize visual content but also to retrieve, integrate, and reason over external information. Much like humans connecting sparse visual cues to rich prior knowledge, intelligent systems must bridge perception with external knowledge to reason effectively. Building on this understanding, we next highlight four representative knowledge-intensive visual tasks that form the focus of this thesis:

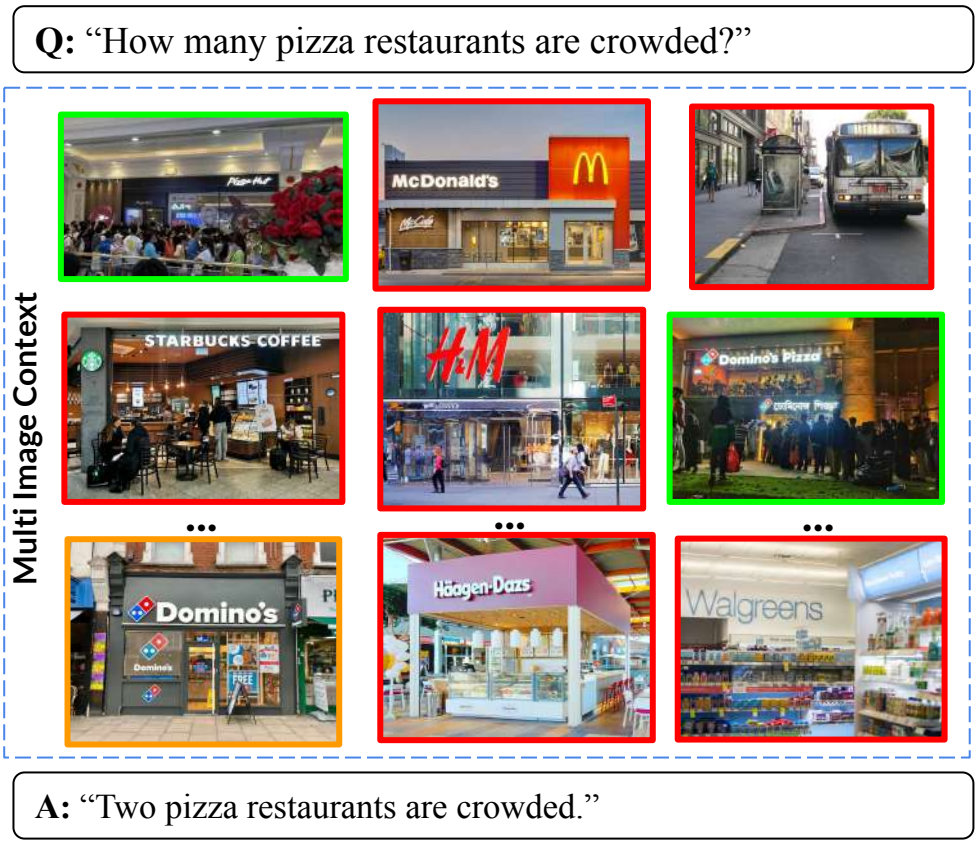


Figure 1.2: An illustrative example of a knowledge-intensive visual task: Consider the question "How many pizza restaurants are crowded?" over the pool of images, the system must first retrieve relevant images (images shown in green) based on both factual cues (e.g., identifying "Domino's" and "Pizza Hut" as pizza outlets) and commonsense cues (e.g., judging visual crowd density). The orange box highlights a partially relevant case: a pizza store that is not crowded. Solving this task requires knowledge-aware retrieval followed by multi-image reasoning and evidence aggregation.

1. ***Knowledge-aware VQA:*** This task [30, 146, 185, 186, 195, 222, 223] involves answering questions about an image that cannot be resolved from the visual content alone, but requires external factual information associated with the visual entities. The model has to first recognize key visual entities, then link them to relevant background knowledge, and finally reason over this knowledge to produce the answer. For example, to answer the question "Where can I place an order online from this store?" (Figure 1.1 (a)), the model needs to identify the store name (Domino's) and use prior knowledge to associate it with the correct website.

2. ***Knowledge-aware Image Retrieval:*** This task [58] requires retrieving images by grounding visual cues, such as scene text or layout, in external knowledge about the entities depicted. Unlike conventional text-to-image retrieval that rely purely on visio-lingual similarity, knowledge-aware retrieval demands representations that integrate both factual and commonsense grounding. For instance, in the query "crowded pizza store" (see Figure 1.2), the model is required to recognize that storefront names like "Domino's" or "Pizza Hut" refer to pizza outlets, which requires factual knowledge. At the same time, it needs to assess whether the scene appears crowded, which involves commonsense visual interpretation.

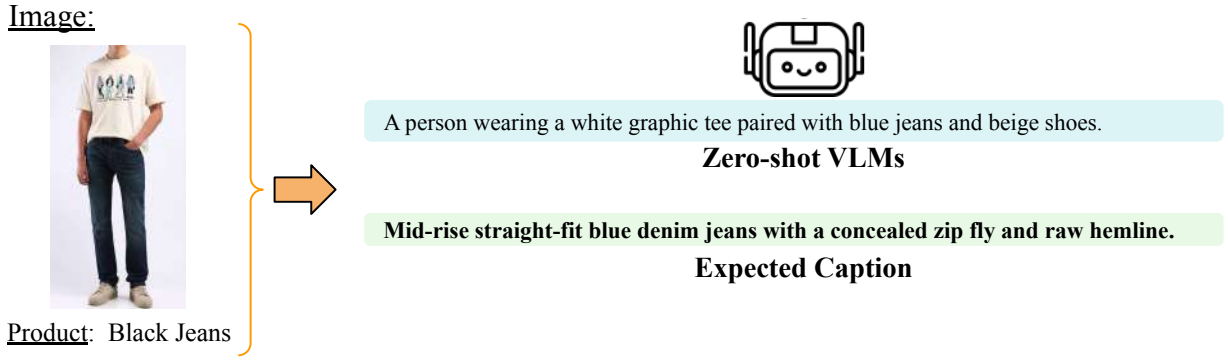


Figure 1.3: Illustration of domain-specific knowledge limitation in zero-shot VLMs. Given a fashion image and product tag ("Black Jeans"), the model generates a generic description that overlooks fine-grained, expert-level attributes; Whereas, the expected caption reflects domain-specific terminology and structured knowledge.

3. ***Multi-image Reasoning:*** This task [15, 24, 164] involves answering questions that require reasoning over a set of images rather than a single input. As discussed in the previous task, the system must first retrieve relevant visual contexts based on both factual and commonsense cues. It must then aggregate information across the retrieved images to produce a coherent answer. Continuing with the example in Figure 1.2, to answer the question "How many pizza restaurants are crowded?", the model must count only those images that satisfy both conditions. A key challenge arises when some images only partially match the query, such as a Domino's outlet that is not crowded (orange box). The model must filter such cases and combine evidence from multiple sources to arrive at the correct answer. This setting requires not only accurate retrieval but also robust multi-step reasoning over visual inputs, which most vision-language models are not equipped to perform natively.
4. ***Knowledge-aware Image Captioning:*** This task focuses on generating captions enriched with fine-grained and domain-specific knowledge, which is especially important in specialized fields like fashion. For example, captioning a fashion image (Figure 1.3) requires recognizing specific types and attributes of garments. Although conventional models produce generic outputs like *"A person wearing a white graphic tee paired with blue jeans and beige shoes"*, knowledge-aware captioning aims for more precise and interpretable descriptions that reflect expert understanding.

1.2 CURRENT APPROACHES AND THEIR LIMITATIONS

Today, most knowledge-intensive visual tasks are approached using large, pretrained vision-language models (VLMs) [31, 35, 40, 122, 126, 133, 137, 138, 158, 169, 227, 252, 269, 271]. These models are trained on massive image-text datasets to align visual and textual modalities, and they attempt to answer questions or generate captions directly from the input image and text prompt. Crucially, they operate in a *parametric-only* manner, relying solely on the knowledge encoded in their model parameters during pretraining, without dynamically retrieving or integrating exter-

nal information at inference time. While this approach has driven much of the recent progress in VLM research, it leaves models unable to adapt to unseen or context-specific knowledge, limiting their applicability in many real-world, knowledge-intensive scenarios. We now outline the key challenges underlying these limitations.

1. ***Insufficient Knowledge Grounding from Large-Scale Pretraining.*** Although large-scale pretraining captures broad visual–textual correlations and improves general perception, it often fails to encode structured, long-tail, or factual knowledge linked to visual entities. For example, the learned representation for a storefront with the word "Nike" may not reflect its identity as a sportswear brand, nor capture facts like its founding year (1964), origin (USA), or headquarters (Beaverton, Oregon). This lack of explicit grounding restricts VLM performance on tasks that require reasoning over latent facts beyond what is visible in the image.

2. ***Hallucinations from Shallow Correlations.*** Despite architectural advances from CNN–RNN pipelines [62, 66, 139] to transformer-based VLMs [31, 35, 40, 80, 122, 133, 137, 138, 269, 271], models often rely on superficial patterns memorized during pretraining. When confronted with unseen or ambiguous cases, they generate brittle or fabricated outputs. For instance, as illustrated in Figure 1.1(b), a model misidentifies a Domino's storefront as "Pizza Hut," relying on weak visual cues (depiction of pizza in the image) and memorized brand associations rather than grounding its prediction in the actual visual entity. Such hallucinations arise not from perception errors alone but from the absence of external factual grounding.

3. ***Lack of Domain-Specific Expertise.*** Most VLMs, though pretrained on broad web corpora, lack grounding in specialized domains such as fashion or scientific imaging. As a result, their outputs are often generic or inaccurate when fine-grained, expert-level reasoning is required. For instance, instead of describing "mid-rise straight-fit denim jeans with a raw hemline," they may produce a vague caption like "a person wearing jeans and a T-shirt." Even advanced prompting techniques such as in-context learning [47, 72] or chain-of-thought reasoning [228, 235], struggle to compensate for this gap, as the missing knowledge is not parametric but domain-specific.

4. ***Trade-Offs Between Scale and Efficiency.*** While large VLMs deliver strong performance, they are computationally expensive and memory-intensive, making them impractical for deployment in real-world or resource-constrained environments. Smaller VLMs offer a practical alternative but suffer from limited capacity and weaker generalization on reasoning-intensive tasks. Current methods to improve them, such as distillation from large models using massive pseudo-labeled datasets [22, 192, 264], are inefficient and often overwhelm small models with trivial or noisy examples, rather than focusing on their true weaknesses. Without targeted supervision, small VLMs remain suboptimal learners, unable to efficiently close the performance gap with their larger counterparts.

Together, these challenges expose two key limitations in current vision-language models. First, existing *VLMs fall short in handling knowledge-intensive visual tasks*. They lack grounding in external or domain-specific knowledge (Challenges: 1–3), struggle to reason beyond what is

visible, and often produce hallucinated or generic outputs. Second, even when high-performing models are available, their scale and training requirements make *VLMs inefficient and impractical for real-world deployment* (Challenge: 4). Improving effectiveness requires methods that augment VLMs with rich, external knowledge, tailored to the knowledge requirements of the task. Improving efficiency, in contrast, requires developing smaller, more cost-effective alternatives that can perform competitively on knowledge-intensive tasks. These two directions form the core research objectives of this thesis.

1.3 OBJECTIVES OF THIS THESIS

This thesis is centered around two key objectives: (1) Making vision-language models (VLMs) more effective by enabling them to reason with external knowledge, and (2) Making VLMs more efficient by reducing their reliance on expensive supervision and large-scale training. We discuss these objectives in detail as follows:

1.3.1 OBJECTIVE 1: MAKING VLMS EFFECTIVE THROUGH KNOWLEDGE AUGMENTATION

This objective focuses on improving the factual reasoning capabilities of vision-language models in knowledge-intensive settings. The aim is to develop modular knowledge augmentation strategies that inject external knowledge, either textual, visual, or multimodal, based on task requirements. These strategies should enhance factual grounding and reasoning ability without extensive re-training. To address this objective, we explore the following research questions.

RQ1. How can textual knowledge be effectively augmented into vision-language models to improve retrieval performance in knowledge-intensive scenarios?

To address this question, we introduce the COFAR dataset, a benchmark designed for the image retrieval tasks that require commonsense and factual reasoning. COFAR focuses on real-world visual entities such as brands, public figures, and landmarks, where retrieval depends not just on visual-text similarity, but on background factual knowledge associated with the entities. We propose KRAMT (Knowledge Retrieval-Augmented Multimodal Transformer), a unified framework that enhances query-image retrieval by integrating external textual knowledge. KRAMT operates in three stages: (i) visual entity linking, where scene text (e.g., store-front names) is matched to Wikipedia entities using a weighted combination of textual and visual similarity, and the top-ranked entity is selected; (ii) knowledge retrieval, where factual descriptions and structured information associated with the linked entity are extracted from Wikipedia; and (iii) Multimodal fusion, where the image, query, and retrieved knowledge are encoded together using a multimodal transformer to compute the query-image relevance, grounded in the retrieved knowledge. This design enables the model to move beyond shallow visual matching and reason over factual associations grounded in external knowledge. Extensive experiments on COFAR demonstrate that KRAMT consistently outperforms competitive retrieval baselines [52, 122, 138, 145, 261] including

Figure 1.4: An illustration of a challenge associated with visual entity linking: "HP" text could refer to (a) Hewlett-Packard or (b) Hindustan Petroleum, requiring both textual and visual context for disambiguation.

models trained on 10x larger data [138, 169], validating that factual knowledge, when effectively retrieved and augmented, significantly improves retrieval in knowledge-intensive scenarios.

We published this work in the following proceedings: *Prajwal Gatti, **Abhirama Subramanyam Penamakuri**, Revant Teotia, Anand Mishra, Shubhashis Sengupta, Roshni Ramnani. "COFAR: Commonsense and Factual Reasoning in Image Search", Asia-Pacific Chapter of the Association for Computational Linguistics- International Joint Conference on Natural Language Processing (AACL-IJCNLP), 2022.* We describe this work in Chapter 3.

RQ2. How can textual knowledge be effectively incorporated into vision-language models to improve accuracy in knowledge-intensive visual question answering?

To investigate this question, we design a two-stage framework that first links visual text to external entities and then integrates the factual knowledge associated with those entities into the VQA pipeline. We propose VisTEL (Visual Text Entity Linker), a module that jointly leverages OCR-extracted scene text and visual context to identify the most plausible entity from a candidate set. For instance, as illustrated in Figure 1.4, when presented with the text "HP" in a storefront image, VisTEL effectively distinguishes between entities like Hindustan Petroleum and Hewlett-Packard by leveraging both textual and visual cues. Once the top-ranked entity is identified by VisTEL, we retrieve its corresponding factual knowledge and feed it into KaLMA (Knowledge-aware Language Model Adapter). KaLMA incorporates this external knowledge alongside the image-question pair, enabling factually grounded answer generation. Importantly, KaLMA also outputs the supporting fact alongside the answer, enabling more interpretable and faithful reasoning. Through extensive experiments, we show that the VisTEL-KaLMA pipeline consistently outperforms prior knowledge-aware VQA methods [195, 229] as well as leading state-of-the-art large vision-language models [40, 133, 252, 271], demonstrating the effectiveness of factual knowledge augmentation in improving answer accuracy for knowledge-intensive VQA.

We published this work in the following proceedings: ***Abhirama Subramanyam Penamakuri***, *Anand Mishra. "Visual Text Matters: Improving Text-KVQA with Visual Text Entity Knowledge-aware Large Multimodal Assistant", Empirical Methods on Natural Language Processing (EMNLP), 2024.* We describe this work in Chapter 4.

RQ3. How can visual knowledge, in the form of retrieved supporting images, be incorporated into vision-language models to improve reasoning in multi-image visual question answering?

To explore this question, we introduce the RetVQA task, where each question is accompanied by a pool of candidate images, and answering requires retrieving and reasoning over a subset that provides the necessary visual evidence. The core challenge lies in identifying the relevant images among many distractors and aggregating their content meaningfully to generate an answer. We curate the RetVQA dataset using images from Visual Genome [110], formulating multi-image questions where supporting information is distributed across multiple images, making retrieval a prerequisite for accurate reasoning. To address this challenge, we develop a retrieval-augmented framework built on MI-BART (Multi-Image BART). Given a question and a pool of images, the system first retrieves a small set of relevant images using a dense retriever trained for semantic alignment on MS-COCO [129]. These retrieved images are then jointly encoded with the question using the MI-BART encoder-decoder architecture, allowing the model to attend across multiple images and reason over the aggregated visual context. Through extensive experiments on the RetVQA benchmark, we show that our retrieval-augmented setup significantly outperforms both competitive single-image VQA baselines [269] and retrieval-free multi-image variants such as image-stitching MI-BART. These results demonstrate that retrieving relevant visual evidence (visual knowledge) provides effective context, enabling more accurate reasoning in multi-image, knowledge-intensive VQA tasks.

We published this work in the following proceedings: ***Abhirama Subramanyam Penamakuri***, *Manish Gupta, Mithun Das Gupta, Anand Mishra. "Answer Mining from a Pool of Images: Towards Retrieval-Based Visual Question Answering", International Joint Conference on Artificial Intelligence (IJCAI), 2023.* We describe this work in Chapter 5.

RQ4. How can multimodal knowledge be effectively incorporated into vision-language models to generate faithful and interpretable captions in domain-specific (fashion) image captioning?

Fashion image captioning requires accurate identification of fine-grained attributes (e.g., sleeve style, fabric type, neckline) and the use of domain-specific terminology. However, the dynamic nature of fashion inventories makes supervised approaches impractical due to their reliance on curated annotations and frequent model retraining. We investigate whether providing effective context, retrieved from visually similar product exemplars, can compensate for the lack of training in frozen vision-language models and enable them to generate accurate, interpretable captions. To this end, we propose RA-CoA (Retrieval-Augmented Chain of Attributes), a training-free and model-agnostic framework tailored for fashion image captioning. RA-CoA decomposes the captioning task into a structured reasoning pipeline: (i) retrieval of visually similar fashion items from a curated knowledge base (ProductKB) to extract relevant attribute types; (ii) prompting frozen vision-language models to infer attribute values independently; and (iii) generating a final cap-

tion conditioned on these attribute-value pairs and the retrieved exemplars. This modular, interpretable, chain-of-attribute reasoning improves faithfulness by grounding generation in both visual cues and retrieved multimodal knowledge (images and structured attributes), while enhancing interpretability through explicit intermediate steps. Extensive experiments on the FashionGen dataset [180] with both open [12, 32, 268] and closed-source VLMs [156] show that RA-CoA significantly outperforms prior training-free paradigms, including in-context learning (ICL) [47, 72], implicit chain-of-thought (CoT-i) [228], and explicit chain-of-thought (CoT-e) [235] prompting, across both automatic and LLM-as-judge metrics, while maintaining zero-shot adaptability for real-world deployment.

This work is under review: ***Abhirama Subramanyam Penamakuri****, Shreya Shukla*, Anand Mishra. "RA-CoA: Training-free Fashion Image Captioning via Retrieval-Augmented Chain-of-Attributes", [Under review at Transactions of Machine Learning Research] (*: equal contribution)*. We describe this work in Chapter 6.

1.3.2 OBJECTIVE 2: MAKING VLMS EFFICIENT THROUGH TARGETED SUPERVISION

This objective addresses the scalability and deployability of vision-language models by reducing their dependence on large-scale labeled datasets and compute-heavy training. The aim is to enable small VLMs to perform competitively on knowledge-intensive tasks without full supervision or fine-tuning. To address this objective, we explore the following research question.

RQ1. How can supervision guided by knowledge disparities enable efficient training of small vision-language models without human-labeled data?

To address this question, we propose Model Parity Aligner (MPA), a label-free pseudo-supervision framework that leverages knowledge disparities between large and small models to drive efficient small model training. MPA operates by first generating pseudo-annotations for a given task using a frozen L-VLM over unlabeled images. It then filters these using a Parity-Identifier (PI) module, which retains only those samples where (i) the L-VLM's answer is correct but the S-VLM fails, or (ii) both models produce different, incorrect outputs. These PI-filtered samples expose the knowledge gaps in the S-VLM and form a focused training subset. This targeted filtering discards trivial or hallucinated cases, ensuring supervision is both efficient and impactful. We evaluate MPA across four diverse VQA benchmarks over ten S-VLM-L-VLM pairs, demonstrating that parity filtering enables effective and scalable knowledge transfer tailored to the unique limitations of small models.

We published this work in the following proceedings: ***Abhirama Subramanyam Penamakuri****, *Navlika Singh*, Piyush Arora*, Anand Mishra. "When Big Models Train Small Ones: Label-Free Model Parity Alignment for Efficient Visual Question Answering using Small VLMs", Empirical Methods on Natural Language Processing (EMNLP), 2025. (*: equal contribution)*. We describe this work in Chapter 7.

1.4 ORGANIZATION OF THIS THESIS

With the above objectives and research questions discussed, the remainder of this thesis is organized as follows. Chapter 2 reviews related work on vision-language models and provides essential background on transformer architectures, the adaptation of large language models to vision-language models, and common training paradigms such as pre-training and supervised fine-tuning. Chapter 3 presents the COFAR benchmark and the KRAMT framework for knowledge-aware image retrieval. Chapter 4 presents VisTEL and KaLMA for knowledge-aware visual question answering. Chapter 5 describes the RetVQA benchmark and the proposed multi-image reasoning framework. Chapter 6 introduces RA-CoA for domain-specific image captioning. Chapter 7 details the Model Parity Aligner for efficient small vision-language models. Finally, Chapter 8 concludes the thesis and outlines future research directions.

2

Vision-Language Models: Background and Overview

In this chapter, we provide a detailed overview of the evolution of vision-and-language (V-L) research over the past decade. We begin by revisiting traditional V-L models, with particular emphasis on two-stream architectures, and then move to state-of-the-art transformer-based approaches for joint visual–textual learning. Specifically, the chapter covers: (i) a survey of visual and textual encoding techniques from the pre-transformer era, (ii) the theoretical foundations of multi-head attention in neural networks, i.e., transformers, and (iii) an in-depth review of modern transformer-based V-L models, including recent advances in large vision-language models. Readers already well-versed in these concepts may skip this chapter without loss of continuity.

2.1 INTRODUCTION

Humans interpretation and interception of the world is multi-modal in nature, i.e., we perceive the world through vision and communicate heavily through natural language. Towards this end, the larger goal of the community is to able an AI system to perceive things and communicate as humans do. To this end, a variety of multimodal tasks have been proposed (key VL tasks are illustrated in Figure 2.1), particularly at the intersection of vision and language, such as:

1. ***Image captioning***: The task is to generate a natural language caption that describes the given image [129, 168].
2. ***Cross-modal retrieval***: The task is to retrieve relevant images given a text query (text-to-image retrieval), or retrieve the corresponding caption given an image (image-to-text retrieval) [129, 168].
3. ***Visual grounding***: The task is to localize a textual expression or phrase within an image by grounding it to the corresponding region [100].

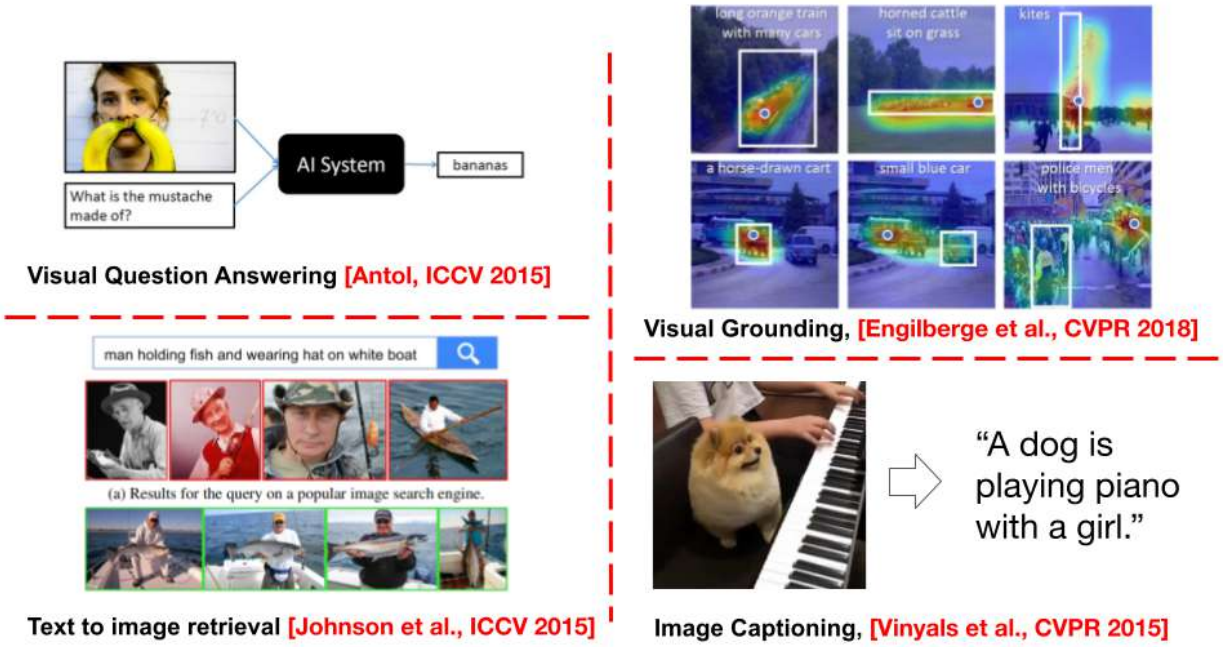


Figure 2.1: Vision-Language Tasks: (i) ***Visual question answering*** [7]: given an image and a question, task is to generate the answer, (ii) ***Visual grounding*** [51]: Given a referring phrase and an image, ground (locate) the phrase in the image, ***Text-to-image retrieval*** [129, 168]: Given a text query retrieve relevant images, ***Image captioning*** [129, 168]: Given an image, task is to generate a caption that describes the image.

4. ***Visual question answering***: The task is to answer a question about an image based on its visual content [7, 62].

5. ***Visual dialog***: The task is to engage in a multi-turn, free-form natural language conversation about an image, resembling human-like dialog [42].

6. ***Visual entailment***: The task is to determine whether a given statement about an image is entailed by, contradicts, or is neutral with respect to the visual content [240].

7. ***Vision-and-Language Navigation (VLN)***: The task is to enable an AI agent to navigate through an environment by following natural language instructions [237].

Representation alignment between visual and textual entities is the key to solving downstream vision-language tasks, e.g., representation for 'image of cat' and the representation for the word 'cat' should be closer in the semantic shared space. This brings us to understand how the information from these two different modalities is encoded into vectors. Starting from non-neural network-based techniques to neural network-based methods there has been growing interest in embedding images [39, 41, 71, 111, 136, 194] and text [17, 94, 141, 149, 167, 181, 204] into vectors. The advent of the transformer architecture [213] demonstrated its ability to model complex dependencies among its inputs. Language models such as BERT [46] and GPT-1 [170] showed that large-scale pretraining of transformers, followed by task-specific fine-tuning, yields state-of-the-art performance on a variety of downstream tasks, even when the supervised datasets are

relatively small. Subsequent works [38, 112, 115, 134, 171, 173] have refined the pretraining objectives to learn more generalizable representations, further strengthening the paradigm. Motivated by these advances, the majority of the prior state-of-the-art methods in vision-language space that use a two-stream network [5, 7, 51, 99, 216, 216, 251] to solve these tasks (CNN-based backbone [71, 179] variants as visual encoder, RNN [18]/LSTM [75] variants as text encoder) have started to leverage transformer architectures by proposing various self-supervised objectives over the large-scale image-caption data.

This chapter is organized as follows: Section 2.2 discusses V-L research prior to transformers, Section 2.3 provides a quick background on transformers, Section 2.4 discusses transformers in vision and language, and finally, Section 2.5 discusses large vision-language models.

2.2 VISION LANGUAGE RESEARCH PRIOR TO TRANSFORMERS

Solving tasks in vision-language space needs one to jointly align representations of both visual and textual inputs. Most of the prior works, have a separate encoder for vision inputs and text inputs, and then learn the alignment using distance-based metrics. To this end, we briefly discuss the progress in visual encoding methods and textual encoding methods, with their limitations.

2.2.1 VISUAL ENCODING

Visual encoding is the task of obtaining a feature vector/embedding for an image, that can be further used to train models for downstream tasks like classification [44]. Prior to deep learning, many vision tasks relied on hand-crafted features such as SIFT [136], SURF, HOG [41], and BoVW [39]. The rise of deep learning, coupled with increased computational capabilities, enabled the training of larger and deeper neural networks. In particular, models pre-trained on ImageNet [44] for image classification [71, 85, 111, 194] have proven highly transferable and are widely adopted for a variety of smaller-scale vision tasks. Following this trend, ImageNet-pretrained convolutional neural networks have become the de facto choice of visual encoders in vision-language applications [7, 129].

2.2.2 TEXTUAL ENCODING

Early approaches to textual encoding treated words as atomic units from a fixed vocabulary. Under this assumption, representations were constructed using one-hot vectors, bag-of-words models, and frequency- or count-based statistics [141], including inverse document frequency [94] and term frequency–inverse document frequency (TF–IDF) [181]. The first neural method for learning word representations in a continuous space was the Neural Network Language Model (NNLM) [17], which learned embeddings through the statistical language modeling objective. Building on this idea, subsequent methods such as GloVe [167], which derives embeddings from

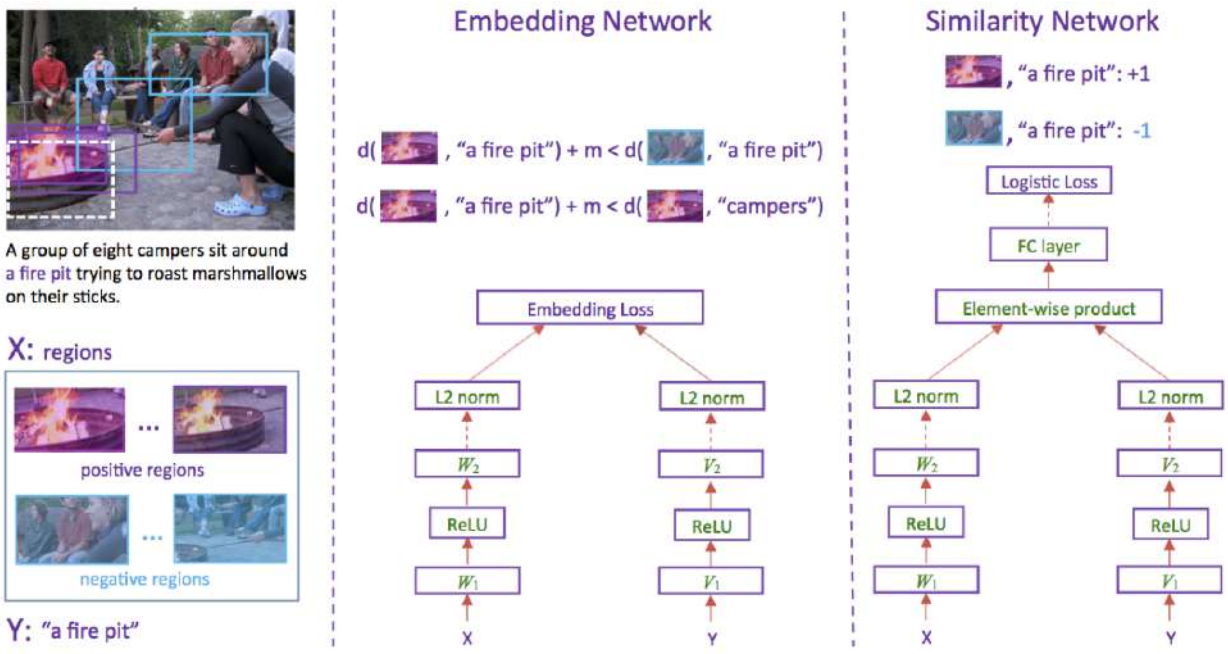


Figure 2.2: An example of two stream network. Image taken from [219]

co-occurrence statistics, and Word2Vec [149], which learns embeddings by predicting words from their surrounding context, became widely adopted for generating distributed word representations.

In the wake of deep learning breakthroughs, particularly following the success of AlexNet [111], Authors in [204] proposed sequence-to-sequence (seq2seq) models based on recurrent architectures such as RNNs [18] and LSTMs [75] for natural language modeling. Although originally introduced for machine translation, seq2seq models soon became widely adopted for encoding textual inputs. LSTMs, in particular, were preferred over vanilla RNNs due to their ability to (i) capture long-range dependencies more effectively, (ii) mitigate the vanishing gradient problem, and (iii) regulate the flow of relevant and irrelevant information through gating mechanisms.

Despite these advantages, recurrent sequence models suffer from notable drawbacks. They are computationally expensive and inherently sequential, making them difficult to parallelize. Furthermore, representing an entire input sequence as a single fixed-length vector often leads to information loss, especially for long sequences where earlier inputs may be inadequately preserved.

2.2.3 ATTENTION

To address these limitations, Bahdanau et al. [10] introduced the attention mechanism, which allows models to dynamically focus on different parts of the input sequence when generating outputs. Unlike fixed-length representations, attention enables the modeling of relationships

among tokens irrespective of their distance in the sequence. Although originally proposed for machine translation, attention has since been widely adopted across NLP and vision-language tasks [54, 105, 244, 251].

2.2.4 TWO-STREAM NETWORKS

A model's performance on vision-language tasks depends on the alignment learned between visual and linguistic entities. To this end, many popular approaches for tasks such as caption-to-image retrieval, image-to-caption retrieval, visual grounding, and visual question answering adopt a two-stream architecture, where one branch encodes the visual input and the other encodes the text. Their interaction is typically realized through similarity measures, embedding alignment, or attention-based mechanisms. In early formulations, the visual stream is often implemented using CNN-based encoders, while the linguistic stream relies on RNN/LSTM encoders to capture sequential dependencies in language [5, 7, 51, 99, 216, 251]. An example of two-stream network [219] is shown in Figure 2.2.

2.3 BACKGROUND ON TRANSFORMER

The transformer [213] was proposed to overcome the limitations of recurrent sequence models [18, 75] by entirely replacing recurrence with self-attention mechanisms [33]. Unlike RNNs and LSTMs, the transformer architecture enables full parallelization during training and captures long-range dependencies without the constraint of fixed-length representation. It follows an encoder–decoder design, where both components are built from stacked layers of self-attention and feed-forward networks. In the following sections, we provide an overview of the key components of the transformer, namely, encoder, decoder, self-attention, multi-head self-attention, and encoder–decoder attention modules. Illustrated in Figure. 2.3 (image taken from [213]).

2.3.1 TRANSFORMER ENCODER

A transformer encoder consists of a stack of identical encoder layers, with the original implementation employing six layers [213]. Each encoder layer is composed of two key sub-modules: (i) a multi-head self-attention (MHSA) mechanism (discussed in Section 2.3.1), and (ii) a position-wise feed-forward network (FFN), which is a simple multilayer perceptron (MLP) with two linear layers separated by a ReLU activation.

The encoder module operates on a sequence of input token embeddings $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$, n is the sequence length, and d is the embedding dimension of each token $\mathbf{x}_i$. Each sub-module is wrapped with a residual connection followed by layer normalization (LN) [8], as illustrated below:

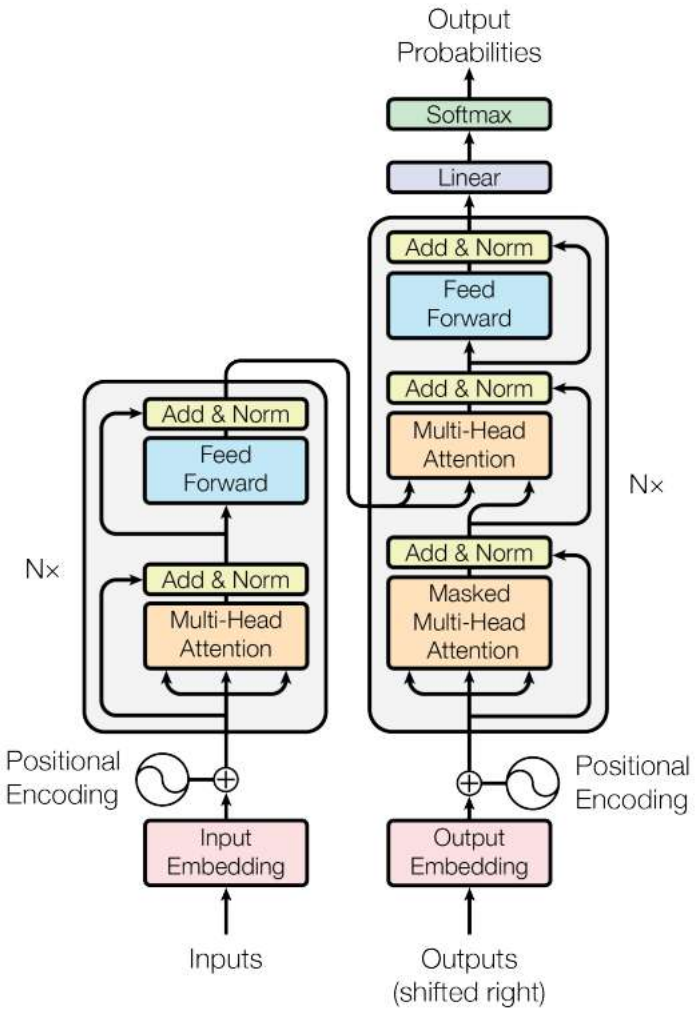


Figure 2.3: An overview of transformer architecture. Image from [213]

$$\begin{aligned}\hat{\mathbf{Z}} &= MHSA(\mathbf{X}),\\ \mathbf{Z} &= LN(\mathbf{X}+\hat{\mathbf{Z}}),\end{aligned} \tag{2.1}$$

The FFN is then applied independently to each position, with weights shared across all positions:

$$\tilde{\mathbf{Z}} = FFN(\mathbf{Z}) = \text{ReLU}(\mathbf{Z}\mathbf{W}_a + \mathbf{b}_a)\mathbf{W}_b + \mathbf{b}_b, \tag{2.2}$$

Finally, a second residual connection and layer normalization are applied to incorporate the FFN output:

$$\mathcal{Z} = LN(\mathbf{Z} + \tilde{\mathbf{Z}}), \tag{2.3}$$

where $\mathbf{W}_a, \mathbf{W}_b$ and $\mathbf{b}_a, \mathbf{b}_b$ denote trainable parameters. The output $\mathcal{Z} \in \mathbb{R}^{n \times d}$ constitutes the updated representation of the input sequence produced by one encoder layer.

SELF-ATTENTION

Self-attention is at the core of the transformer architecture. The objective of the self-attention mechanism is to model dependencies among all tokens in the input sequence. In particular, it computes a context-aware representation of each token by incorporating information from all other tokens in the sequence. Unlike additive attention [10], the transformer employs scaled dot-product attention, which is both conceptually simpler and computationally more efficient.

In each encoder layer, self-attention is parameterized by three learnable projection matrices: $\mathbf{W}_Q \in \mathbb{R}^{d \times d_{out}}$, $\mathbf{W}_K \in \mathbb{R}^{d \times d_{out}}$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d_{out}}$. Given input token embeddings $\mathbf{X}$, the queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) are obtained as:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V. \tag{2.4}$$

For each token, the attention mechanism computes similarity scores between its query vector and the key vectors of all tokens in the sequence. These raw dot-product scores are then scaled by a factor of $\sqrt{d_k}$, where d_k is the key dimension. This scaling prevents the dot products from growing too large in magnitude, which could otherwise push the softmax function into regions with very small gradients and hinder training stability. The scaled scores are normalized with a softmax operation to produce attention weights. Finally, the representation of each token is computed as a weighted sum of the value vectors of all tokens, where the attention weights determine the relative contribution of each token:

$$\hat{\mathbf{Z}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{2.5}$$

MULTI-HEAD SELF-ATTENTION

Multi-head self-attention (MHSA) extends the basic self-attention mechanism by running multiple attention operations, referred to as attention heads, in parallel. Each head attends to the input sequence independently, enabling the model to capture token interactions in different representation subspaces. In the original transformer, $h = 8$ attention heads were used [213].

Formally, each head $i \in \{1, 2, \ldots, h\}$ is parameterized by its own set of projection matri-

ces $\{\mathbf{W}_Q^i, \mathbf{W}_K^i, \mathbf{W}_V^i\}$, and produces an output $\hat{\mathbf{Z}}^i \in \mathbb{R}^{n \times d_{out}}$. The outputs from all heads are concatenated and linearly projected back to the model dimension d using a projection matrix $\mathbf{W}_O \in \mathbb{R}^{hd_{out} \times d}$, as shown in Eq. 2.6:

$$\hat{\mathbf{Z}} = \mathbf{W}_O \, [\, \hat{\mathbf{Z}}^1 \parallel \hat{\mathbf{Z}}^2 \parallel \cdots \parallel \hat{\mathbf{Z}}^h \,], \tag{2.6}$$

where $\parallel$ denotes concatenation along the feature dimension. The combined output $\hat{\mathbf{Z}}$ is then passed through the residual connection and layer normalization, as described in Eq. 2.1.

POSITIONAL ENCODING

Since the self-attention mechanism does not rely on recurrence or convolution, it is inherently permutation-invariant; that is, it computes attention scores among tokens without considering their order in the sequence. To incorporate positional information, each input token embedding is augmented with a positional encoding. In the original transformer [213], sinusoidal position encodings were employed to provide absolute positional information.

The sinusoidal formulation defines the encoding for a token at position pos and dimension i as:

$$PE_{(pos,2i)} = \sin\Big(\frac{pos}{10000^{2i/d_{in}}}\Big), \tag{2.7}$$

$$PE_{(pos,2i+1)} = \cos\Big(\frac{pos}{10000^{2i/d_{in}}}\Big), \tag{2.8}$$

where d_{in} is the input embedding dimension. These encodings inject position-dependent variations into the embeddings and allow the model to generalize to longer sequences by enabling the representation of relative positions through linear combinations of sinusoidal functions.

2.3.2 TRANSFORMER DECODER

Similar to the transformer encoder, the transformer decoder is composed of a stack of identical layers, with the original implementation employing six layers [213]. Each decoder layer contains three sub-modules, as opposed to two in the encoder: (i) multi-head masked self-attention (MHMSA), (ii) encoder–decoder attention, and (iii) a position-wise feed-forward network (FFN).

The key distinction between the encoder and the decoder lies in the self-attention mechanism. While the encoder computes attention over the entire input sequence, the decoder is restricted to attending only to preceding positions. This masking enforces causality, ensuring that predictions

at a given position cannot incorporate information from future tokens, consistent with the next-word prediction training objective.

MASKED SELF-ATTENTION

Masked self-attention modifies the standard self-attention mechanism by preventing tokens from attending to future positions. This is achieved by assigning $-\infty$ to the attention logits corresponding to future positions prior to the softmax operation, thereby forcing their attention weights to zero.

ENCODER–DECODER ATTENTION

The encoder–decoder attention module is structurally similar to multi-head self-attention, with the difference that the queries are computed from the decoder's previous layer, while the keys and values are derived from the encoder outputs. This allows the decoder to incorporate context from the source sequence (encoder) while generating target sequence representations.

The transformer architecture was originally proposed for machine translation, where the encoder processes the source sequence and the decoder generates the target sequence. However, owing to its ability to learn powerful contextualized representations through self-attention, the transformer framework has since been widely adopted in multimodal learning. In particular, our discussion in the following sections focuses on two major directions: (a) vision–language transformers, which aim to learn effective cross-modal alignment between visual and textual entities, and (b) large vision–language models (LVLMs), which extend this paradigm to large-scale pretraining for unified multimodal reasoning.

2.4 TRANSFORMERS IN VISION-LANGUAGE

Building on the success of transformers in natural language processing, where large-scale pretraining followed by supervised fine-tuning [46, 134, 171] has become the dominant paradigm, researchers have extended this approach to the multimodal setting. In particular, Vision–Language Pretraining (VLP) refers to training transformer-based models on large-scale multimodal datasets with the objective of learning joint representations of images and text. The central goal of VLP is to encode effective cross-modal alignment between visual and textual modalities, enabling transfer to downstream vision–language applications (as discussed in Section 2.1) and yielding state-of-the-art performance. A summary of representative VLP methods is provided in Table 2.1.

In the remainder of this section, we first review the pretraining objectives commonly employed in VLP, followed by an overview of the datasets that enable large-scale vision–language

Table 2.1: An Overview of key works in Vision-Language Pretraining

Study	*Architecture type*	*Data*	*Key pretraining objectives*
		Visual Backbone: VG pre-trained Faster R-CNN	
ViLBERT [137]	Two-stream	Conceptual Captions	Masked multi-modal modeling Multi-modal alignment
VisualBERT [122]	Single Stream	MS-COCO	Masked language modeling with image Image-text matching
B2T2 [4]	Single stream	Conceptual Captions	Masked language modelling Imposter identification (similar to ITM)
Unicoder-VL [118]	Single Stream	Conceptual Captions SBU Captions	Masked language modeling Masked object classification Visual-linguistic matching
LXMERT [206]	Two-stream	MS-COCO Visual Genome	Masked cross-modal LM RoI Feature Regression Masked Object Detection Cross Modality Matching Image Question Answering
VL-BERT [199]	Single Stream	Conceptual Captions	Masked language modeling with visual cues Masked RoI classification with linguistic cues
VLP [269]	Single Stream	Conceptual Captions	Masked language modeling Seq2seq objective
UNITER [31]	Single Stream	MS-COCO Conceptual Captions SBU Captions Visual Genome	Masked language modeling Image-text matching Word region alignment Masked region feature regression Masked region classification Masked region classification with KL Divergence
OSCAR [126]	Single stream	MS-COCO Conceptual captions Flickr30k SBU captions GQA	Masked token loss Contrastive Loss
		Visual backbone - ImageNet [44] pre-trained CNN [71]	
PixelBERT [86]	Single stream	MS-COCO Visual Genome	Masked language modeling Image-Text matching
CLIP [169]	Two stream	LAION-400M	Image-Text contrastive learning
		No convolution	
ViLT [106]	Single stream	MS-COCO Visual Genome SBU Captions Conceptual Captions	Image-Text matching Masked language modeling (whole word masking)
ALBEF [121]	Two stream (followed by single stream) (momentum-based model)	MS-COCO Conceptual Captions SBU Captions Visual Genome	Masked language modeling Image-text matching Image-Text contrastive learning
SimVLM [226]	Single stream	ALIGN	Prefix language modeling (PLM)
VLMo [225]	Single stream (different FFN for each modality)	MS-COCO Conceptual Captions SBU Captions Visual Genome	Masked language modeling Image-Text matching Image-Text contrastive learning
TCL [248]	Two stream (followed by single stream) (momentum-based model)	MS-COCO Conceptual Captions SBU Captions Visual Genome	Cross modal alignment Intra modal contrastive Local mutual information maximization Image text matching Masked language modeling

pretraining.

2.4.1 COMMON PRETRAINING OBJECTIVES IN VLP

1. ***Masked Language Modeling with Vision (MLM)***: Analogous to the MLM objective in BERT [46], a fixed proportion of tokens (typically 15%) in the text input are masked, and the model is trained to predict them using both the surrounding text and the associated image context.
2. ***Image–Text Matching (ITM)***: Inspired by the next sentence prediction task in BERT [46], the model is presented with both aligned (image–caption) and misaligned pairs, and is trained to distinguish between the two [122]. This objective is also referred to as multi-modal alignment, visual–linguistic matching, or cross-modal matching.
3. ***Masked Object Classification***: Region-of-interest (RoI) tokens are randomly masked, and the model is trained to predict the corresponding object labels, typically obtained from a detector such as Faster R-CNN [179]. This objective is also called masked region classification or masked RoI classification.
4. ***Masked RoI Feature Regression***: Instead of predicting discrete labels, the model reconstructs the continuous feature vectors of masked RoIs, encouraging richer visual representation learning [31].
5. ***Masked Object Classification with KL Divergence***: The distribution of predicted object classes from the VLP model is aligned with the distribution produced by a Faster R-CNN detector using KL divergence as the training signal [206].
6. ***Image–Text Contrastive Learning***: For each aligned image–caption pair, all other image–caption combinations in the batch are treated as negatives. The objective is to maximize similarity for positive pairs while minimizing it for negatives [121, 169].
7. ***Masked Token Loss with Object Tags***: OSCAR [126] introduces object tags obtained from Faster R-CNN as additional text tokens. The objective extends MLM to these object tokens, strengthening visio–linguistic alignment.
8. ***Prefix Language Modeling***: Given the image context and the prefix of a sentence, the model is trained to autoregressively generate the remaining part of the caption [226].
9. ***Intra-Modal Alignment***: The model maximizes agreement between two different augmented views of the same input (either image or text), often within momentum-based contrastive frameworks [121, 248].
10. ***Local Mutual Information (MI) Maximization***: The model encourages local region representations from one view of an image to align closely with the global representation (e.g., [CLS] token) of another view, facilitating fine-grained alignment [121, 248].

2.4.2 DATASETS

The majority of vision-language models (VLMs) are pretrained on large-scale image-caption pairs, which are cheaply available (over the web) with minimum human annotation efforts. Some of the

Table 2.2: Selected Vision Language Pretraining Datasets (not exhaustive)

Dataset	*Number of Image-caption pairs*
MS-COCO [129]	1.2M
Visual Genome [110]	5M
Conceptual Captions [188]	3.3M/12M
SBU [157]	1M
Flickr30k [168]	0.1M
GQA [87]	113K
ALIGN [91]	1.8B
LAION [184]	400M/5B

widely used datasets in this context are summarized in Table 2.2.

2.5 LARGE VISION-LANGUAGE MODELS

2.5.1 FROM LLMS TO GENERATIVE L-VLMS

The emergence of large language models (LLMs) such as GPT-3 [21], PaLM [36], LLaMa [211], Vicuna [34] and QwenLM [11] has demonstrated that scaling in terms of data, model parameters, and compute unlocks emergent abilities such as in-context learning, zero-shot generalization, and instruction following. Motivated by these advances, the multimodal community has investigated whether similar scaling laws can yield models that *see and talk*. To this end, Large Vision–Language Models (L-VLMs) integrate pretrained LLMs with visual encoders, enabling open-ended multimodal generation and dialogue. In this thesis, we focus on generative L-VLMs, which extend LLM backbones to reason jointly over visual and textual inputs.

2.5.2 ARCHITECTURAL EXTENSIONS OF LLMS

Generative L-VLMs extend pretrained LLMs into multimodal settings through three key components: a visual encoder, an adapter or projection module, and the LLM backbone. Although this decomposition is common across L-VLMs, different models vary significantly in how each component is designed and integrated, reflecting trade-offs between accuracy, efficiency, and extensibility. An overview of a standard L-VLM architecture is shown in Figure 2.4.

Visual encoders. Most generative L-VLMs employ Vision Transformers (ViTs) [48] as image encoders, owing to their strong representational capacity and architectural compatibility with transformer-based LLMs. Typically, these encoders are initialized from large-scale pretrained checkpoints such as CLIP ViT-L/14 [169], ensuring robust visual features without requiring task-specific supervision. Extending the scaling laws of LLMs, several works have scaled up the vision backbone to enhance perceptual ability. For example, InternVL [32] employs InternViT encoders with

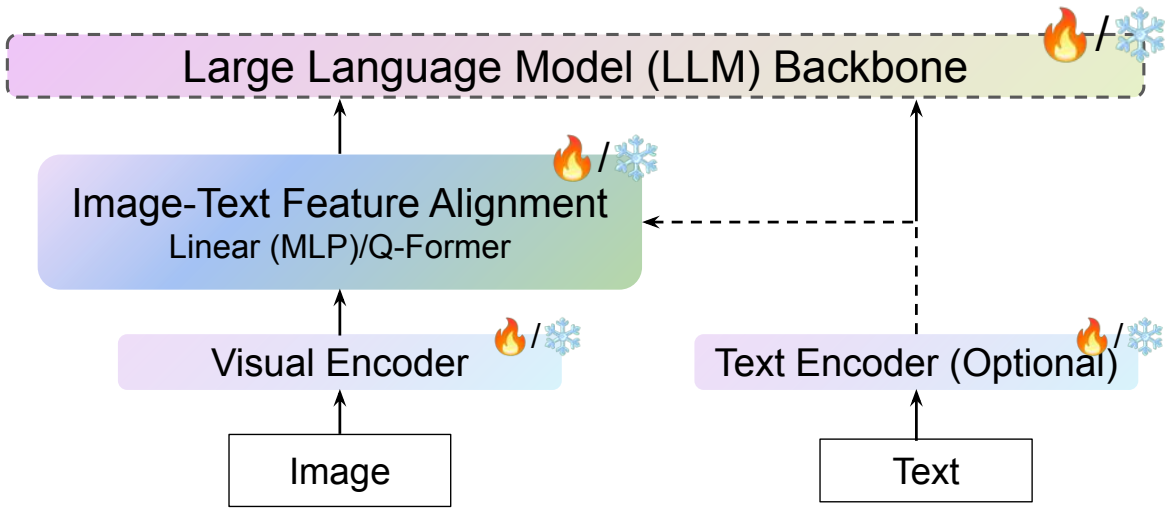


Figure 2.4: Overview of a standard generative L-VLM architecture.

billions of parameters (e.g., InternViT-6B) to align with large LLM backbones. Similarly, EVA-CLIP [202] provides ViT-L and ViT-G checkpoints that improve upon CLIP in recognition and retrieval performance, and has been adopted in models such as MiniGPT-4 v2 [26]. Such large-scale designs aim to capture fine-grained visual details and support complex multimodal reasoning.

In parallel, a complementary trend emphasizes lightweight encoders for efficiency. SigLIP [259] offers compact ViT-B/16 style models trained with a sigmoid-based contrastive loss, providing strong generalization while reducing compute requirements. Lightweight VLMs like SmolVLM-500M [144] and TinyLLaVA [268] pair the SigLIP vision encoder with compact language backbones such as Phi-2 [1], Gemma [209], and Qwen-0.5B [11], enabling instruction-following under resource constraints. Small-scale variants of Qwen-VL and InternVL also rely on reduced ViT backbones for deployment on mobile or edge devices. Together, these two design philosophies: (i) scaling up for perception quality and (ii) scaling down for efficiency, highlight the flexibility of ViTs and their derivatives as the default choice for L-VLM vision encoders.

Adapters and projection layers. Since LLMs operate over text embeddings, visual features must be mapped into a compatible latent space. Two common approaches are (i) direct projection and (ii) query-based adaptation. The simplest method is a linear or MLP projection, as in LLaVA [133], which directly aligns vision embeddings with the LLM token space and enables efficient training. Alternatively, BLIP-2 [120] introduced the Querying Transformer (Q-Former), which learns a fixed set of query tokens to selectively attend to dense vision features and pass compact representations to the LLM. MiniGPT-4 [271] adopts the same Q-Former strategy, combining it with CLIP encoders and Vicuna for instruction following. mPLUG-Owl [252] also leverages a query-based adapter, whereas its successor mPLUG-Owl2 [253] introduces a modality-adaptive module that dynamically modulates interactions across modalities. A different line of work is represented by Flamingo [3], which interleaves gated cross-attention layers throughout the LLM, allowing visual

tokens to influence generation at multiple depths. Overall, these adapter designs reflect distinct philosophies: lightweight direct projection for efficiency, query-based adaptation for structured feature extraction, modality-adaptive layers for flexible fusion, and interleaved cross-attention for deep integration.

LLM backbone integration. L-VLMs differ in how they incorporate the LLM backbone. BLIP-2 [120] kept the LLM entirely frozen and trained only lightweight adapters (the Q-Former and projection layers), enabling faster adaptation. LLaVA-1.0 [133], by contrast, froze the LLM during the vision–language alignment stage but later performed supervised fine-tuning with the LLM unfrozen, allowing deeper adaptation to multimodal data. More recent open-source VLMs, such as Qwen-VL [11] and InternVL [32] support partial or complete fine-tuning of the LLM backbone, achieving stronger integration at the cost of substantially higher compute requirements. Overall, these strategies span a spectrum from frozen backbones with lightweight adapters to full backbone fine-tuning, reflecting trade-offs between efficiency and multimodal performance.

Training stages. The development of generative L-VLMs typically follows a three-stage pipeline:

- ***Stage I: Visual feature alignment.*** The vision encoder is first aligned with the LLM embedding space, typically using large-scale web-scraped image–text pairs (e.g. CC [188], LAION-5B [183]). Pretraining objectives at this stage include contrastive alignment, image–text matching, or lightweight projection learning through image-conditioned generation. This stage equips the model with a basic ability to ground visual tokens in the LLM space.
- ***Stage II: Generic instruction tuning.*** The aligned model is then fine-tuned on curated multimodal instruction datasets (e.g., LLaVA-Instruct [133], ShareGPT4V [27]), which combine human-annotated and GPT-generated conversations. Different L-VLM families (e.g., Qwen2-VL, InternVL2/3) adopt their own mixtures of such resources to endow models with broad, general-purpose instruction-following capabilities.
- ***Stage III: Task-specific instruction fine-tuning.*** Finally, the model is adapted to specialized application domains such as visual question answering, document understanding, or medical imaging using smaller but higher-quality datasets. This stage refines task performance while leveraging the broad generalization abilities acquired in earlier stages.

Parameter-efficient fine-tuning. Given the scale of modern LLM backbones, full fine-tuning is often computationally expensive. To address this, many L-VLMs adopt parameter-efficient fine-tuning (PEFT) techniques. The most widely used is Low-Rank Adaptation (LoRA) [81], which inserts low-rank trainable matrices into frozen weight matrices. LoRA and its variants make it feasible to adapt billion-parameter LLMs for multimodal instruction following at manageable cost, and have consequently become the de facto standard in recent L-VLM development.

Table 2.3: Overview of generative L-VLMs commonly referenced in this thesis, arranged chronologically.

Model	*Adapter / Fusion*	*Vision Encoder*	*LLM Backbone*	*Scale*
InstructBLIP (2023) [40]	Q-Former	BLIP-2 ViT-G/14	Flant5-XL (3B) / XXL (11B), Vicuna-7B/13B	7B–13B
MiniGPT-4 v1 (2023) [271]	Q-Former	CLIP ViT-L/14	LLama-7B	7B
MiniGPT-4 v2 (2023) [26]	Q-Former	EVA-CLIP ViT-L/14	LLama2-7B/13B	7B–13B
LLaVA-1.5 (2023) [133]	Linear projection	CLIP ViT-L/14	Vicuna-7B/13B, LLaMa-7B	7B–13B
mPLUG-Owl (2023) [252]	Query-based adapter	CLIP ViT-L/14	LLaMA-7B	7B
mPLUG-Owl2 (2023) [253]	Modality-adaptive module	CLIP ViT-L/14	LLaMA-7B/13B	7B–13B
Qwen-VL (2023) [11]	Q-Former variant	OpenCLIP ViT-G	Qwen-7B	7B
InternVL v1.0 (2023) [32]	Query-based adapter	InternViT-6B	InternLM-7B	13B
LLaVA-Next (2024) [132]	Linear projection	CLIP ViT-L/14	Vicuna / Qwen / Mistral	7B–34B
InternVL v2.0 (2024) [32]	Modality-adaptive module	InternViT-300M–6B	InternLM2 (1B–20B), LLaMA3-70B	1B–76B
Qwen2-VL (2024) [221]	Linear projection	ViT-675M	Qwen2 (1.5B / 7.6B / 72B)	2B / 7B / 72B
TinyLLaVA (2024) [268]	Linear projection	SigLIP	Phi-2 / Gemma / OpenELM / Qwen2 (0.5B–3B)	0.9B–3.1B
SmolVLM-2 (2024) [144]	Linear projection	SigLIP	SmolLM2	0.5–3B

2.5.3 GENERATIVE L-VLMS RELEVANT TO THIS THESIS

The landscape of generative L-VLMs has expanded rapidly, with a diverse set of models exploring different design choices for visual encoders, adapters, and LLM backbones. Table 2.3 provides a chronological overview of representative L-VLMs that are most relevant to this thesis. The table highlights their adapter or fusion strategy, vision backbone, language backbone, and released model scales. While not exhaustive, these models illustrate the main trajectories of L-VLM development—ranging from early Q-Former based systems such as InstructBLIP and MiniGPT-4, to projection-based approaches like LLaVA and Qwen2-VL, to more advanced designs such as InternVL with modality-adaptive fusion. Together, they demonstrate how scaling and design innovations have shaped both the capabilities and limitations of current L-VLMs.

The progression from early vision–language pretraining models to modern L-VLMs underscores the remarkable advances enabled by large-scale data and transformer scaling. These models exhibit impressive general-purpose multimodal abilities such as zero-shot reasoning, open-ended generation, and multimodal dialogue. However, as outlined in Chapter 1, they also inherit significant limitations: a tendency to hallucinate, over-reliance on parametric memory, limited grounding in domain-specific knowledge, and inefficiencies in training and inference pipelines, particularly for knowledge-intensive tasks. This thesis positions itself within this landscape by addressing two central objectives. First, we explore how L-VLMs can be made more *effective* through external knowledge augmentation, improving factuality and grounding. Second, we investigate how they can be made more *scalable and efficient*, reducing compute overheads while retaining performance. The subsequent chapters build on this background to develop methods that advance L-VLMs toward more reliable, knowledge-aware, and practical multimodal reasoning.

3
Augmenting VLMs with Textual Knowledge for Image Retrieval

In this chapter, we present a unified framework, namely Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT), that treats named visual entities in an image as gateway to encyclopedic knowledge and leverages them, together with the natural language query, to ground relevant external knowledge. KRAMT seamlessly integrates visual content and grounded knowledge within a multimodal transformer to learn robust alignment between images and queries. This unified design enables image retrieval that requires both commonsense reasoning and factual grounding. We evaluate the retrieval performance of KRAMT against competitive baselines on the COFAR benchmark and show that it consistently improves retrieval accuracy in knowledge-intensive scenarios.

3.1 INTRODUCTION

Retrieving relevant images for a natural language query has been an exciting field of research in the vision-and-language community [93, 220, 224]. Most of the available literature focuses on querying visually-evident aspects in the images, such as searching for objects or their interactions in natural scenes. However, as illustrated in Figure 3.1, users often require an image search engine that can perform commonsense reasoning and leverage facts (world knowledge) about the image content. To fill this gap, we propose a novel image search task requiring commonsense and factual reasoning associated with named visual entities.

To study this problem, a suitable dataset is required. While many text-to-image search datasets are publicly available [129, 193, 255], they have not been explicitly created to study our proposed task. Few of the recently introduced knowledge-enabled VQA datasets such as OK-VQA [146], KVQA [186], Text-KVQA [195], FVQA [222] require either factual or commonsense or a combination of both. However, they may not be well-suited for studying the "image search" task we are interested in. Note that in the conventional VQA task, a query (question) is evaluated against a single image which is often directly relevant to the query; whereas, in image search, a query needs

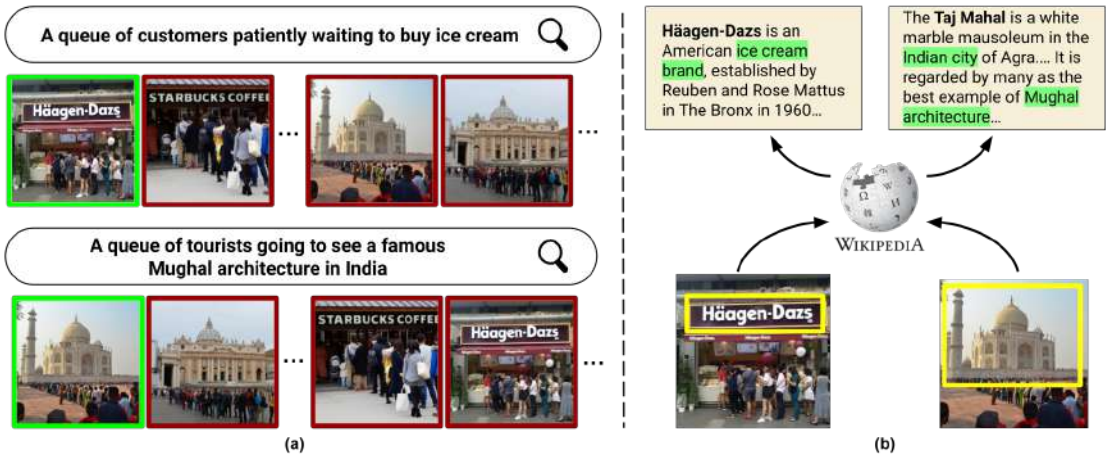


Figure 3.1: Consider the two natural language queries shown in (a). Retrieving images relevant to these queries (shown using a green bounding box) requires a model that has the ability to interpret images beyond just what is visually apparent, such as interpreting – who are customers vs. who are tourists? Who are waiting to buy vs. who are going to see? in other words, visual commonsense. Additionally, the model would need to interpret facts of world knowledge, such as Häagen-Dazs is an ice cream brand and the Taj Mahal in India is an example of Mughal architecture. This can be enabled by linking visual entities in the image to an encyclopedic knowledge source such as Wikipedia. Our work presents such a model, namely KRAMT. ***[Best viewed in color]***

to be evaluated against several thousands of images, including distractors and then needs to rank the relevant image as the top result. Moreover, to our knowledge, there is no dataset available that includes natural scene images containing a diverse set of visual named entities (such as business brands, celebrities, and world landmarks), visual details of the natural scene along with annotations that demands commonsense and factual reasoning associated with the images. To meet these requirements, we present COFAR, which contains manually annotated English language queries for natural scenes containing named visual entities. (A selection of samples in COFAR dataset are shown in Figures 3.2, 3.3, and 3.4).

A plausible approach to addressing our image search problem on COFAR is large-scale vision-language pretraining [138, 169] and learning the associations between commonsense-factual concepts and images. This can be successful in learning popular associations, e.g., Starbucks to Coffee, Eiffel tower to Paris if it has seen such samples during training. However, such methods often require large data and generalize poorly to unseen or rare entities. In contrast, we take a distinct path in this work and ground external knowledge associated with entities in the images to perform commonsense and factual reasoning. To this end, we present a unified model, namely Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT), that retrieves relevant knowledge from Wikipedia by performing query-knowledge similarity-guided visual entity linking. It then encodes the retrieved knowledge, query and visual features, and learns image-query alignment using a multimodal transformer to perform knowledge-aware image search.

The contributions of this work are twofold. The first contribution is the design of a knowledge retrieval–augmented multimodal transformer (KRAMT), a unified framework that aligns queries with the relevant images by performing visual entity linking, retrieving relevant knowledge, and

(a) *Query*: Two people getting married in front of a tower in Paris.
Commonsense: Two people in white gown and suit holding hands leads to the commonsense that they are getting married.
Visual named entity: The Eiffel Tower
Fact: The landmark is Eiffel Tower, which is located in Paris, France.

(b) *Query*: The captain of the Argentina national football team celebrating after scoring a goal.
Commonsense: The person is running cheerfully next to a goalpost leads to commonsense that they are celebrating after scoring a goal.
Visual named entity: Lionel Messi
Fact: Lionel Messi is the captain of the Argentina national football team.

(c) *Query*: Two people showing an interest to purchase a watch.
Commonsense: People looking into the display of a watch store implies they could be interested to purchase a watch there.
Visual named entity: Rolex
Fact: The store Rolex sells watches.

Figure 3.2: A selection of examples from COFAR showing query, relevant image, associated visual named entity, commonsense and fact.

Query: Visitors standing in rain admiring a temple dedicated to the Greece goddess Athena

Visual Named Entity: Parthenon

Knowledge Text: The Parthenon is a former temple on the Athenian Acropolis, Greece, dedicated to the goddess Athena, whom the people of Athens considered their patroness.

Query: A young fan asking the author of the Harry Potter series for an autograph

Visual Named Entity: J. K. Rowling

Knowledge Text: Joanne Rowling (born 31 July 1965), also known by her pen name J. K. Rowling, is a British author and philanthropist. She wrote a seven-volume children's fantasy series, Harry Potter, published from 1997 to 2007.

Query: A white truck parked outside a grocery store waiting to pick up orders

Visual Named Entity: Walmart

Knowledge Text: Walmart Inc. is an American multinational retail corporation that operates a chain of hypermarkets (also called supercenters), discount department stores, and grocery stores from the United States, headquartered in Bentonville, Arkansas.

Figure 3.3: A selection of examples from COFAR along with the ground truth visual named entities present in the images and the associated knowledge texts extracted from their respective Wikipedia articles.

seamlessly integrating it with visual content. The second contribution is a comprehensive evaluation demonstrating that KRAMT, in addition to visual reasoning, is capable of commonsense and factual reasoning, and achieves state-of-the-art performance on the COFAR benchmark against competitive baselines.

The remainder of this chapter is organized as follows. In Section 3.2, we review the related literature on image retrieval, commonsense reasoning, and knowledge-grounded vision-language methods. Section 3.3 introduces our proposed framework, namely KRAMT, and explains its major components. The experimental setup, evaluation protocols, and results are presented in Section 3.4, along with comparisons to strong baselines and ablation studies. Finally, Section 3.5 concludes the chapter with key takeaways, and Section 3.6 discusses ethical considerations.

Query: A person taking home groceries after shopping at a supermarket

Query: A grey car waiting to refuel at a gas station

Query: Celebration at a prehistoric monument known for a ring of standing stones

Query: A crowd of people posing for pictures near a tower famously known for its unstable foundation

Query: The 44th President of the United States of America celebrating his birthday

Query: Kids learning to play the game of chess from a former World Champion

Figure 3.4: Additional selection of examples from COFAR

3.2 RELATED WORK

3.2.1 IMAGE SEARCH BY VISIO-LINGUAL ALIGNMENT

The performance of image search using natural language query has been significantly improved in the last few years. Typically, the methods in this space learn the semantic visio-lingual (V-L) alignment; during retrieval, rank the images according to the learned similarity function. Early works [52, 220] learn to project image representations and text embeddings into a joint space. Recently, multimodal transformers have become a de facto model for V-L tasks. Their different avatars [137, 261] tackle multiple V-L tasks jointly by using multi-headed self-attention to encode word tokens and visual objects and are the current state of the art for text-to-image retrieval. However, these methods focus only on the visual cues to represent images and do not encode any external knowledge in their framework. Consequently, any explicit crucial information associated with the image is also ignored.

3.2.2 COMMONSENSE AND FACTUAL REASONING

Bringing commonsense in vision and language tasks is one of the exciting areas of research. The works in this area primarily address: (i) tasks where commonsense reasoning is purely visio-lingual data-driven [160, 242, 254, 257] and (ii) tasks where commonsense is enabled by associating the images with external knowledge [146, 185, 186, 195, 222, 234]. Our proposed task falls in the latter category. However, it is distinctly different from others as none of these works address *image search* requiring detailed visual, commonsense as well as factual reasoning *associated to a diverse set of named entities appearing in the image* including business brands, celebrities, and landmarks. Concerning using named visual entities and associated factual reasoning, the only works closest to ours are [186, 195]. However, compared to ours, these works restrict themselves to only celebrities or business brands and have weaker annotations for visual and commonsense reasoning. Despite its importance and many real-world applications on the Web such as news-search, named visual entity linking and its utility towards downstream tasks have been under-explored in the literature. We aim to fill this gap.

3.3 KNOWLEDGE RETRIEVAL-AUGMENTED MULTIMODAL TRANSFORMER (KRAMT)

Given a natural language query and a large gallery of images each containing a visual named entity, our goal is to retrieve relevant images. To this end, we present Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT) – an unified framework that contains two major modules: (i) visual entity and query-aware knowledge retrieval and (ii) knowledge-infused multimodal transformer as illustrated in Figure 3.5.

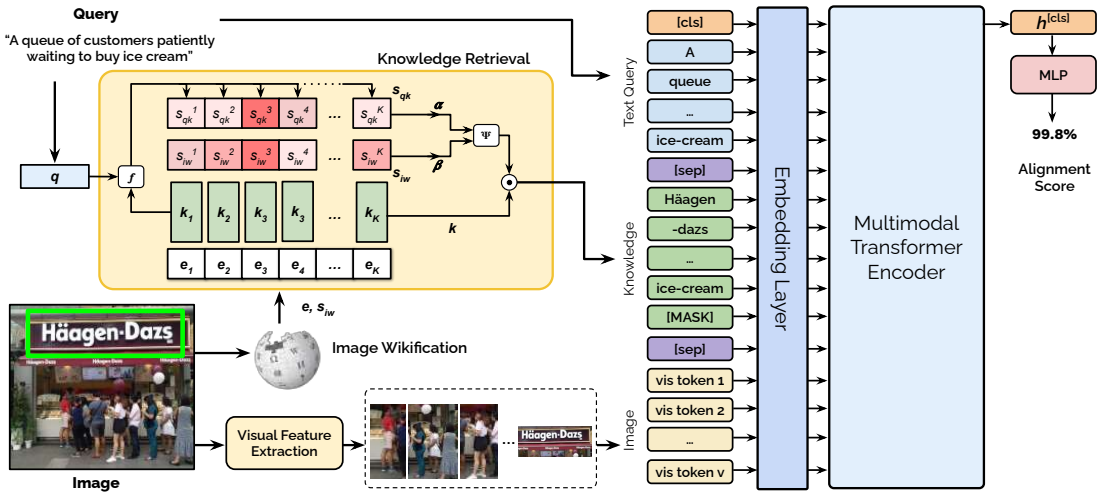


Figure 3.5: Overview of proposed Knowledge Retrieval Augmented Multimodal Transformer (KRAMT): Given a query and a ranked list of visual entities identified in the image, KRAMT grounds the relevant knowledge. This grounded knowledge, along with visual objects and natural query, is fed to a multimodal transformer that learns to align query and relevant image. Please refer Section 3.3 for more details. ***[Best viewed in color]***.

3.3.1 VISUAL ENTITY AND QUERY-AWARE KNOWLEDGE RETRIEVAL:

We posit that visual entities appearing in the image act as a gateway to the encyclopedic knowledge, and its integration to an image retrieval system has the potential to bring commonsense and factual reasoning ability. Therefore, to associate visual entities appearing in the given image to their corresponding Wikipedia page, we perform *visual entity linking* or Image Wikification which is an analogous task to Wikification [191] of text corpora, i.e. linking entity mentions in text documents to their corresponding Wikipedia page. More formally, given an image, a set of m candidate entities $\mathcal{E} = \{e_1, e_2, \cdots, e_m\}$ containing business brands, celebrities, and world landmarks, and associated knowledge text (obtained from Wikipedia articles of these entities) $\mathcal{K} = \{k_1, k_2, \cdots, k_m\}$; Image Wikification aims to rank these entities with respect to their image wikification likelihood (s_{iw}). Here, for an image, s^u_{iw} denotes likelihood of uth entity in that image. We obtain these likelihood scores by using off-the-shelf approaches such as CRAFT+CRNN [9, 189] for detecting and recognizing business brand mentions in the image, VGG face [161] for comparing celebrity faces appearing in the images against a set of reference faces, and landmark recognition [230] for recognizing world landmarks. The image wikification process is illustrated in Figure 3.6.

If we link images to only that entity which corresponds to the highest likelihood score, linking may be incorrect (especially due to look-alike faces or similar world landmarks or noisy text recognition). This is also evident from the experiment, which clearly shows the gap between top-1 and top-K performance of visual entity linking (Refer to Table 3.1). To resolve any error in visual entity linking and subsequently retrieving relevant knowledge, we further leverage the natural language query. To this end, we compute the similarity between query and knowledge text associated with top-K entities using a trainable BERT model f and denote these similarity scores

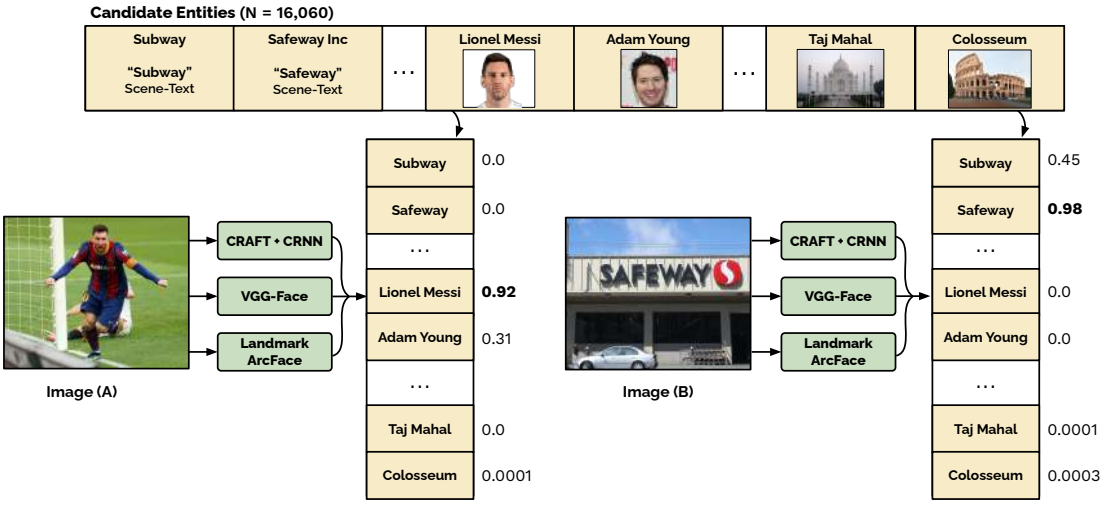


Figure 3.6: Overview of Image Wikification (visual entity linking) method in KRAMT. To recognize named visual entities in images, we use available methods such as CRAFT+CRNN, VGG-Face, and Landmark ArcFace for brands, celebrities, and landmarks respectively. Using these experts, we measure similarity against several thousands of reference entities to obtain a set of high ranking candidates. This open-set recognition approaches allow for addition or removal of any number of reference entities without a need to re-train.

COFAR Category	Top 1 (%)	Top 5 (%)
Brand	60.8	79.6
Landmark	63.5	70.2
Celeb	80.1	83.0

Table 3.1: Results of Image Wikification (visual entity linking) on different categories of COFAR test data.

as s_{qk} where s^u_{qk} denotes the similarity between query and knowledge text corresponding to uth entity. Further, relevance of each entity with respect to image and given query is computed as follows: $s = \Psi(\alpha s_{iw} + \beta s_{qk})$, here Ψ is argmax. The choice of argmax over softmax is intuitive as only one knowledge text is relevant for a given query and image in our task. Once we obtain s, we perform element-wise multiplication to $\mathcal{K} = \{k_1, k_2 \cdots k_K\}$ and feed this knowledge to a multimodal transfer as described next. This query-aware knowledge retrieval process is illustrated in Figure 3.7.

3.3.2 KNOWLEDGE-INFUSED MULTIMODAL TRANSFORMER:

Once we obtain relevant knowledge from our knowledge retrieval module, we use Knowledge-infused Multimodal Transformer – a simple and effective architecture to learn alignment between natural language search queries and images along with their associated external knowledge. KRAMT seamlessly integrates these three input modalities in a unified end-to-end trainable architecture. To achieve this, we first encode the query text, knowledge text, and visual regions as three sequences of features. We then project these features to a shared embedding space before using them as input to the KRAMT. These features then attend to each other through multiple

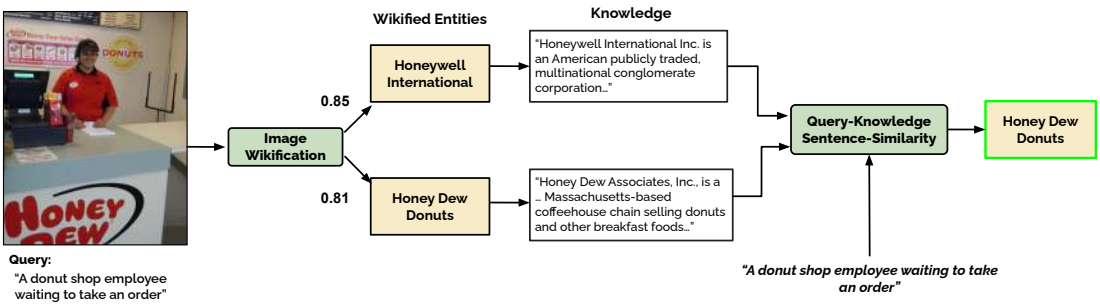


Figure 3.7: Using query-based guidance in knowledge-retrieval for KRAMT. Taking the set of top-ranked candidate entities, we use the search query to select the most appropriate entity by measuring sentence-similarity between the query and entity's knowledge text.

self-attention layers [213]. The output of a special class token from the final layer's output is then used to predict the alignment between the query and image along with its knowledge text.

3.3.3 PRETRAINING:

We learn a strong vision-language grounding capability in KRAMT through pretraining on MS-COCO [129] with the objective tasks of masked language modelling (MLM) and image text matching (ITM).

3.3.4 QUERY AND KNOWLEDGE ENCODER:

We fine-tune pretrained BERT [46] to encode the text of the query and external knowledge. For a given search query Q containing L words and a given knowledge k_i containing M words, we embed them into sequences of d-dimensional BERT feature vectors $\{q_l\}_{l=1}^{L}$ and $\{k_{ij}\}_{j=1}^{M}$ respectively.

3.3.5 IMAGE ENCODER:

Given an image, we detect a fixed set of N visual objects using Faster R-CNN [179] pretrained on Visual Genome [110]. Each image I is represented as an unordered sequence of the N object proposals $\{R_i\}_{i=1}^{N}$ where each R_i is represented as (R_i^{cnn}, R_i^{bbox}), which denote *2048*-dimensional region feature and *4*-dimensional spatial feature, respectively.

We project regional feature R_i^{cnn} and spatial feature R_i^{bbox} into the same d-dimensional space as the search query and the knowledge text using two different learnable transformation matrices $\mathbf{W}_{cnn}$ and $\mathbf{W}_{bbox}$. We apply layer normalization $L(\cdot)$ [8] to each transformed feature, and add them

to get the final visual object feature F_{R_i}.

$$F_{R_i} = L(\mathbf{W}_{cnn} R_i^{cnn}) + L(\mathbf{W}_{bbox} R_i^{bbox}). \tag{3.1}$$

3.3.6 QUERY-IMAGE ALIGNMENT LEARNING:

Besides learning d-dimensional embeddings for the three inputs, we also learn it for three special tokens, namely [SEP] to separate the input modalities, [CLS] to calculate the final alignment score and [$MASK$] to replace the text tokens during MLM. We then allow all the $L + M + N + 3$ input token features to attend to each other through T transformer encoder layers to obtain a joint representation.

As the final step, a multi-layer perceptron that takes d-dimensional $[CLS]$ output feature and produces an alignment score $Out^{[CLS]}$ indicating if the given pair of a search query and the image with associated knowledge are aligned or not, is used. During training, we create positive pairs by selecting images and their corresponding queries from the dataset and negative pairs by randomly changing either the image or the query of the selected pair with another random choice in the dataset. We train the model using binary classification loss. Further, to make the image-query alignment robust, we also train the model with the MLM objective wherein each iteration of training, we replace text input tokens at random with a special token [$MASK$] with a probability of 0.15 and predict the masked tokens based on the context of image, query, and knowledge. During retrieval, for a given query, we rank all the images in the gallery based on the predicted alignment scores. Further implementation details of KRAMT are provided in the Appendix.

3.4 EXPERIMENTS AND RESULTS

3.4.1 KRAMT PRE-TRAINING

To train our full KRAMT model, we initially pretrain on the COCO captions dataset [129] for the objective task of image-caption alignment and masked language modelling. COCO presents a huge diversity of visual content and serves as a good dataset for improving visual reasoning abilities in KRAMT. Further, the model is finetuned on the trainset of COFAR.

We group image retrieval baseline approaches into three categories: (i) Knowledge-only, (ii) Vision-only, and (iii) Knowledge-aware vision and language (V-L) models to investigate the following questions respectively:

- How much impact does external knowledge have? Can it alone drive performance in COFAR without any visual cues?
- Is there a need for integrating external knowledge in COFAR?
- How do other knowledge-aware baselines perform on COFAR?

Under ***Knowledge-only***, we utilize BERT [46] to perform query-knowledge sentence-matching. In ***VL models***, we use modern text-to-image retrieval methods, namely VSE++ [52], and competitive vision-and-language transformers such as VisualBERT [122], ViLBERT [137], and VinVL [261].

Knowledge-aware VL models: As there are no directly comparable knowledge-aware image-retrieval methods in current literature, we implement a few knowledge-aware visual question answering-based models with appropriate modifications to make them compatible for our task: ***(i) Modified Memory Network:*** Memory networks, and their variations have shown to yield state-of-the-art performance on knowledge-aware VQA benchmarks [186, 200]. We implement this baseline by using top-K knowledge texts. These texts are scored with a query, and the weighted sum of this representation, CNN features of the image, and query representation are passed to a binary classifier that classifies if the image is relevant to the query. ***(ii) KRISP-inspired model:*** KRISP [145] addresses open knowledge-based VQA using implicit and symbolic knowledge stored in a graph data structure. In our setting, we use unstructured knowledge text in place of symbolic knowledge. We model implicit knowledge using MM-BERT, similar to KRISP, and for unstructured text, we use BERT embedding of the knowledge text. The output of these representations along with BERT-based query representation is fed to an MLP for learning alignment. ***(iii) KQIA***: Here, knowledge text, along with queries and images, are encoded using gated recurrent units and CNN, respectively, and are then projected into a common space to learn alignment. All baselines are pretrained on the COCO dataset unless mentioned otherwise.

3.4.2 ABLATIONS:

To evaluate the effect of different components of KRAMT, we present the following ablations: ***KRAMT (w/o Knowledge):*** where knowledge text is omitted, ***KRAMT (w/o vision):*** where only query and retrieved knowledge is used, and ***KRAMT (Oracle)*** that assumes ground-truth knowledge is available to the model.

3.4.3 RESULTS AND DISCUSSIONS

We quantitatively evaluate KRAMT on COFAR and compare it against related approaches in Table 3.2. We report recall (R1, R5 and, R10) and median rank (MdR) averaged over all the test queries. Note that higher values for recall and lower values for median rank are desired. The poor performance of knowledge-only models confirms that image search in COFAR is non-trivial and external knowledge about the entities in images alone is insufficient. Further, we observe that the vision-only models such as VisualBERT, ViLBERT, and VinVL, without access to external knowledge, do reasonably well solely through visual reasoning. However, it falls short to KRAMT. By virtue of its seamless integration of search query, visual content, and unstructured knowledge, KRAMT clearly outperforms other baselines, including other Knowledge-aware V-L baselines. These results show the effectiveness of transformer-based methods in COFAR task. The results of ablations are also reported in Table 3.2. Here, we observe that KRAMT that leverages harvested

Method	COFAR (Unified)				COFAR (Brand)				COFAR (Celeb)				COFAR (Landmark)			
	R1	R5	R10	MdR	R1	R5	R10	MdR	R1	R5	R10	MdR	R1	R5	R10	MdR
1K Gallery																
Knowledge-only																
Sentence similarity	3.1	8.7	19.0	84	2.4	9.3	18.8	68	3.0	8.2	16.9	143	4.2	9.1	19.3	97
Vision-only																
VSE++ [52]	7.4	19.2	23.8	68	6.9	19.5	27.6	60	6.0	25.1	38.5	27	21.8	48.0	59.0	9
VisualBERT [122]	22.7	50.0	62.5	5	24.0	50.9	63.3	5	8.0	29.3	37.3	22	32.4	64.5	70.0	4
ViLBERT [137]	29.8	57.9	71.0	5	28.1	55.4	68.6	4	16.5	34.4	42.0	15	36.0	66.9	74.0	4
VinVL [261]	30.5	62.1	74.3	4	31.2	64.8	75.7	4	18.3	38.9	46.5	10	38.7	68.0	76.3	3
Knowledge-aware V-L Models																
Modified Memory Network	15.2	35.0	50.3	5	14.4	34.9	48.6	18	6.1	26.8	39.4	23	24.5	51.1	60.3	5
KQIA	22.0	52.4	64.5	5	19.9	48.2	57.5	9	10.1	29.2	40.5	19	31.9	57.8	67.0	5
KRISP-inspired model	28.1	53.8	69.0	4	26.8	51.5	67.6	5	13.6	32.5	39.8	17	34.3	65.9	74.2	3
Ours																
KRAMT (w/o Vision)	1.9	6.6	12.6	57	1.1	7.4	12.4	35	2.6	6.6	17.1	164	2.7	10.9	14.5	100
KRAMT (w/o Knowledge)	19.8	39.1	49.8	14	19.4	38.3	49	15	11.8	26.3	35.5	25	35.5	67.3	74.5	2
KRAMT	***31.6***	***64.4***	***76.2***	***3***	***32.9***	***66.5***	***78.6***	***3***	***19.7***	***44.7***	***51.3***	***8***	***40.0***	***69.1***	***80.0***	***2***
KRAMT (Oracle)	40.0	73.2	84.5	2	38.5	72.0	83.3	2	26.3	48.7	61.8	6	42.7	76.4	87.3	2
5K Gallery																
Vision-only																
VSE++ [52]	4.7	11.2	18.0	119	3.9	9.2	17.4	128	2.9	9.1	12.5	274	8.8	20.4	33.6	49
VisualBERT [122]	11.4	28.6	40.0	19	11.1	28.0	38.8	20	6.7	13.3	20.0	95	13.6	31.0	40.1	18
ViLBERT [137]	13.6	31.7	43.5	12	13.0	30.8	41.5	10	9.1	15.8	25.0	67	12.2	43.6	54.0	8
VinVL [261]	15.9	35.6	49.2	10	14.9	33.6	44.5	9	11.2	17.7	30.4	31	14.2	44.9	58.0	6
Knowledge-aware V-L Models																
Modified Memory Network	7.3	21.8	34.6	40	6.8	19.9	30.1	46	3.8	10.1	14.6	143	9.3	26.8	37.9	38
KQIA	9.8	25.3	36.2	21	9.1	24.9	35.4	24	7.7	14.9	20.8	79	10.8	28.1	37.4	28
KRISP-inspired model	14.1	36.6	45.9	10	13.3	32.4	43.7	10	8.8	14.1	23.9	61	12.0	41.4	53.7	7
Ours																
KRAMT	***17.1***	***42.9***	***57.2***	***8***	***16.7***	***42.2***	***56.5***	***8***	***11.8***	***18.4***	***34.2***	***28***	***12.7***	***45.5***	***58.2***	***6***
KRAMT (Oracle)	18.9	45.8	59.9	8	18.5	45.0	58.9	7	15.8	25	38.2	18	18.2	52.7	65.5	5

Table 3.2: Comparison of retrieval performance on COFAR (with 1K and 5K gallery each) with baselines and ablations. We report mean recall (R) at top 1, 5, and, 10 retrievals and median rank (MdR) over all the test queries.

knowledge for enabling commonsense and factual reasoning is significantly superior to KRAMT (w/o knowledge).

3.4.4 MODELS PRETRAINED ON LARGE-SCALE DATASETS

We note it may not be fair to compare our model with those which use very-large-scale datasets for pretraining due to significant differences in size of training data. Moreover, there is possibility of overlap of images in their train sets and COFAR-test set; for the sake of a comprehensive comparison, we compare KRAMT with two modern transformer-based models namely CLIP [169] and 12-in-1 [138] in Table 3.3. Please note that they use 400M and 6.3M images, respectively, for pretraining as compared to 125K images (COCO) in our model. We see KRAMT surpasses CLIP and 12-in-1 despite being a smaller model.

We show a selection of visual results for top-3 retrievals for two queries in Figure 3.8. The retrieved images by KRAMT (w/o knowledge) may contain the relevant image, but often ranked lower due to their inability to recognize the entities and perform factual reasoning. On the con-

Method	# of Pre-train Images	COFAR-1K R1	R5	R10	MdR
CLIP [169]	400M	26.4	58.1	72.8	6
12-in-1 [138]	6.3M	30.2	59.9	74.3	4
KRAMT	125K	***31.6***	***64.4***	***76.2***	***3***

Table 3.3: Results using external knowledge over very large-scale pretraining on COFAR 1K.

Figure 3.8: Top-3 retrieved images using proposed KRAMT(w/o Knowledge) and KRAMT on COFAR-1K for two queries. We see that models without access to external knowledge often fail to interpret commonsense such as a financial transaction or protest, and factual information, such as the world's most visited museum, present in the query. On the contrary, KRAMT retrieves semantically more coherent images. Here green colored bounding box indicates the ground truth image.

trary, the proposed KRAMT consistently retrieves relevant images, confirming our hypothesis.

3.4.5 KRAMT IMPLEMENTATION DETAILS

We implement the code in PyTorch [162]. The transformer layers of KRAMT are implemented using Hugging Face's transformers library [231]. We use three transformer encoder layers, with 8 attention heads. The hidden dimension of each block of the transformer layer, as well as the input token feature dimension, is the same as the standard BERT [46] model's hidden dimension of 768.

To encode the query, we use pretrained BERT ('bert-base-uncased') provided by Hugging Face. We keep the sequence length of query text to 40, by truncating the longer sequences and padding the shorter ones. To encode knowledge text, we use the same pretrained BERT, however, this time we keep the sequence length to 80 to accommodate the Wikipedia summary of a page (typically at most 70 words long). This BERT is further fine-tuned during the training of KRAMT with 0.1 times smaller learning rate than that of the KRAMT layers.

To encode images, we extract visual objects using Faster R-CNN [179] pretrained on Visual Genome [110]. We use top-50 most confident visual object proposals for each image, and represent the visual object's appearance features using Faster R-CNN's 'fc6' features of 2048 dimensions. For spatial features, we use 4-dimensional normalized bounding box representation as mentioned

Type	Number of Named Entities	Avg. Length of Knowledge (words)	Avg. Length of Queries (words)	Number of Countries	Number of Entity types
Brand	1060	44.2	11.7	79	39
Celeb	2000	39.0	14.0	92	150
Landmark	2000	41.7	13.6	40	463

Table 3.4: Statistics about the three categories of data in COFAR.

	COFAR-1K (Unseen entities)				COFAR-1K (Seen entities)			
Method	R1	R5	R10	MdR	R1	R5	R10	MdR
KRAMT	31.6	64.4	76.2	3	***35.1***	***72.6***	***88.6***	***3***

Table 3.5: Performance of KRAMT on two COFAR-1K versions comprising of entities previously unseen during training and entities seen during training. We observe that performance of KRAMT is higher for already-seen entities.

in our approach 3.3. To represent special tokens $[CLS]$ and $[SEP]$ we learn 768-dimensional embedding for each of them during training.

To get alignment scores from the output embedding of the $[CLS]$ token, we learn a multi-layer-perceptron (MLP) with one hidden layer of size 512 and a ReLU activation. For pretraining on COCO, the knowledge text input is masked and trained for 42 epochs using Adam [107] optimizer, with a constant learning rate of 1e-4. Before we finetune KRAMT on COFAR for the task of query-image alignment, we finetune KRAMT on text of COFAR with just masked language modelling objective for 10 epochs using Adam [107] optimizer, with a constant learning rate of 5e-5. Finally, we finetune KRAMT on COFAR with the task of query-image alignment for 15 epochs using Adam [107] optimizer, with a constant learning rate of 0.00002. The model is trained with the binary cross-entropy loss for query-image alignment task, and cross-entropy loss over vocabulary for masked language modelling task. The model was trained using two Nvidia RTX 5000 GPUs (each having 16GB of GPU memory) with a batch size of 64 while training and 128 while testing. KRAMT pretraining takes approximately four days on the two GPUs, whereas KRAMT finetuning on COFAR takes lesser time. Further details of the implementation can be found in the code which we provide in the project page.

3.4.6 LIMITATIONS AND FUTURE SCOPE

We observe the following limitations of our work: (i) for the introduction of COFAR, we have chosen natural scenes that contain only one visual named entity. This may not be the case in a real-world setting, (ii) restricted by the budget, current version of COFAR contains only 25K images of 5K named entities in all. However, in an open-set scenario, a much larger and diverse set of visual named entities can be considered, and Image Wikification can be a promising research challenge. In fact a contemporary work [267] poses this as a stand-alone task, and (iii) explicit external knowledge associated with common objects has not been leveraged. We leave addressing

these limitations as a future work of this work.

3.5 CONCLUSION

In Information Retrieval and NLP community, knowledge bases are instrumental in enabling commonsense and semantic search. However, their utility in semantic image search has not been extensively explored in the literature. We have drawn the attention of the vision and language community towards this issue through our work and presented a novel multimodal transformer namely KRAMT which seamlessly combines image, query, and knowledge encoding to learn alignment between the image with associated knowledge and query. We firmly believe that image search requiring commonsense and factual reasoning and the new dataset viz. COFAR introduced in this work will open up several future research avenues.

3.6 ETHICAL CONSIDERATIONS

One caveat of COFAR is that the images have been collected from various publicly available sources that may contain geographical bias inherently present in them that were undetected in this work. This problem is common with many public vision benchmarks. A more rigorous inspection is indeed required before deploying the proposed model for real-world applications.

4

Augmenting VLMs with Textual Knowledge for Visual Question Answering

In this chapter, we focus on augmenting vision–language models with textual knowledge for visual question answering, with an emphasis on the Text-KVQA task. Building on Chapter 3, where we introduced a vanilla visual entity linker, we extend this idea to jointly exploit both textual and visual cues for more accurate disambiguation. Specifically, we propose VisTEL (Visual Text Entity Linker), which leverages OCR-extracted scene text together with surrounding visual context to link entities to the correct knowledge base entries. Further, we propose KaLMA (Knowledge-aware Large Multimodal Assistant), which integrates the retrieved factual knowledge into large multimodal models to generate answers that are both accurate and interpretable. A key strength of KaLMA is its ability to output supporting facts alongside answers, making the reasoning process more transparent. Through extensive experiments on Text-KVQA, we show that the VisTEL-KaLMA pipeline achieves a new state-of-the-art, surpassing prior knowledge-aware approaches and competitive large vision–language models by an absolute margin of 23.3%.

4.1 INTRODUCTION

In the past few years, the research community has shown significant interest in visual question answering based on text appearing in images, as evidenced by the emergence of OCR-VQA [150], ST-VQA [20] and TextVQA [196]. Giving another aspect to these problems by leveraging external knowledge for text-based visual question answering, [195] introduced a task called Text-KVQA. The Text-KVQA presents a unique challenge: given an image containing textual entities like business brands, book titles, or movie titles, the task is to answer questions that require external knowledge about these entities. Addressing Text-KVQA involves detecting text in images, recognizing it, linking it to a knowledge base, and employing visual context and knowledge base for reasoning to provide an answer. Since the introduction of this problem, several advancements

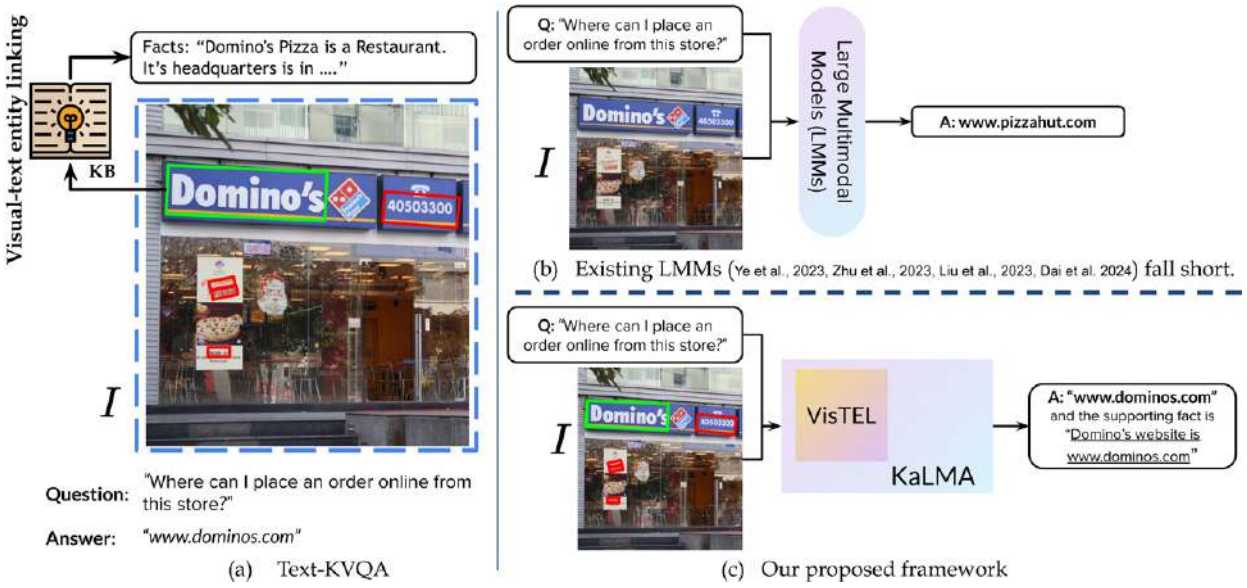


Figure 4.1: (a) Text-KVQA [195]: Given an image containing a named entity as visual text, e.g., "Domino's" in this illustration, the aim is to answer the question by leveraging explicit knowledge about the visual text entity. (b) Large Multimodal Models are one obvious choice for solving such tasks today. However, they alone are insufficient as they hallucinate on visual objects. (c) We propose a novel approach – KaLMA that augments an LMM with specialized visual text recognition and retrieved relevant knowledge obtained using visual text entity linking by proposed VisTEL. Our approach establishes a new state-of-the-art for this task.

have happened in visual text understanding as well as vision and language models. In this work, we revisit Text-KVQA by leveraging these modern advancements and propose a framework that judiciously integrates various components of contemporary architecture.

The emergence of large multimodal models (LMMs)[1] represents a significant trend in the literature on vision and language [37, 133, 158, 166, 211, 252, 263, 271]. Over the past few years, many large-scale language and vision models have been developed, demonstrating exceptional performance across various tasks including, but not limited to, image captioning, visual question answering, multimodal reasoning, and visual grounding. We believe that pretrained LMMs hold great potential for addressing Text-KVQA. These models are rich in the implicit knowledge learned by large-scale pretraining. However, despite their numerous advantages, they are not without drawbacks, notably hallucinations. This challenge becomes particularly apparent in Text-KVQA, where precise reasoning about entities depicted in images and associated knowledge is required. Consider the following scenario where a customer, after finishing their meal at a restaurant store, takes a picture of the store signboard and enquires about a possible future online delivery, asking, 'Where can I place an online order from this store?' (Figure 4.1(a)). Existing LMMs often hallucinate over the pizza present in the image and points to the website of 'Pizza Hut' instead of 'Domino's' (Figure 4.1(b)); whereas complementing the LMM with an explicit visual text entity linking followed by knowledge-retrieval helps overcome hallucination (Figure 4.1(c)), thereby generating an accurate answer to the given question. Our model is developed on this

[1]We refer to both large multimodal model and large vision and language models as LMM in this work.

hypothesis.

We address Text-KVQA by introducing an architecture, namely KaLMA – knowledge-aware large multimodal assistant that first invokes our proposed visual text entity linker or VisTEL – an LMM-architecture that links visual text entities to the associated knowledge base (Illustrated in Figure 4.1 (c)). Once the entities are linked to the knowledge base, the associated knowledge is retrieved and augmented to a large multimodal model to answer visual questions.

The first contribution of this work is a revisit of Text-KVQA [195] in the light of recent advancements in large multimodal models (LMMs). To this end, we benchmark state-of-the-art LMMs on Text-KVQA and show that, although powerful, these models often ignore visual text present in the images, resulting in hallucinations. The second contribution is a principled approach, VisTEL, for linking visual text entities in images to a knowledge base. VisTEL is an LMM-based architecture that leverages the surrounding OCR-extracted texts obtained using a specialized recognition module along with the visual context within the image to perform highly accurate entity linking for visual text entities. The third contribution is KaLMA, a Knowledge-aware Large Multimodal Assistant, which enhances an existing LMM, specifically LLaVA [133], by integrating retrieved knowledge from our proposed VisTEL. This augmentation facilitates robust vision and language reasoning, thereby enabling superior knowledge-aware text-based visual question answering. The fourth contribution is extensive experimentation and ablation to demonstrate the superior performance of our proposed framework over competitive approaches and state of the art. We also provide insights into the design choices, the attribution ability of KaLMA, and its effectiveness in addressing hallucination issues of LMMs. Our framework advances state of the art on Text-KVQA by 18.2% on scene images, 19.6% on book covers, and 32.2% on movie poster splits of the dataset on an absolute scale.

The remainder of this chapter is organized as follows. In Section 4.2, we review related literature on knowledge-aware VQA tasks, large multimodal models, and visual entity linking. In Section 4.3, we present our methodology in detail: we first introduce VisTEL, a visual text entity linker (Section 4.3.1), and then describe KaLMA, our knowledge-aware large multimodal assistant (Section 4.3.2). The experimental setup, datasets, evaluation protocols, and results are discussed in Section 4.4, along with comparisons, ablations, and qualitative analyses. We conclude our study in Section 4.5, followed by a discussion of limitations in Section 4.6, and ethical considerations and broader impact in Section 4.7.

4.2 RELATED WORK

KVQA Tasks: Visual Question Answering is a well-studied task [7, 62]. This task has been extended to scenarios that require the ability to read text within images, leading to the development benchmarks such as st-vqa [19, 20], textvqa [196], docvqa [148], and ocr-vqa [150]. While these benchmarks were successful in their intent of integrating reading and reasoning abilities in VQA, they are often restricted to reasoning around what is visually apparent. To address this

gap and encourage models to perform reasoning beyond visually apparent facts, [195] introduced knowledge-aware Text-based VQA task. Distinctively different from other knowledge-aware visual question answering tasks such as kb-vqa [223], fvqa [222], kvqa [186], ok-vqa [146], and Infoseek [30], Text-KVQA deals with reasoning over visual text entities and associated knowledge to arrive at answer.

Methods Prior to Large Multimodal Models: Early methods to solve knowledge-aware VQA tasks focus on leveraging knowledge in the form of triplets [152, 153, 234], or sub-knowledge-graph [195, 265] or memory facts [229]. Later, transformer architectures [213] owing to their ability to encode intrinsic knowledge using large-scale pretraining, have become defacto for addressing kvqa.

Inspired by the hybrid models, e.g. [67, 116] where intrinsic knowledge of transformer architectures is complemented with explicit external knowledge; researchers proposed hybrid methods such as conceptbert [57], krisp [145], and reveal [84] which augment the multimodal transformers with explicitly retrieved external knowledge.

Emergence of Large Multimodal Models: The early success of large-scale pretraining on the downstream tasks demonstrated by the foundation models, e.g., bert [46] and gpt [171] paved the way for the researchers to scale the model and the data used for pretraining. gpt-3 [21] is an early large language model (llm) demonstrating reliable performance on many downstream tasks. Following this, several llm variants [37, 166, 211, 232, 263] have been introduced. Researchers adopted these llms to vision-language research, with the key idea being aligning the visual information with the linguistic information of the llms to come up with large multimodal models (LMMs) [120, 133, 166, 212, 252, 271]. Recently, LMMs have become first-hand solutions for many downstream vision-language tasks, making them an obvious choice to solve Text-KVQA. Authors in [102, 250] prompt the llms with visual information via dense captions, object tags, object-level bounding box coordinates, and OCR tags. These methods rely heavily on the implicit knowledge learned by these llms. Further, kat [65] improves upon such methods by augmenting external knowledge via retriever before prompting the llm. However, it ignores the explicit visual information, which revive [130] aims to fix. Although these methods show significant success, they have limitations such as hallucination and ignoring visual texts for reasoning. We aim to fill these gaps by proposing a novel solution for Text-KVQA.

Visual Entity Linking: Entity linking has traditionally been a well-established focus area within the NLP community [95]. In contrast, the problem of visual entity linking has only garnered attention in the last decade [82, 186, 203]. [203] have proposed a novel dataset and benchmark for visual named entity linking. [186] drew attention to the need for visual entity linking for addressing knowledge-based visual question answering. Open-domain Visual Entity Recognition has also been studied in the literature [23, 82, 239]. However, most of these works have focused on linking entities such as persons, landmarks, and other named entities, while neglecting visual text such as business brand names and movie or book titles. In this work, we address this gap by proposing a principled solution for visual text entity linking and demonstrate its utility as a precursor to Text-KVQA.

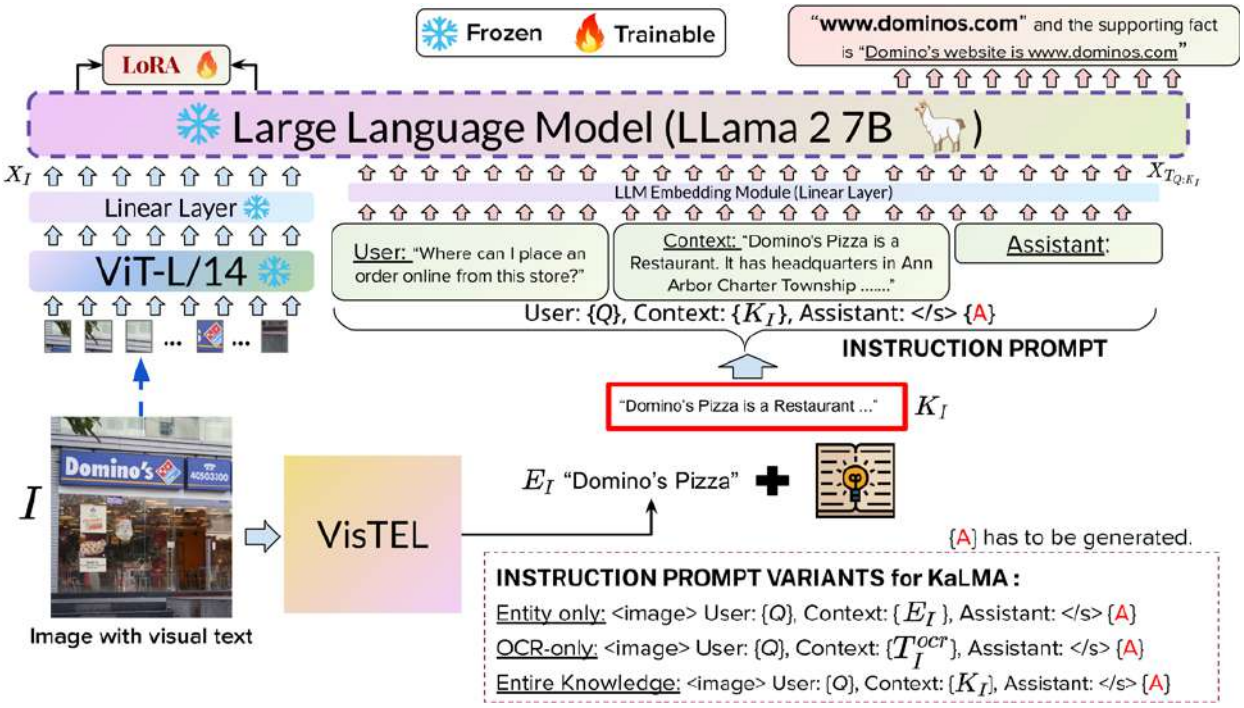


Figure 4.2: Overview of our proposed framework KaLMA. We first link the visual text in the image I to the entity E_I using VisTEL (Section 4.3.1) and its associated knowledge K_I is fetched. Then, we frame an instruction prompt with the question Q and the knowledge K_I, and encode it using the $LMM_{embedding}$ module f to obtain textual features $X_{T_{Q:K_I}}$. We encode the image I using a vision encoder to obtain visual features X_I. Then, we concatenate X_I and $X_{T_{Q:K_I}}$ and feed them to the LMM to generate an accurate answer A to the question Q. Instruction prompt templates used in our ablation study are shown in the bottom right box, where T_I^{ocr} is the visual text of the image I.

4.3 METHODOLOGY

Problem Statement: Text-KVQA [195] is a knowledge-intensive visual question-answering task that requires a system to read and interpret the visual text in an image and leverage it as a gateway to access and reason over external knowledge to answer the question. The external knowledge base $\mathcal{K}$ consists of a set of n entities $\mathcal{E} = \{E_1, E_2, ..., E_n\}$ and their corresponding knowledge $\mathcal{K} = \{K_1, K_2, ..., K_n\}$, where each K_i is a set of facts. For example, *Domino's Pizza* is an entity whose associated knowledge facts, obtained in the form of triplets from Wikidata, are concatenated to form simple sentences such as *"Domino's Pizza is a restaurant", "Its headquarters are in Ann Arbor Charter Township", "It belongs to the fast food industry", and so on*. In this section, we describe our approach, whose overall architecture is illustrated in Figure 4.2. Our approach first links visual text entities using the proposed VisTEL module and retrieves relevant knowledge to the entity (Section 4.3.1), it then reasons over the image and the retrieved knowledge to answer the question (Section 4.3.2).

Figure 4.3: Challenges associated with Visual Text Entity Linking: (a) Visual text entity may appear as abbreviation instead of the entity name directly, e.g. "RBS" instead of "The Royal Bank of Scotland", (b) Visual text with varying font and stylized orientation pose a challenge to the recognizer, (c) Example of homonyms where visual text *HP* may refer to 'Hewlett Packard' (left) or 'Hindustan Petroleum' (right).

4.3.1 VISTEL: VISUAL TEXT ENTITY LINKER

Entity linking is a well-studied task [95], where given a sentence, the named entities need to be identified and linked with their corresponding entities in a knowledge base. In this work, we study an analogous task, where the input is no longer a sentence, but instead an image containing visual text entities and the task is to link them to a corresponding external knowledge base.

One plausible solution, as shown in [195], is to extract the visual text in these images using visual text recognition engines and then leverage distance-based text similarity methods between the recognized text and the candidate entities for the entity linking task. However, such methods are highly sensitive to the following challenges: (i) Noisy or imperfect OCR may lead to wrong entity linking, and (ii) visual text might contain abbreviations instead of the entity names, e.g. *"RBS"* for the entity *"The Royal Bank of Scotland"*, (iii) The problem of homonymy, e.g. visual text *HP* may refer to *'Hindustan Petroleum'* or *'Hewlett Packard'*. Furthermore, unlike entity linking which often benefits from larger textual contexts; visual text entity linking has limited textual context, e.g., surrounding visual texts, and often must infer correct entities based on visual context. Please refer to Figure 4.3 for a selection of challenges associated with visual text entity linking. The other plausible solution is to use large multimodal models (LMMs). By virtue of large-scale pretraining, they have strong abilities to reason and infer correct entities based on visual cues.

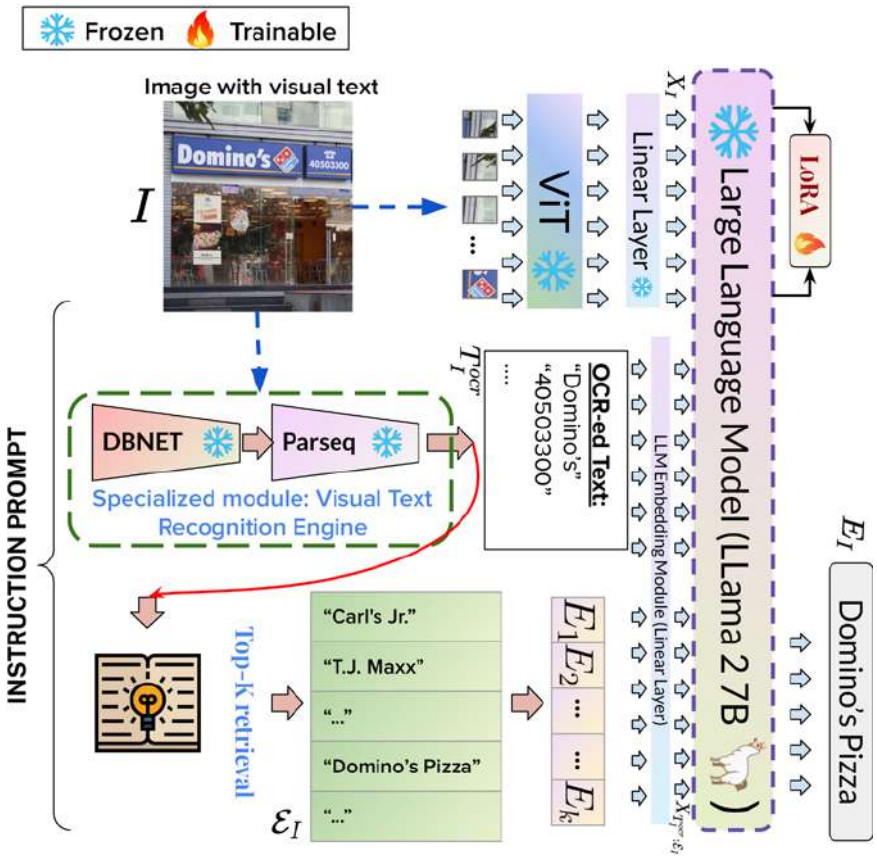


Figure 4.4: Illustration of VisTEL. We extract visual text from the given image using visual text recognition engine and, based on textual similarity, obtain k candidate entities from the knowledge base. We fit OCRed text and the candidate entities into an instruction prompt template and encode the image using a visual encoder and the text prompt using an LMM embedding module to obtain X_I and X_T, respectively. Once encoded, LMM generates the entity associated with the visual text in the image. Please refer to the Section 4.3.1.

However, we observe that feeding only the image without the surrounding OCRed text often results in hallucinations. To address these shortcomings, we propose Visual Text Entity Linker (VisTEL) that links the visual text present in an input image to its corresponding entity by jointly reasoning on textual context obtained using an explicit specialized visual text recognition engine and visual context obtained using a vision encoder of a large multimodal model. The architecture for VisTEL is illustrated in Figure 4.4.

Visual Text Recognition Engine: Given an image I, we extract text $T_I^{ocr} = \{t_1^{ocr}, t_2^{ocr}, ..., t_r^{ocr}\}_I$ using specialized visual text detection and recognition methods. We, then find a set of k candidate entities $\mathcal{E}_I$ based on the normalized edit-distance (NED) score between the entity name in the knowledge base with T_I^{ocr}. We use state-of-the-art text detection and text recognition approaches, namely dbnet [127] and ParSeq [16], respectively.

Vision encoder: We use the output of the last transformer layer of a pretrained CLIP visual encoder ViT-L/14 [169] as our patched image features $\tilde{X}_I \in \mathbb{R}^{p \times d_v}$, where p and d_v are the number of patches and encoding dimension of ViT, respectively. Further, these image features are projected

to d_{lmm} dimension using a linear layer g to obtain the final sequence of image features $X_I \in \mathbb{R}^{p \times d_{LMM}}$, i.e., $X_I = g(\tilde{X}_I)$.

Large Multimodal Model: Once we obtain the OCR-ed text T_I^{ocr} and candidate entities $\mathcal{E}_I$, we frame the following instruction prompt:

Instruction prompt template for VisTEL

<image>
USER:Given an image. The task is to link the visual text $\{T_I^{ocr}\}$ to one of the following entities: $\{\mathcal{E}_I\}$
ASSISTANT:$\{E_I\}$

Then, we feed the prompt to the embedding module h of the LMM to obtain text tokens $X_{T_I^{ocr}:\mathcal{E}_I} \in \mathbb{R}^{l \times d_{LMM}}$ i.e., $X_{T_I^{ocr}:\mathcal{E}_I} = h(prompt(T_I^{ocr} : \mathcal{E}_I))$, where l and d_{lmm} are the number of text tokens and input embedding dimension for the LMM, respectively. We, then concatenate image features X_I and text features $X_{T_I^{ocr}:\mathcal{E}_I}$, and feed it as an input to the large multimodal model. VisTEL auto-regressively predicts the probability of the next token E_{I_t} in the target entity E_I by attending to the input prompt tokens and the previously generated entity tokens $E_{I<t}$. We train VisTEL by optimizing the language modeling loss for generating the target entity conditioned on the inputs X_I and $X_{T_I^{ocr}:\mathcal{E}_I}$.

4.3.2 KALMA: KNOWLEDGE-AWARE LARGE MULTIMODAL ASSISTANT

We present Knowledge-aware Large Multimodal Assistant (KaLMA) for addressing Text-KVQA. The KaLMA is an effective architecture that seamlessly integrates questions and images in the context of external knowledge in a trainable architecture to generate accurate answers.

We use visual features X_I from the vision encoder. Further, we concatenate question Q and the knowledge K_I via instruction prompt template (as shown in the Figure 4.2) and feed to the embedding module f of the LMM to obtain text tokens $X_{T_{Q:K_I}} \in \mathbb{R}^{m \times d_{lmm}}$ i.e., $X_{T_{Q:K_I}} = f(prompt(Q : K_I))$, where m is the number of text tokens. Then, we concatenate image features X_I, and text features $X_{T_{Q:K_I}}$ and feed to the large multimodal model to generate the accurate answer A. Further, to bring attribution ability, we model KaLMA to generate the supporting fact S that contributed to the answer along with answer generation. From here onwards, we will refer answer and supporting fact together as A. KaLMA predicts the probability of the next token A_{a_t} in the answer A_a in an auto-regressive manner. It does so by attending to the prompt inputs and the previously generated tokens $A_{a<t}$. We train by minimizing the generative language modeling loss $\mathcal{L}_{ans\_gen}(\theta)$, which aims to generate the target tokens based on the inputs X_I and $X_{T_{Q:K_I}}$ (Eq. 4.1). Note that target tokens comprise both the answer and the supporting fact. During training, we leverage the ground truth entity and its corresponding knowledge K_I, while during inference, we obtain it using our VisTEL module. We reuse the weights of VisTEL to initialise KaLMA.

$$\mathcal{L}_{ans\_gen}(\theta) = -\left[\sum_{t=1}^{|A|} \log(P_\theta(A_{a_t}|A_{a<t}, X_I, X_{T_{Q:K_I}}))\right], \tag{4.1}$$

where θ are the trainable parameters, $A_{a<t}$ represents the answer tokens already generated before predicting the token A_{a_t} at the current time step t.

4.4 EXPERIMENTS AND RESULTS

4.4.1 DATASET, METRICS AND COMPARISONS

We conduct our experiments on Text-KVQA [195] dataset[2]. The questions in this dataset span across three splits, namely, scene, book, and movie containing natural scene images, book covers, and movie posters, respectively. These splits have (50K questions, 10K images, 500 entities), (1M questions, 207K images, 207K entities), (222K questions, 34K images, 34K entities), respectively. Further, each of these splits comes with its own knowledge base, namely *KB-business* containing knowledge facts about business brand entities harvested from Wikidata, *KB-book* containing knowledge facts about books harvested from a book catalog, and *KB-movie* containing knowledge facts about movies harvested from IMDB, respectively. For each split, we follow the similar train-test division as [195] where entities in train and test sets are disjoint. We evaluate the methods using an accuracy metric.

Along with traditional VQA baselines, we compare the question answering performance of our proposed approach KaLMA with methods from the following three major categories: (i) ***Pre-LMM Approaches***: here, we choose classical transformer-based baselines, namely, GPT-2 [171] (text-only), GPT-2 (with BLIP-2 [120]-extracted captions as visual context), ViLT [106] and VL-Bart [35]. For an encoder-only model like ViLT, we treat Text-KVQA as a classification-style visual question answering where the task is to predict the answer from a set of all possible answers. (ii) ***LMM-based Approaches:*** restricting ourselves to open-source models, we choose four popular LMMs, namely, mPlug-Owl [252], MiniGPT4v2 [271], LLaVA-1.5 (7B) [133] and InstructBLIP [40] for comparison. Prompts used and other fine-tuning details for these LMMs are discussed in the Appendix. (iii) ***SOTA approaches:*** we also compare against memory network [229] and graph neural network-based approach [195] which are the current state of the art. In addition to these comparisons, we compare the visual text entity linking performance of our proposed VisTEL against recent multimodal retrievers from UniIR [227], specifically CLIP-SF and BLIP-SF, where we use image and visual text to retrieve entities from the knowledge base.

[2]Available at: `https://textkvqa.github.io`

Method	Accuracy on Text-KVQA		
	scene	book	movie
Traditional VQA Baselines			
BiLSTM	17.0	12.4	11.3
BoW+CNN	11.5	8.7	7.0
BLSTM+CNN [7]	19.8	17.3	15.7
HiCoAttenVQA [139]	22.2	20.4	18.4
BAN [105]	23.5	22.3	20.3
Pre-LLM Approaches			
GPT-2 [171]	22.8	22.3	31.8
GPT-2 (w/ Visual Context)	25.4	43.2	38.5
ViLT [106]	38.2	31.1	40.1
VLBart [35]	35.1	38.6	41.5
Previous SOTA			
Memory Network [229]	49.0	57.2	42.0
Singh et al. [195]	54.5	62.7	45.2
LLM-based Approaches			
mPlug-Owl [252]	21.3	26.7	8.2
LLaVA-1.5 [133]	39.2	37.0	46.1
MiniGPT4v2 [271]	48.2	47.7	47.6
InstructBLIP [40]	31.5	30.3	29.9
Ours (KaLMA)			
w/ NED retrieval	54.9	63.4	70.8
w/ VisTEL	***72.7*** (↑ 18.2%)	***82.3*** (↑ 19.6%)	***77.4*** (↑ 32.2%)
Oracle	99.3	92.8	99.4

Table 4.1: Results on T***ext***-KVQA***:*** Various methods on the three data categories of Text-KVQA dataset, namely, scene, book and movie.

4.4.2 IMPLEMENTATION DETAILS

We implemented our method using PyTorch and the Huggingface Transformers library [231]. We used LLaVA-1.5 as our foundation model for both VisTEL and KaLMA models. Note that, LLaVA-1.5 is trained on CC3M [188] and MS-COCO [129]. We have carefully examined these datasets for duplicates and found no overlap with the evaluation set of Text-KVQA. Further, dbnet [127] and ParSeq [16] are used as visual-text detection and visual-text recognition modules in the visual text recognition engine, respectively. We fine-tuned VisTEL with LoRA for 10 epochs with a learning rate of 1e-5 with a batch size of 128. Similarly, we fine-tuned KaLMA with LoRA for 6 epochs with a learning rate of 2e-5 with a batch size of 64. LoRA details are as follows: rank: 16, alpha: 32, dropout: 0.05, for both the models. Our experiments are conducted on a machine with three A6000 GPUs (48 GB each). We make our implementation publicly available at our project website[3].

[3] https://vl2g.github.io/projects/LMM4Text-KVQA/

Method	Visual Context	Textual Context	scene	book	movie
Text-only					
Direct match	✗	✓	54.8	63.6	58.1
NED	✗	✓	57.1	66.5	60.1
Multimodal retrievers					
UniIR (*CLIP-SF*)	✓	✓	64.5	78.8	45.2
UniIR (*BLIP-SF*)	✓	✓	60.6	78.5	50.1
Ours					
VisTEL	✓	✗	73.2	76.9	66.6
VisTEL	✗	✓	31.5	9.8	11.6
VisTEL	✓	✓	***76.5***	***80.6***	***71.6***

Table 4.2: Visual Text Entity Linking Results. We report Recall@1. Text-only retrievers: direct match and normalized edit distance-based methods and Multimodal retrievers: CLIP-SF and BLIP-SF from UniIR [227] fall short. On the contrary, the proposed VisTEL, which leverages both visual and textual context (surrounding OCRed text) in an LMM framework, shows impressive visual text entity linking performance over both text-only as well as multimodal retrievers.

4.4.3 RESULTS ON TEXT-KVQA

We quantitatively evaluate our proposed framework KaLMA on Text-KVQA and compare against relevant methods in Table 4.1. We report accuracy averaged over the entire test set for all the three splits of Text-KVQA. It is no surprise that traditional VQA baselines perform poorly as they do not have the ability to read and reason over visual text. Pre-LMM language models (GPT-2) and vision-language models (GPT-2 w/ visual context, ViLT, VisualBert) along with LMM baselines (mPlug-Owl, LLaVA-1.5, MiniGPT4v2, InstructBLIP) outperform traditional methods, but fail to outperform knowledge-aware methods including the state-of-the-art method [195]. We observe that on knowledge-intensive tasks like Text-KVQA, the OCR-free capabilities acquired by LMMs are due to heavily correlated hallucinations of visual objects, thereby fall short to our proposed approach by a significant margin. Our proposed framework seamlessly integrates knowledge associated with visual text entity (extracted using our proposed VisTEL) and significantly enhances the performance on Text-KVQA. To be specific, we advance the state-of-the-art by 18.2%, 19.6%, and 32.2% on scene, book, and movie splits of Text-KVQA on an absolute scale. This superiority of our approach demonstrates its efficacy in knowledge-aware text-based visual question answering.

4.4.4 VISUAL TEXT ENTITY LINKING RESULTS

We report them in Table 4.2. Here, we observe that the proposed VisTEL clearly outperforms both (i) Text-only retrievers, such as a direct match or normalized edit distance-based match of OCRed text and entity name, and (ii) Multimodal retrievers, CLIP-SF and BLIP-SF from UniIR [227]. By virtue of LMM and joint reasoning of visual and textual (OCR) context for linking visual text, VisTEL yields reasonably advanced performance. Nevertheless, there is still scope of improvement which we believe can be achieved by further improving visual text recognition, and performing detailed visual reasoning such as logo recognition. We leave these extensions as future work.

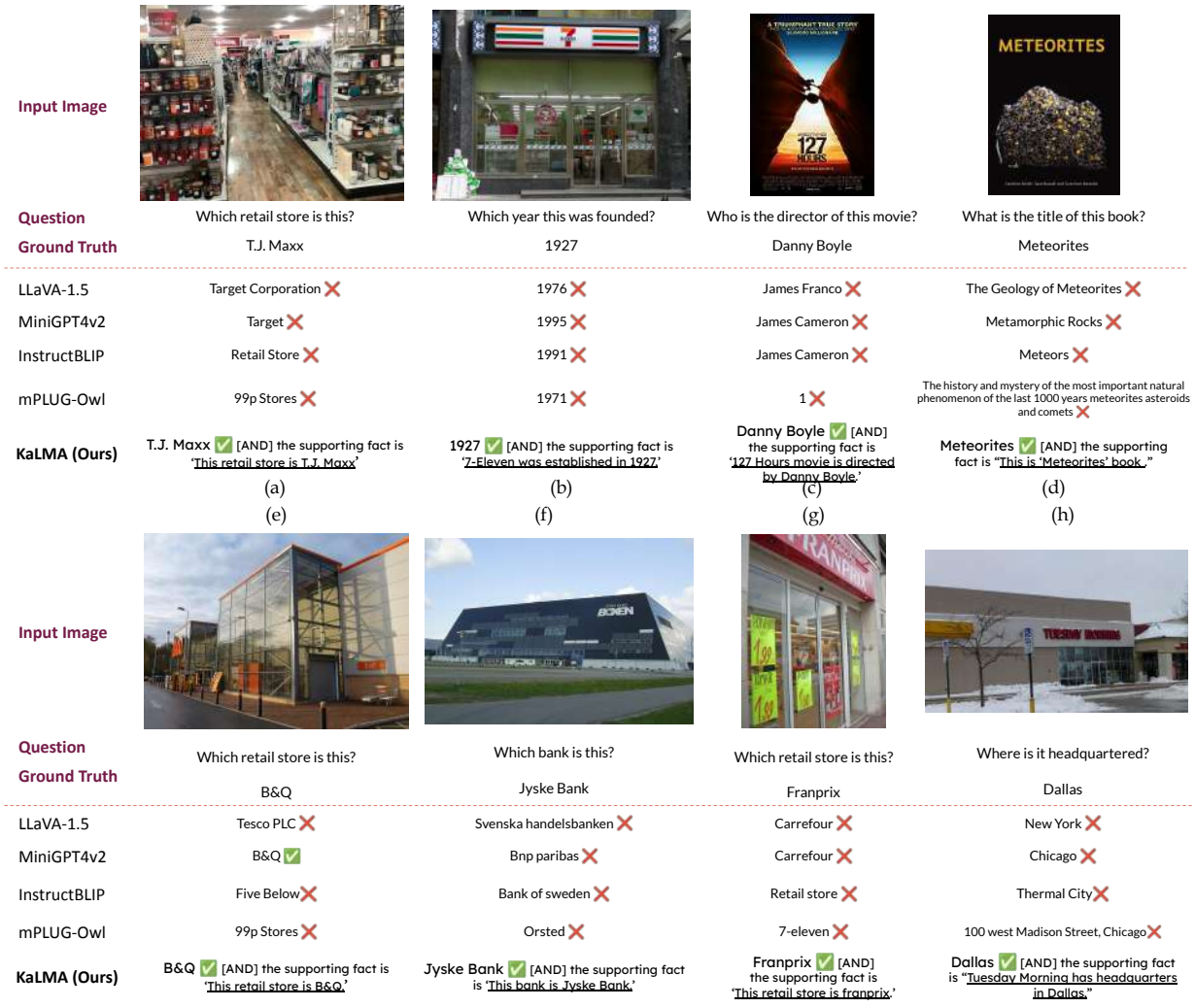


Figure 4.5: A selection of our results as compared to implicit knowledge-based LMM approaches. Please refer Qualitative Results in Section 4.4.3 for observations. More results in Section 4.4.9.

4.4.5 QUALITATIVE RESULTS

We show a selection of results for text-based knowledge-aware visual question answering and visual text entity linking in Figure 4.5 and Figure 4.6, respectively.

In Figure 4.5, LMM models exhibit hallucination over visually apparent objects. In (a), all LMMs incorrectly identify *T.J. Maxx* as popular retail stores *Target* and *99p Stores*. In (b), they provide a random year. In (c), these models are confused over the keyword *James*, mixing up the director and actor names on the poster. In (d), LMMs hallucinate and suggest non-existent book titles. Similar hallucinations can be seen in the other examples (e-h). Our proposed method owing to visual-text entity linking capabilities and reasoning over explicit knowledge, provides accurate answers. In Figure 4.6, we observe that our proposed model accurately links visual text in the images to the correct entity despite noisy OCR in (a), abbreviations in (b), and ambiguous visual text in (c).

Figure 4.6: Comparison of visual text entity linking results. The VisTEL infers the correct entity based on visual context as well as textual context in the form of surrounding text in the image. Please refer to Qualitative Results in Section 4.4.3 for more details.

Model	Visual text EL	Knowledge	scene	book	movie
KaLMA	✗	✗	39.2	37.0	46.1
	✗	OCR only	52.2	49.8	51.7
	✓	Entity name only	53.2	59.1	59.2
	✓(w/o VisTEL)	Knowledge facts	54.9	63.4	70.8
	✓(w/ VisTEL)	Knowledge facts	***72.7***	***82.3***	***77.4***
MiniGPT4v2 (best LMM method)	✗	✗	48.2	47.7	47.6

Table 4.3: Ablations for showing the importance of visual text entity linking, explicit knowledge facts and VisTEL. Also, note that the first-row result corresponds to LLaVA-1.5 result from Table 4.1, as KaLMA without VisTEL and knowledge is equivalent to LLaVA-1.5. Please refer to Section 4.4.6 for more details.

4.4.6 ABLATIONS AND ANALYSIS

We conduct the following ablations and analysis of the proposed work:

(i) What is the need for VisTEL?: To study the performance of our model in the absence of the pro-

Detection	Recognition	scene	book	movie
EAST	CRNN	67.2	81.3	66.4
CRAFT	CRNN	67.7	81.9	75.1
DBNet	ParSeq	***72.7***	***82.3***	***77.4***

Table 4.4: Effect of Different Text Detection and Recognition Approaches in our approach.

posed VisTEL module, we replace it with traditional edit-distance-based entity linking where entities are sorted based on the normalized edit-distance between extracted OCRs and the entity name. The results of this ablation, as shown in Table 4.3 further support our claim that the superior visual text entity linking capabilities of the proposed VisTEL, enhances the downstream performance of KaLMA.

(ii) What is the need for visual text entity linking and explicit knowledge in T***ext***-KVQA***?:*** We show these ablation results in Table 4.3. We, ***first***, skip visual entity linking in the KaLMA, and feed only the extracted OCRed text to KaLMA. The drop in performance shows the utility of visual text entity linking. ***Second***, we perform visual entity linking, but, we feed the visual text linked entity name from the VisTEL as input to KaLMA. Our observations indicate that although entity names give some hints about the associated knowledge and reduce hallucination to some extent, it is not as useful as using explicit knowledge in our full model.

(iii) How much does choice of visual text recognition engine matter?: In this ablation, we replace db-net [127] and parseq [16] used in KaLMA with craft [9], east [270] and crnn [189], and report the results of KaLMA on Text-KVQA in Table 4.4. Although effective visual text recognition is critical to the performance, our model that jointly reasons on visual and textual context, performs reasonably well even with sub-par visual text recognition.

(iv) Attribution ability of KaLMA: To study the impact of support fact generation (SFG) along with the answer generation on the performance of KaLMA, we train KaLMA without support fact generation, and report the results in Table 4.5. We observe that KaLMA's performance drops slightly, further supporting our claim that support fact generation elicits chain-of-thought reasoning, thereby improving the performance of answer generation along with adding attribution abilities to the model.

(iv) Cost analysis: We provide a comparison of KaLMA with a traditional non-LLM-based approach (ViLT). Our approach takes on average 5.6s per sample, which includes 4s for visual text recognition, 0.8s for entity linking using VisTEL and 0.8s for VQA using KaLMA as compared to ViLT which takes on average 0.2s per sample during inference. The training time (finetuning) of both these models are 36 and 8 hrs, respectively. Furthermore, the trainable parameters for both these models are 20M (Total size: 14B) and 114M (Total size: 114M), respectively. We achieved speed-up in our LMM components through parameter-efficient fine-tuning (LoRA) with 16-bit precision and 8-bit quantization during inference. As anticipated, traditional models have a notable advantage in terms of computational efficiency compared to our LMM-based approach. Nonetheless, we substantially surpass them in Text-KVQA accuracy.

SFG	scene	book	movie
✓	***72.7***	82.3	***77.4***
✗	71.4	***83.5***	76.9

Table 4.5: Performance of KaLMA w/ and w/o supporting fact generation (SFG).

	Text-KVQA (scene)					Text-KVQA (book)					Text-KVQA (movie)					
Method	B	D	P	L	OE	B	D	P	G	OE	B	D	P	G	L	OE
Pre-LLM Methods																
GPT-2	54.8	0.2	0.0	13.7	15.4	54.5	43.8	0.1	4.3	0.6	74.5	2.1	0.0	15.2	63.7	0.0
GPT-2 (w/ Visual Context)	57.1	0.3	0.0	16.1	17.0	80.1	63.8	5.2	45.1	7.5	75.4	3.2	0.0	24.3	66.8	29.3
ViLT	75.9	0.0	0.0	33.9	28.7	68	63.3	0	21.3	0.9	85	4.4	0.2	42.1	76.7	0.0
VLBart	78.9	0.2	0.0	18.8	27.4	79.2	62.0	1.7	34.9	0.9	***85.4***	6.3	0.0	43.7	76.7	0.0
LLM Methods																
mPlug-Owl	22	8.9	0.0	45	9.8	19.5	69.7	38.7	43.8	12	7.8	17.5	0.7	9.7	6.2	5.5
LLaVA-1.5	81.1	0.0	2.0	38.7	23.4	79	70.6	19.3	57.3	2.7	84.8	13.5	0.3	1.6	72.7	9.9
MiniGPT4v2	***81.7***	2.7	1.3	49.9	41.7	80.1	71.9	18.2	54.2	6.6	79.9	13.6	1.2	53.7	78.4	30.4
InstructBLIP	50.0	0.1	6.6	29.7	32.8	49.8	70.3	22	15.2	12.8	50.0	6.6	0.3	1.4	76.5	39.5
Ours																
KaLMA	77.2	***69.0***	***76.8***	***67.8***	***69.9***	***88.5***	***72.9***	***80.0***	***80.2***	***79.6***	84.2	***69.6***	***74.8***	***70.6***	***91.5***	***69.1***
KaLMA (Oracle)	83.9	95.8	95.4	91.9	91.8	98.0	96.4	98.2	99.9	98.2	99.9	99.8	95.9	100.0	100.0	99.7

Table 4.6: QA accuracy performance breakdown for various methods by question categories on Text-KVQA. Categories are ***B***: binary, ***D***: date, ***P***: people, ***L***: location, ***G***: genre and ***OE***: open-ended.

4.4.7 RESULTS WITH QUESTION CATEGORISATION

We show the visual question-answering results over concretized sub-categories under each of the scenes, book and movie split in Table 4.6. We observe that our proposed model shows remarkable performance across diverse question categories, particularly in the challenging categories such as date, people, and open-ended question categories.

4.4.8 FINETUNING DETAILS OF LMMS

In this section, we explain the hyperparameters and prompts used to finetune the LMMs. Note that we conduct all our experiments on a machine with 3 48GB A6000 GPUs. For mPlug-Owl and MiniGPT4v2, we have used hyperparameters as per the original papers.

mPlug-Owl: We finetuned mPlug-Owl with LoRA for 6 epochs with a learning rate of 2e-5 with a batch size of 256. LoRA details: rank: 8, alpha: 32, dropout: 0.05.

Instruction prompt template for mPlug-Owl

The following is a conversation between a curious human and an AI assistant. The assistant gives accurate and crisp answers to the user's questions.
Human: <image>
Human: {Q}
AI: {A}.

MiniGPTv4v2: We finetuned MiniGPTV4v2 with LoRA for 6 epochs with a learning rate of 3e-5

with a batch size of 128. LoRA details: rank: 16, alpha: 64, dropout: 0.05.

Instruction prompt template for MiniGPT4v2

<image>
{vqa} Based on the image, respond to this question with a short answer: {Q}, ASSISTANT: {A}

InstructBLIP: We finetuned InstructBLIP for 3 epochs with a learning rate of 1e-5 with a batch size of 128.

Instruction prompt template for InstructBLIP

<image>
USER: {Q}. ASSISTANT: {A}

LLaVA-1.5: We finetune LLaVA with LORA for 6 epochs with a learning rate of 5e-5 with a batch size of 64. LoRA details: rank: 16, alpha: 32, dropout: 0.05.

Instruction prompt template for LLaVA-1.5

<image>
USER: {Q}. ASSISTANT: {A}

4.4.9 ADDITIONAL SPLIT-WISE QUALITATIVE RESULTS

More qualitative results on movie and book splits of Text-KVQA are shown in Figure 4.7 and Figure 4.8, respectively.

4.5 CONCLUSION

We have revisited the Text-KVQA and significantly advanced state of the art on this task. Our findings suggest that visual text entity linking, combined with seamless reasoning using both visual and textual cues, as well as explicit external knowledge via LMM, is key to our success. We performed extensive ablation studies and analyses to support our claims. The future scope of this work is to expand the dataset with more visual-intensive queries and address Text-KVQA for multilingual societies.

Input Image				
Question	Who is the director of the movie?	In which year this movie was released?	Is this a biography?	What is the release year of this movie?
Ground Truth	Michele Lupo	2015	No	2014
LLaVA-1.5	John Huston ❌	1975 ❌	Yes ❌	2007 ❌
MiniGPT4v2	John Cleese ❌	2014 ❌	Yes ❌	2008 ❌
InstructBLIP	Shin'ichirō Sawazaki ❌	2013 ❌	No ✅	2013 ❌
mPLUG-Owl	Nigoi ❌	2010 ❌	Yes ❌	2020 ❌
KaLMA (Ours)	Michele Lupo ✅ [AND] the supporting fact is 'Sette volte sette is directed by michele lupo.'	2015 ✅ [AND] the supporting fact is 'Welcome to Leith was released in 2015.'	No ✅ [AND] the supporting fact is 'The sound and the shadow is a comedy movie.'	2014 ✅ [AND] the supporting fact is "Exile nation: the plastic people was released in 2014."
	(a)	(b)	(c)	(d)

Figure 4.7: A few more selection of our results as compared to implicit knowledge-based LMM approaches on the movie subset of Text-KVQA.

Input Image				
Question	Who wrote this book?	Is this a games related book?	Is this a pharmaceutical book?	What type of this book is this?
Ground Truth	Polybius	No	Yes	Travel
LLaVA-1.5	Anonymous ❌	Yes ❌	No ❌	Literature Fiction ❌
MiniGPT4v2	Pylos ❌	Yes ❌	No ❌	Literature Fiction ❌
InstructBLIP	Sophus Clausius ❌	No ✅	No ❌	Reference ❌
mPLUG-Owl	Strabo ❌	Yes ❌	Yes ✅	Fiction ❌
KaLMA (Ours)	Polybius ✅ [AND] the supporting fact is 'The rise of the roman empire (penguin classics) is written by polybius.'	No ✅ [AND] the supporting fact is 'The great cholesterol con: the truth about what really causes heart disease and how to avoid it book's genre is health, fitness and dieting.'	No ✅ [AND] the supporting fact is 'The mad dogs and englishmen genre is medical books.'	Travel ✅ [AND] the supporting fact is "Finding the center book's genre is travel."
	(a)	(b)	(c)	(d)

Figure 4.8: A few more selection of our results as compared to implicit knowledge-based LMM approaches on the book subset of Text-KVQA.

4.6 LIMITATIONS

We observe the following limitations in our work: (i) Existing visual text recognition pipelines suffer on low-resolution images where it is challenging to extract visual text, which further impacts the performance of our VisTEL (ii) In the dataset we use, it was assumed that each image contains only one visual text entity which may not be always true in a real-world scenario. (iii) Current state-of-the-art visual text recognition engines are not effective enough over multi-lingual text in the wild; Hence, in this work, we further assume the visual-text is English which again might not hold in a realistic setting. (iv) The temporal nature of knowledge, such as the entity "Statoil" being renamed "Equinor" over time, is not handled by our current models. We leave addressing

these limitations as a future work.

4.7 ETHICAL CONSIDERATIONS AND BROADER IMPACT

This work is based on the publicly available Text-KVQA dataset, which predominantly contains English visual text, and the associated knowledge base, questions, and answer pairs are also in English. The dataset may have some geographic bias that went undetected in this work, a common issue with many public computer vision and NLP benchmarks. Additionally, our work uses large multimodal models (LMMs), which can inherit and potentially amplify biases from the large-scale pretraining data used.

We are mindful of the environmental impact of using LMMs due to their heavy computational requirements. To mitigate this, we judiciously used LMMs by reusing pre-existing checkpoints wherever appropriate.

We open-source our implementation to facilitate reproduction and further study. Nevertheless, a more rigorous inspection is indeed required before deploying the proposed model in real-world applications to ensure ethical considerations are comprehensively addressed.

Broader Impact: The proposed work has the following broader impact: (i) The ability to link visual text entities to knowledge bases and leverage this linked knowledge for answering questions can improve the accuracy and relevance of information retrieval systems. Although not studied in this work, this may be particularly valuable in content recommendation systems and search engines. (ii) This research contributes to advancing the capabilities of AI systems to understand and interact with multimodal information (text and images), which can benefit applications in fields such as virtual assistants, content understanding, and automated decision-making. (iii) Methodologically, contributions such as VisTEL provide new frameworks and techniques for visual text entity linking, which can inspire further innovations in Visual NLP.

5

Augmenting VLMs with Visual Knowledge for Retrieval-based Visual Question Answering

In this chapter, we augment vision–language models with visual knowledge for retrieval-based visual question answering. While the previous chapters addressed knowledge integration in single-image settings using textual knowledge, many real-world scenarios require aggregating evidence from multiple images. To address this challenge, we introduce the task of Retrieval-based Visual Question Answering (RetVQA), where each question is paired with a pool of candidate images containing both relevant and irrelevant content. Solving this task requires identifying the supporting images and reasoning over them collectively. To benchmark progress on this setting, we curate the RetVQA dataset and highlight its distinct characteristics, including multi-image reasoning, retrieval dependence, and the need for both classification-style and open-ended answers. To solve the task, we propose MI-BART (Multi-Image BART), a retrieval-augmented encoder–decoder framework that first selects relevant images using a dense retriever and then attends across the retrieved set to generate fluent and accurate answers. Extensive experiments show that our framework achieves state-of-the-art performance on RetVQA and delivers strong gains on the image subset of the WebQA dataset, demonstrating the effectiveness of visual knowledge retrieval in improving reasoning for multi-image VQA.

5.1 INTRODUCTION

Question Answering (QA) over textual as well as visual data has been an active area of research [66, 79]. In text-based QA, the research focus has recently shifted from highly-explored QA on a single paragraph such as SQuAD [174] to a setting where mining answers from a huge corpus of documents is a requirement [2, 79]. On the contrary, visual question answering (VQA) [7] literature has so far largely restricted itself to answering questions about a given relevant visual context (often a single image). However, this does not necessarily suffice to satisfy our information needs

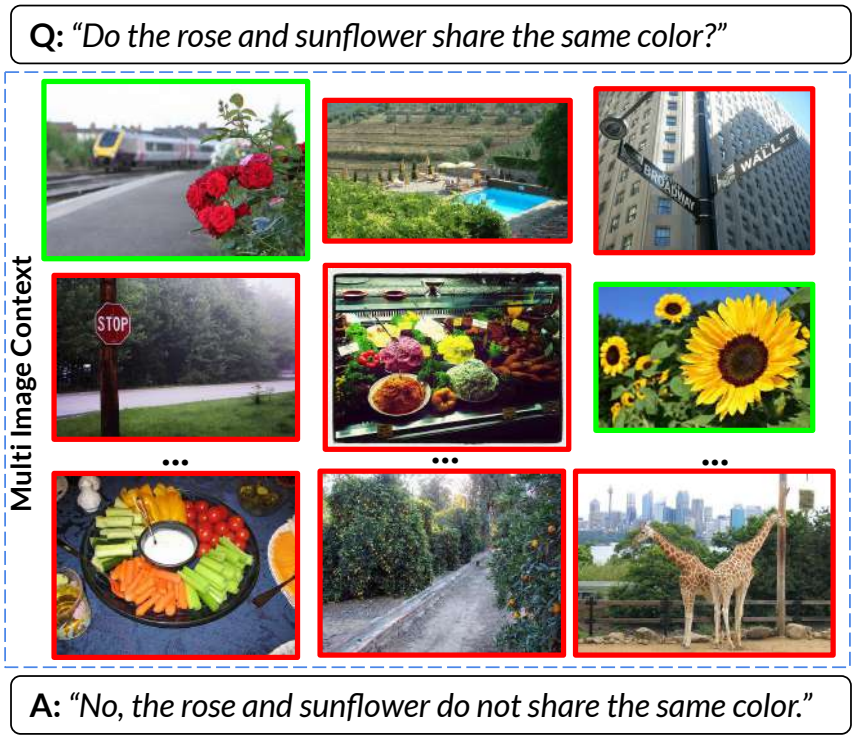


Figure 5.1: Given a question and a pool of images (multi-image context), RetVQA task involves two stages: (i) retrieve the relevant images from the pool, and (ii) generate a free-form natural language answer by reasoning over the retrieved relevant images as context.

since the information may be spread across multiple images and may not be present in some images. For example, consider a natural language question '*Do the rose and sunflower share the same color?*', answering such a question from a pool of images as visual context (refer Figure 5.1), requires a model to first retrieve relevant images and then perform visio-lingual reasoning on the retrieved images to arrive at a fluent free-form natural language answer. We refer to this problem as RetVQA or retrieval-based visual question answering. The RetVQA setting has potential applications in question answering on web images, e-commerce, environmental monitoring, and health care, among others, e.g., multiple images of a particular area can be analyzed to monitor environmental changes over time; multiple MRI or CT scans of a patient's brain need to be analyzed to detect abnormalities, such as tumors.

For RetVQA, the input is a pool of images with only a few images being relevant to the question. Close to our setting, there is some exciting progress in the recent literature [15, 24, 197, 205]. However, these works assume one or more of the following constraints: "without requiring explicit retrieval", "having classification-type fixed-vocabulary answers", "assuming the availability of meta-data like *WikiEntities, captions*", "having a homogeneous yet limited number of images in the pool", and "having only a small set of questions that need multiple images". Such constraints in the existing datasets point towards a need for a large-scale benchmark to study RetVQA. To this end, we present a *derived* dataset prepared from Visual Genome [110], leveraging its questions and annotations of images. We curate questions under different categories: (i)

Measurement	*Value*
#Distinct questions	418K
#Distinct precise answers	16,205
Train set questions	334K (80%)
Val set questions	41K (10%)
Test set questions	41K (10%)
Avg question length (words)	8.7
Avg answer length (words)	8.5
#Distinct words in questions	10,868
#Distinct words in answers	9,278
#Avg relevant images per question	2
#Avg irrelevant images per question	24.5

Table 5.1: Key statistics for RetVQA dataset.

common attributes such as color, shape, and count, (ii) other object-attributes that include non-common attributes, e.g. length, material, and (iii) subject-object relationships, e.g., 'eats', 'left of'. Further, to facilitate benchmarking capabilities of the VQA models over open-ended answers, we curate questions under binary (yes/no) and open-ended answer categories. Note that the answers are free-form fluent in both the answer categories, e.g. '*No, rose and sunflower do not share the same color*' (a binary answer); '*The color of rose and sunflower is red and yellow, respectively*' (an open-ended generative answer). RetVQA dataset statistics and distribution across the question-answer categories are shown in Table 5.1 and Table 5.2, respectively.

Further, to solve the RetVQA task, a model must first retrieve the relevant images for the question and then consume the retrieved images as the context to answer the question. Towards this end, we present a unified Multi Image BART that takes in the question along with the multi-image context retrieved using a relevance encoder to generate the free-form fluent natural-language answer. Our proposed framework, MI-BART, allows joint reasoning over multiple retrieved images along with the question to capture better semantics.

The first contribution of this work is the introduction of RetVQA, a dataset that is 20$\times$ larger than the closest existing dataset [24] in this setting. RetVQA is constructed by leveraging questions and image annotations from Visual Genome, with a particular focus on multi-image, metadata-independent questions over a heterogeneous pool of images. The dataset is designed to encourage both classification-oriented and generative answers, thereby promoting deeper reasoning capabilities. We believe that the proposed task, dataset, and benchmarks will serve as a strong foundation for future research in this direction. The second contribution is MI-BART, a unified method that jointly reasons over the retrieved multi-image context and the question in order to generate a fluent, free-form answer. Unlike single-image or metadata-driven approaches, MI-BART explicitly integrates signals from multiple retrieved images, enabling robust reasoning across heterogeneous visual content. The third contribution is extensive experimentation to evaluate the effectiveness of our proposed framework on both RetVQA and the image segment of WebQA. Our approach significantly outperforms baseline methods on RetVQA and achieves state-of-the-art results on the WebQA image benchmark.

Question Category	*Binary*	*Open-ended*	*Total*
Color	50K	50K	100K
Shape	49K	50K	99K
Count	50K	50K	100K
Object-attributes	80K	-	80K
Relation-based	-	38K	38K
Total	229K	188K	418K

Table 5.2: Distribution of questions by various categories in RetVQA dataset. The answers are of two types: binary-generative and open-ended generative.

The remainder of this chapter is organized as follows. In Section 5.2, we review related literature on visual and multi-modal QA, as well as multimodal models. Section 5.3 introduces our newly curated RetVQA dataset and analyzes its key properties. Section 5.4 presents our retrieval-based QA methodology, including the problem formulation, multimodal relevance encoder, and the proposed MI-BART framework. Experimental setup, results, and ablation studies are discussed in Section 5.5, followed by qualitative analysis. Finally, Section 5.6 concludes the chapter with key takeaways and future research directions.

5.2 RELATED WORK

Visual and Multi-modal QA. Visual Question Answering (VQA) aims at answering a natural language question in the context of a relevant image [7]. This area has seen significant progress partly due to the introduction of several challenging datasets [7, 62, 92, 110, 142, 178, 272]. Most methods for VQA either use a multimodal fusion of language and image embeddings [56, 101, 154, 178], attention-based multimodal fusion [54, 139, 190, 243, 251] or neural module networks [6, 83]. More recently, knowledge-based VQA [146, 186] has gained attention where external knowledge is used for answering visual questions. Contrary to these exciting works in VQA literature, our problem setting is distinctively different as we need to mine the answer from a collection of relevant as well as irrelevant images.

Sharing a similar motivation as ours, the following tasks and accompanying datasets have been recently introduced in the literature: (i) MultimodalQA [205], (ii) ISVQA [15], and (iii) WebQA [24]. In MultimodalQA [205], only a small part of the dataset (ImageListQ) is relevant to our setting; however, even on this subset, MultimodalQA assumes the availability of extra image metadata, i.e., table or WikiEntity linkage. Similarly, in ISVQA [15], every question has a small set of homogeneous images as context. Since images are homogeneous, there is no need for explicit retrieval. Recently, WebQA [24] dataset has been proposed to target such practical QA scenarios, where a question has to be answered in the context of multimodal sources; however, the images in the dataset have associated captions. All of these settings have differences from RetVQA in either usage of additional context, constraints on images in the collection, answer schema, or

Dataset	#Questions	Retrieval required	Heterogeneous images	Multi-image reasoning	No Meta-data assumption	Answer type	% of questions that need multiple images
MultimodalQA [205]	2K	✗	✗	✓	✗ (WikiEntities)	Classification	6.1%
ISVQA [15]	141K	✗	✗	✓	✓	Classification	33%
WebQA [24]	18K	✓	✓	✓	✗ (Captions)	Open-ended	44%
RetVQA (Ours)	**327K**	✓	✓	✓	✓	Open-ended	**100%**

Table 5.3: Comparison of our curated dataset RetVQA with other relevant QA datasets. For Multimodal QA and WebQA datasets, we have considered their image-only modality questions subset.

classification instead of generation. In another recent work MIMOQA [197], only extractive question answering is performed. In terms of reasoning on more than one image, another related work is NLVR [201]. However, it does not involve any retrieval and open-ended answer generation. Further, in terms of QA over multiple document images or video frames, related works are DocVQA [148] and VideoQA [113, 114, 207]. DocVQA focuses on text-heavy document images, limiting visual reasoning, whereas the VideoQA task does not involve explicit retrieval and open-ended answer generation. Contrary to these works, our newly curated dataset is significantly larger, has no assumption of meta-data availability, and requires retrieval and reasoning over multiple images to arrive at an answer. A comparison of RetVQA with the relevant datasets is shown in Table 5.3.

Multi-modal modeling. Recently, multi-modal transformers such as VisualBERT [122], VilBERT [138], VILT [106], LXMERT [206], OSCAR [126], UNITER [31] have shown strong results on the downstream vision and language tasks. However, these encoder-based models are more suited for classification-style VQA settings. Multimodal transformers like VLP [269] and VLBart [35] are pre-trained with sequence-to-sequence objectives and hence are more suitable for the current setting that requires the model to generate free-form natural language answers. We follow a similar approach by devising an encoder-decoder framework to jointly reason over multiple images along with the question.

5.3 RETVQA DATASET

Traditionally, VQA datasets [7, 62, 196, 205] assume that the context provided is always relevant to the question. Recently, [24] proposed a benchmark where given a question and a pool of multimodal sources (containing both image and text snippets as context), only a few of these sources are relevant to the question. Similar to our problem setup, it requires first retrieving the relevant context and then using it to answer the question. However, after carefully observing the WebQA dataset, we found that most questions include rare entities like 'Maracana Stadium', 'Minnetonka Rhododendron', etc. Such questions make the retrieval task of the problem non-trivial without auxiliary information about these images. Methods proposed in [24] leverage metadata like image captions, which contain information like the name of the entity in the image. Further, a rule-based retrieval using word overlap of the question with the image caption for the retrieval task has an F1 score of 37, asserting our claim that retrieval is over-dependent on image metadata and not sig-

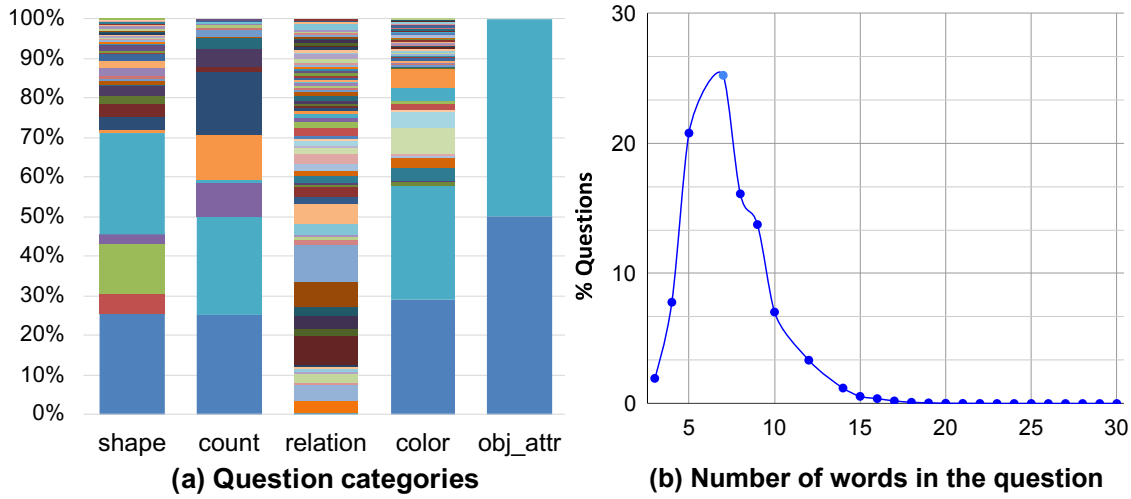


Figure 5.2: RetVQA questions and answers analysis: (a) Answer distribution over various question categories, (b) Distribution of the number of words across questions.

Figure 5.3: Word cloud of Top-80 frequent answers.

nificantly on visual data. Further, the image-based subset of WebQA has a majority of samples (55.6%) with one relevant image per question, thereby, a single image VQA method may perform equally well given the relevant image is retrieved.

To overcome such limitations, we curate a dataset RetVQA from Visual Genome, where we emphasize multi-image, metadata-independent questions over a pool of heterogeneous images, expecting a mix of classification-oriented and open-ended generative answers. We leverage question-answer and object annotations of Visual Genome to curate the dataset. We curate truly multi-image questions spanning over five different categories, namely, color, shape, count, object-attributes, and relation-based. For each question category, we curate binary-generative and open-ended gen-

Figure 5.4: Word cloud for color, shape, count type question categories respectively (left to right) in the RetVQA dataset.

erative answers. Dataset statistics are shown in Tables 5.1 and 5.2.

The questions in RetVQA are curated as follows. We start by extracting subjects and relations of the existing question-answer pairs from Visual Genome; for example, consider these two question-answer pairs: q_1 (over image I_1): "What is the cow eating in the image?" where the answer is a_1: "grass"; and, q_2 (over image I_2): "what is the sheep eating?" where the answer is a_2: "grass". Given q_1 and q_2, we extract subjects ("cow" and "sheep"), relations ("eating (eat/eats)"), and then we frame combined questions using templates (over images I_1 and I_2) as follows. q_3: "what else eats the same thing as cow does?" with answer a_3: "sheep eats the same thing as cow". Another question could be q_4: "Does cow and sheep eat the same thing?" where the answer is a_4: "Yes, cow and sheep eat the same thing". Thus, we curate binary-generative (like q_4) and open-ended generative (like q_3) types of answers. We further associate negative images for each of the curated questions using their object annotations as follows. A negative image is one where both the subject and object (used to generate the question) do not exist together in the image. This enforces that the answer has to be inferred only when all the relevant images are correctly retrieved and the negatives serve as sufficiently hard negatives.

We use a random 80%-10%-10% train-validation-test split. All the questions in our dataset have at least two relevant images and 24.5 irrelevant images on average. Further, Figure 5.2 shows the distribution of unique answers across question types and question length distribution.

5.3.1 WORD CLOUDS FOR RETVQA

Figure 5.3 shows the word cloud of top-80 frequent answers. We observe that most questions are in the 5–10 words range, and there is no noticeable bias towards the majority of answers in the dataset. Figure 5.4 shows word clouds for popular short answers for the color, shape and count question categories, respectively (left to right) in the RetVQA dataset. There is a large coverage across different unique values showing that the dataset does not suffer from any majority bias across various question categories.

5.4 RETRIEVAL QA METHODOLOGY

5.4.1 RETVQA PROBLEM FORMULATION

The RetVQA problem is defined as follows. Given a natural language question Q, a set of N heterogeneous images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, the task is to generate an answer (A) for the question Q based on $\mathcal{I}$ where only a few images are relevant for the question. To answer the question Q using $\mathcal{I}$, we need a method that retrieves the relevant images $\mathcal{I}^r \subseteq \mathcal{I}$ and then leverages the retrieved context $\mathcal{I}^r$. Accordingly, our labelled dataset consists of quadruplets $(Q, \mathcal{I}, \mathcal{I}^r, A)$.

5.4.2 RETVQA FRAMEWORK

The proposed framework solution: (i) multi-modal relevance encoder for retrieval of relevant sources $\mathcal{I}^r$ from $\mathcal{I}$ for the given question Q and (ii) a unified Multi Image BART (MI-BART) to generate fluent free-form natural language answer for the question Q using the retrieved images $\mathcal{I}^r$ as context.

Image representation. Inspired by recent vision-language pretraining literature [31, 122, 126], for every image I_i in $\mathcal{I}$ where $i \in \{1, 2, \ldots, N\}$, we first detect a fixed set of $\mathcal{P}$ objects using Faster R-CNN [179] pretrained on Visual Genome [110]. For every object p, where $p \in \{1, 2, \ldots, \mathcal{P}\}$, we obtain 2048-dimensional regional feature $\mathbf{o}_p^{reg}$ and 4-dimensional bounding box co-ordinates $\mathbf{o}_p^{bbox}$. Thereby, each image I_i is represented by a set of $\mathcal{P}$ object proposals $\{(\mathbf{o}^{reg}, \mathbf{o}^{bbox})_p\}_i$. Following [122], for every region p we project both 2048-dimensional regional representation and 4-dimensional bounding box coordinates into the d-dimensional space using a linear projection to obtain $\{\mathbf{o}_p\}_i$ and then concatenate across all regions within the image to obtain image embedding $\mathbf{o}_i$ as follows.

$$\mathbf{o}_i = \{\mathbf{o}_p\}_i, \text{ where } p \in \{1, 2, \ldots, P\}, i \in 1, 2, \ldots, N. \tag{5.1}$$

Question representation. We encode the textual question Q containing M words using a pretrained BERT [46]. This results into a sequence $\mathbf{q}$ of M d-dimensional vectors, $\mathbf{q} = \{\mathbf{q_m}\}$ where $m \in \{1, 2, \ldots, M\}$. Note that if any additional metadata is available (e.g. captions in WebQA dataset), we augment it to the question.

$$\mathbf{q} = \{\mathbf{q_m}\} = BERT(Q), \text{where } m \in \{1, 2, \ldots, M\}. \tag{5.2}$$

5.4.3 MULTIMODAL RELEVANCE ENCODER FOR IMAGE RETRIEVAL

Pretraining. Our multi-modal Relevance Encoder (RE) consists of three transformer encoder layers followed by a multi-layered perceptron (MLP) with a sigmoid unit over the final representation of the $[CLS]$ token. We pretrain our relevance encoder on MS-COCO [129] using two unsupervised objectives, Image Text Matching (ITM) and Masked Language Modelling (MLM) similar to [122].

Question-Image relevance learning. Each sample in our dataset contains a question Q and N images $I_1, I_2, \ldots, I_N$ of which some have been labelled as positive and others negative. Further, for each image, we have P regions. We use each question-image pair (Q, I_i) to learn question-image relevance using our multi-modal Relevance Encoder (RE). Our pretrained multi-modal relevance encoder is fed with question-image pairs, along with two special tokens, $[CLS]$ and $[SEP]$; in short, the input to our relevance encoder is $[[CLS], \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}\}, [SEP], \{\mathbf{o_1}_i, \mathbf{o_2}_i, \ldots, \mathbf{o_p}_i\}]$. Our encoder then allows the input $M + P + 2$ token sequence of question-image features to attend to each other and produces a sequence of contextualized embeddings. The d-dimensional contextualized embedding of $[CLS]$ token is further fed to an MLP with a sigmoid unit to produce a relevance score ($\hat{s}_i$) between 0 and 1 (Eq. 5.3), indicating whether the given question-image pair (Q, I_i) is relevant or not. We finetune our multi-modal relevance encoder parameters ϕ by minimizing binary cross-entropy loss $\mathcal{L}_{REL}(\phi)$ (Eq. 5.4).

$$\hat{s}_i = RE_\phi(Q, I_i). \tag{5.3}$$

$$\mathcal{L}_{REL}(\phi) = -\mathbb{E}_{(Q,I_i)\sim D}[s_i \log(\hat{s}_i) + (1 - s_i) \log(1 - \hat{s}_i)]. \tag{5.4}$$

Given a question Q and a set of N images $\mathcal{I}$ sampled from our dataset D, we obtain relevance scores $S = \{\hat{s}_i\}_{i=1}^N$ for each question-image pair $(Q, \{I_i\}_{i=1}^N)$ using our fine-tuned relevance encoder (Eq.5.5). To choose the final set of relevant images $\mathcal{I}^r$ from the pool of images $\mathcal{I}$, we rank all the images in the pool using S and choose top-K images as our relevant context $\mathcal{I}^r$ for the given question Q (Eq. 5.6).

$$S = \{\hat{s}_i\}, \text{ where } \hat{s}_i = RE(Q, I_i), i \in 1, \ldots, N. \tag{5.5}$$

$$\mathcal{I}^r = \{I_k\} \text{ where } k \in \text{top-}K(S). \tag{5.6}$$

5.4.4 MULTI IMAGE BART FOR QUESTION ANSWERING

Given the question and the retrieved images $\mathcal{I}^r$, the goal of MI-BART is to generate an accurate yet fluent free-form natural language answer for the question. Towards this end, we propose an encoder-decoder architecture similar to SimVLM [226]. MI-BART encoder is a stack of six transformer layers [213], where each transformer layer comprises a self-attention layer, followed by a fully connected linear layer with a residual connection. Similarly, the MI-BART decoder is also a stack of six transformer layers [213], with an additional cross-attention layer in each transformer

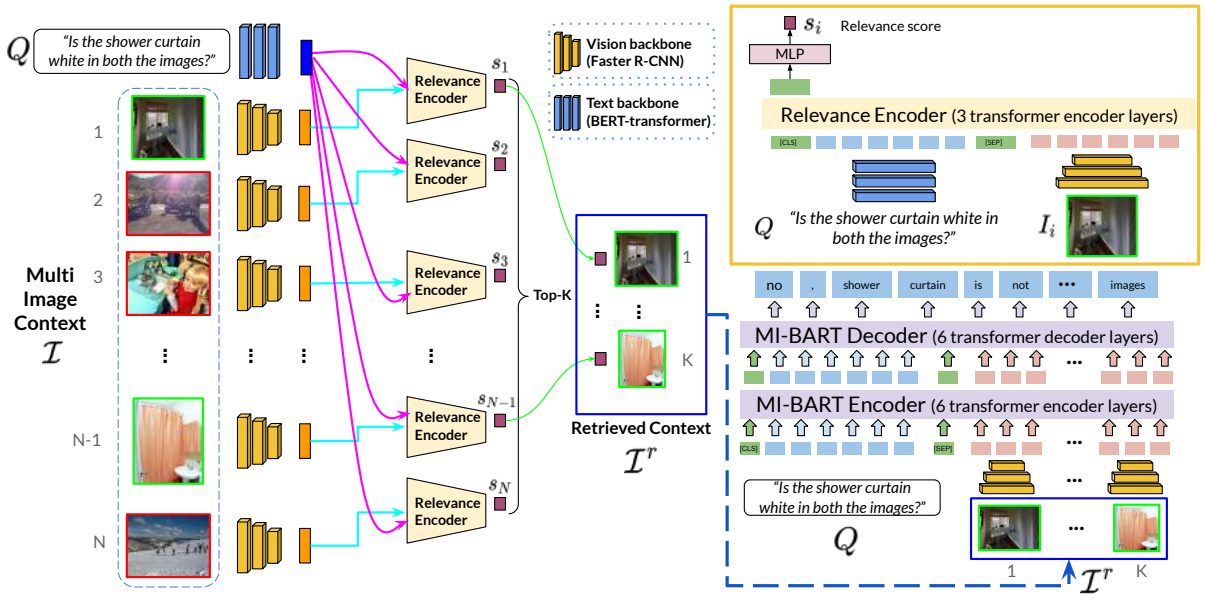


Figure 5.5: An overview of our proposed framework for retrieval-based VQA. Given a question Q and a pool of images $\mathcal{I}$, we encode the question and each image using a pretrained BERT and a pretrained Faster R-CNN, respectively. Once encoded, our multimodal relevance encoder (shown in the yellow box at the top right) generates relevance scores S for all images in the set with the question. We choose top-K scoring images as the retrieved relevant images $\mathcal{I}^r$. We encode the images in $\mathcal{I}^r$ using Faster R-CNN and feed them to our MI-BART encoder along with Q to facilitate joint reasoning over the multi-image context with respect to the question. Once the MI-BART encoder encodes the question in the context of retrieved images, the MI-BART decoder generates the free-form natural language answer A to the question.

layer. We concatenate question embedding $\mathbf{q}$ with image embeddings of each image I_k in $\mathcal{I}^r$ with a special token $[SEP]$ in between. Also, to distinguish the image features belonging to different images in the retrieved image set $\mathcal{I}^r$, we assign image order ids to image features from different images. Note that image order ids are not meant for assigning a sequence number to images in the retrieved set. Their sole purpose is to differentiate image features from different images. In short, the input to our MI-BART encoder is $[[CLS], \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}\}, [SEP], \{\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_p}\}_1, [SEP], \ldots, \{\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_p}\}_k]$, where $m \in \{1, 2, \ldots, M\}$, $p \in \{1, 2, \ldots, P\}$ and $k \in \{1, 2 \ldots, K\}$. These inputs attend over each other through various self-attention layers of the MI-BART encoder and produce a sequence of contextualized embeddings $\mathbf{z} = \{\mathbf{z_j}\}$, where $j \in \{1, 2, \ldots, M + (P \times k) + k\}$ (Eq. 5.7).

$$\mathbf{z} = \{\mathbf{z_j}\} = \text{MI-BARTEncoder}(Q, \mathcal{I}^r). \tag{5.7}$$

MI-BART decoder auto-regressively predicts the probability of the next token A_t in the answer A by attending to these encoder outputs $\mathbf{z}$ and previously generated answer tokens $a_{<t}$ through cross-attention and self-attention layers, respectively (Eq. 5.8). We train MI-BART decoder parameters θ by minimizing the generative loss $\mathcal{L}_{GEN}(\theta)$ for generating the target answer token conditioned on the question Q and retrieved image context $\mathcal{I}^r$ (Eq. 5.9). During training, we leverage the ground truth relevant images as retrieved image context $\mathcal{I}^r$, while during inference,

we obtain it from our relevance encoder.

$$P_\theta(A_t|A_{<t}, Q, \mathcal{I}^r) = \text{MI-BARTDecoder}(\mathbf{z}, A_{<t}). \tag{5.8}$$

$$\mathcal{L}_{GEN}(\theta) = -\mathbb{E}_{(Q,\mathcal{I}^r)\sim D}\left[\sum_{t=1}^{|A|}\log(P_\theta(A_t|A_{<t}, Q, \mathcal{I}^r))\right]. \tag{5.9}$$

To summarize, our proposed framework works as follows, given a question Q and a pool of N images $\mathcal{I}$, we (i) obtain question-image relevance scores S for each question-image (Q, I_i) pair in $(Q, \mathcal{I})$ using our multimodal relevance encoder, (ii) rank all images in the pool based on S, and choose top-K images as retrieved images context $\mathcal{I}^r$, and (iii) question Q along with the retrieved image context $\mathcal{I}^r$ is fed to MI-BART which encodes the provided context and generates the free-form natural language answer A to the question Q. The proposed framework is illustrated in Figure 5.5.

5.4.5 IMAGE-STITCH MI-BART

Inspired by [15], we consider stitching the retrieved images along the width into a single joint image and use single image VQA on this joint image. We use our proposed MI-BART for this baseline; however, we feed the stitched joint image instead of feeding K images. While the MI-BART combines information across images in the embedding space, the image-stitch MI-BART combines information across images in the input space.

5.5 EXPERIMENTS AND RESULTS

We conduct our experiments on RetVQA as well as on WebQA. Since our task deals with the image set as a given context, we consider the image-only subset of the WebQA dataset. Following [24], we use accuracy (A), fluency (F), and $F \times A$, as metrics to evaluate the generated answers. Accuracy validates whether the correct answer is present in the generated answer, whereas fluency measures the quality of the answer paraphrase. Fluency is computed using a recently proposed natural language generation metric called BARTScore [256]. Further, an F1 score is used for retrieving relevant images from a pool of images.

5.5.1 BASELINES

We compare our proposed method (MI-BART) and its variant image-stitch MI-BART with the following baselines: ***(i) Popularity-based Baselines***: To check for prior biases associated with frequent answers globally or per question category, we use two popularity-based baselines. (a) Global popularity: the most frequent answer in the training set is always considered as the answer by

the model, and (b) Per-category popularity: the most frequent answer for each question category is always considered the answer for questions in the corresponding question category. ***(ii) Aggregate VQA***: RetVQA task involves VQA over multiple images. In the Aggregate VQA baseline, we use the traditional single-image VQA method [7] for each image and aggregate the results. Given a question Q and its corresponding K retrieved images, $\mathcal{I}^r$ from our relevance encoder, we feed each retrieved image along with the question Q to a single image VQA model to get a joint representation. We concatenate joint representations of all retrieved images into a single representation F and feed to a linear layer (MLP) to predict the final answer, i.e., $A = MLP(F)$. Since traditional VQA methods follow a classification-style answer prediction approach, we use the 1000 most frequent answers as classes in the softmax layer. To generate a fluent answer, we prepend the predicted answer to the question after removing the first word from the question. This baseline is not benchmarked on the WebQA dataset, as the dataset does not provide precise answer annotations for the trainset questions. ***(iii) VLP***: As the RetVQA task requires the model to generate the text, encoder-only multimodal transformer models like ViLBERT [138], VisualBERT [122], OSCAR [126], ViLT [106] and UNITER [31] are not directly suitable. Hence, we use VLP [269] which is a unified encoder-decoder multimodal transformer as our baseline. We finetune a pretrained VLP on our datasets for evaluation.

5.5.2 ABLATIONS

We perform the following ablations to better understand the various components of our proposed model. (i) *Question-only*: To study the role of the images in generating accurate and fluent answers, we ignore the images and use questions only as input to our model. (ii) *Single-image retrieval*: To study the importance of reasoning over multiple images to generate an answer to the question, we use top-1 retrieved image as our only context instead of a multi-image context. (iii) *Missing captions*: To study the role of image metadata in the relevant source image retrieval and, thereby, the answer generation, we conduct experiments on WebQA without leveraging the image metadata (captions). In this ablation, we augment captions (available in WebQA) as part of the textual input in both the relevance encoder and MI-BART.

5.5.3 IMPLEMENTATION DETAILS

We have implemented our framework in PyTorch [162] and Hugging Face's transformers [231] library. Our relevance encoder has three transformer layers, each having eight attention heads. We pretrain our relevance encoder on MS-COCO [129] with a constant learning rate of 1e-4 using Adam optimizer [107]. Using the same optimiser, we finetune the relevance encoder on both datasets with a constant learning rate of 2e-5. Our MI-BART has six standard transformer encoder layers and six standard transformer decoder layers [213]; we initialize our MI-BART with VLBart [35] pretrained weights to leverage the strong visual-textual learning of VLBart. We further finetune MI-BART on a multi-image QA task with a learning rate of 5e-5 using Adam optimizer with a linear warm-up of 10% of the total steps. Our relevance encoder and MI-BART

	RetVQA						WebQA					
	Oracle Images			Retrieved Images			Oracle Images			Retrieved Images		
Method	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A
Popularity-based Baselines												
Global popularity	27.4	14.5	7.9	36.2	16.7	9.8	17.7	1.3	0.4	17.7	1.3	0.4
Per-category popularity	27.8	16.0	7.6	27.8	16.0	7.6	25.2	1.3	0.5	25.2	1.3	0.5
Other Baseline Approaches												
Question only	62.4	15.3	10.4	62.4	15.3	10.4	22.2	34.9	13.4	22.2	34.9	22.2
Aggregate VQA	69.2	17.1	13	66.6	16.2	11.9	*	*	*	*	*	*
VLP [269]	65.1	70.2	58.8	65.1	70.2	58.8	45.7	42.2	25.9	44.2	38.9	24.1
MI-BART (Ours)												
Image stitch ***MI-BART***	78.2	74.7	70.7	72.1	76.6	66.8	49.6	50.5	27.5	***49.1***	50.3	27.4
MI-BART	***84.2***	***85.6***	***79.8***	***76.5***	***79.3***	***70.9***	***49.8***	***51.1***	***28.1***	48.7	***50.7***	***27.6***

Table 5.4: Performance comparison of various methods on RetVQA and image segment of WebQA. * WebQA only provides full-sentence answers rather than answer category annotations. Therefore, classification model like AggregateVQA cannot be trained for WebQA.

	Color			Shape			Count			Object-attributes			Relation-based		
Method	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A
Popularity-based Baselines															
Global popularity	25.4	8.2	0.6	24.5	9.2	1.3	25.6	13.4	6.7	49.5	35.2	30.5	0.0	4.8	0.0
Per-category popularity	25.3	9.1	0.5	24.5	9.2	1.3	24.5	14.4	4.5	49.5	35.2	30.5	6.0	15.6	2.3
Other Baseline Approaches															
Question-only	58.0	12.2	6.1	86.9	13.3	11.8	51.6	15.3	9.6	74.9	21.8	18.3	12.4	14.7	4.1
Aggregate VQA	60.1	12.4	6.7	91.3	14.4	13.5	54.6	15.6	10.3	75.4	21.9	18.6	32.2	20.8	11.9
VLP [269]	62.0	67.3	52.8	84.0	81.0	75.7	50.8	70.0	50.8	76.8	78.2	74.8	36.8	33.8	18.5
MI-BART (Ours)															
Image stitch ***MI-BART***	71.8	***76.8***	63.7	***96.2***	***94.4***	***91.1***	62.7	***80.1***	62.6	***81.6***	***87***	***81.4***	52.0	39.5	26.4
MI-BART	***72.1***	75.7	***63.9***	92.4	90.3	87.7	***66.0***	80.0	***66.0***	78.5	83.4	78.4	***69.5***	***50.4***	***43.1***

Table 5.5: Performance breakdown for various methods by question categories on RetVQA with the retrieved images.

were trained using 3 Nvidia RTX A6000 GPUs with a batch size of 96 and 256 while training and a batch size of 360 and 480 during testing, respectively.

5.5.4 RESULTS ON RETVQA

We conduct our experiments in two settings, namely, (i) Oracle images: Here, we use ground-truth relevant images for answer generation, and (ii) Retrieved images: Here, relevant images are retrieved using our relevance encoder. We show the results under both these settings in Table 5.4. We observe that the popularity-based methods perform poorly. This result is expected as popularity-based methods do not use any question or image context. Methods that involve either questions or images perform better than the popularity-based baselines. However, the question-only baseline has a F×A score of 10.4 on RetVQA, showing that image context is needed to generate accurate yet fluent answers. Transformer-based baseline VLP and image-stitch MI-BART reach a F×A score of 58.8 and 70.7 on our dataset, respectively, in the oracle setting, compared to 79.8 of our proposed MI-BART framework. Image-stitch MI-BART outperforms transformer-based VLP by 12% on our dataset and 1.5% on the WebQA dataset, which shows that having

	Binary			Open-ended		
Method	Acc.	F	F×A	Acc.	F	F×A
Popularity-based Baselines						
Global popularity	49.9	17.3	14.3	0.0	11.1	0.0
Per-category popularity	49.5	17.1	13.3	1.2	14.6	0.5
Other Baseline Approaches						
Question-only	74.2	11.5	9.2	48	20	12
Aggregate VQA	75.6	11.5	9.5	55.5	22	14.9
VLP [269]	73.2	80	72.5	55.1	58.2	42
MI-BART (Ours)						
Image stitch ***MI-BART***	***80.4***	***88.3***	***80.3***	69.4	70.2	57
MI-BART	78.7	85.6	78.7	***73.7***	***71.7***	***61.5***

Table 5.6: Performance breakdown by answer categories for various methods on RetVQA with the retrieved images.

Retrieval	Acc.	F	F×A
Top-1 retrieved image	59.8	61.1	48.8
All retrieved images	***76.5***	***79.3***	***70.9***

Table 5.7: MI-BART performance on RetVQA using different retrieval strategies.

a separate decoder in the proposed MI-BART baseline has better reasoning capabilities than a unified encoder-decoder like VLP. We further present the QA results over the retrieved images setting using our relevance encoder, which has an F1 score of 71 at the top-2. All the approaches involving image context outperform question-only baseline, emphasizing that RetVQA has a reasonable utility to develop and benchmark methods capable of jointly reasoning over multi-image context and the question.

Further, in Table 5.5, we show the QA results over various question categories, and in Table 5.6, we show the results over answer categories under the retrieved images setting. Our framework outperforms baselines, especially in questions with open-ended generative answers, which constitute nearly half of our dataset. As expected, open-ended generative questions are more challenging than binary ones. However, compared to the baselines, MI-BART provides better improvement for open-ended questions than binary ones by jointly reasoning over multi-image context. Results in Table 5.7 further emphasize our hypothesis of requiring multiple images to answer the given question. We show the missing caption ablation results on the image-subset of WebQA in Table 5.8; this result further affirms our claims that the performance of methods on the WebQA dataset depends on the image metadata like captions.

5.5.5 QUALITATIVE ANALYSIS

We illustrate a selection of results using our proposed approach and one of the most competitive baselines viz. VLP in Figure 5.6. In both these results, our approach correctly answers the question in a large heterogeneous visual context. Further, to understand the importance of the multimodal

	Retrieval			***QA***
Captions	P	R	F1	F×A
w/ captions	32.3	44.7	37.5	19.7
w/o captions	***79.7***	***86.3***	***77.4***	***28.1***

Table 5.8: Effect of w/ and w/o captions in WebQA: Performance of MI-BART on retrieval and QA (retrieved images setting).

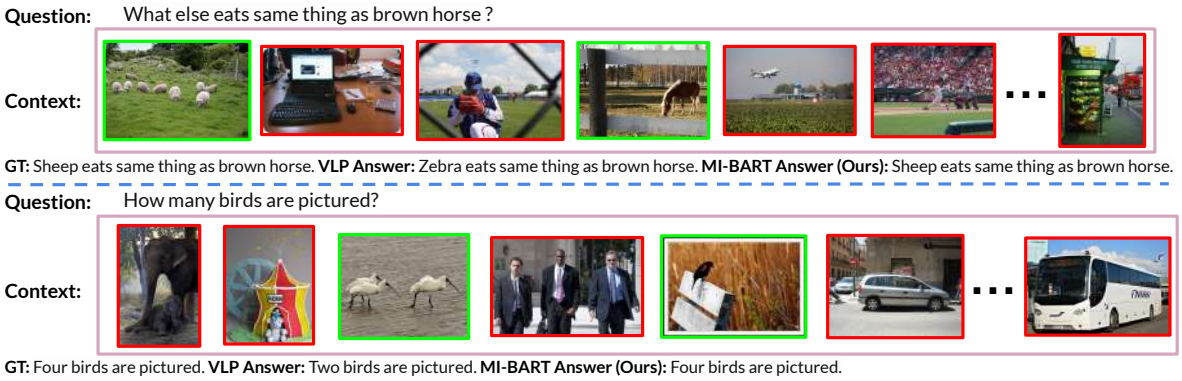


Figure 5.6: A selection of QA using MI-BART (Ours) and VLP (best baseline) from RetVQA test dataset. The images in the green and red bounding boxes show relevant and irrelevant images, respectively. Our approach consistently performs better than the baseline on different types of questions requiring multiple images to arrive at an accurate answer. (Best viewed in color).

input for question answering, we plot the multimodal attention map over the retrieved images and the question during the answer generation in Figure 5.7. The figure shows that our proposed MI-BART model attends to the relevant regions of both images while generating the main answer word 'sheep'. Further, we observe that it attends to brown horse regions of the first image along with the corresponding question parts while generating 'brown horse'. Thus, both images are needed and paid attention to when generating the right answer. We further conducted a detailed error analysis on 50 randomly chosen samples where our model failed to generate a correct answer. We categorize the errors into four major categories: (i) partial retrieval: images retrieved by relevance encoder are partially relevant (52%). (ii) Incorrect retrieval: images retrieved by the relevance encoder are entirely irrelevant (26%). (iii) Incorrect reasoning: model generating a partially incorrect answer despite all the retrieved images being relevant (40%).

5.5.6 VARYING IRRELEVANT IMAGES IN RETVQA EXPERIMENT

We evaluate the proposed framework MI-BART on RetVQA with the varying numbers of irrelevant images in the pool. We report F1 on the retrieval task, along with accuracy, fluency and F × A metrics for question answering in Table 5.9.

As expected, we observe that retrieval F1 drops as we increase the pool size. However, thanks to the robustness of our proposed question answering approach, MI-BART, F×A does not drop by the same extent.

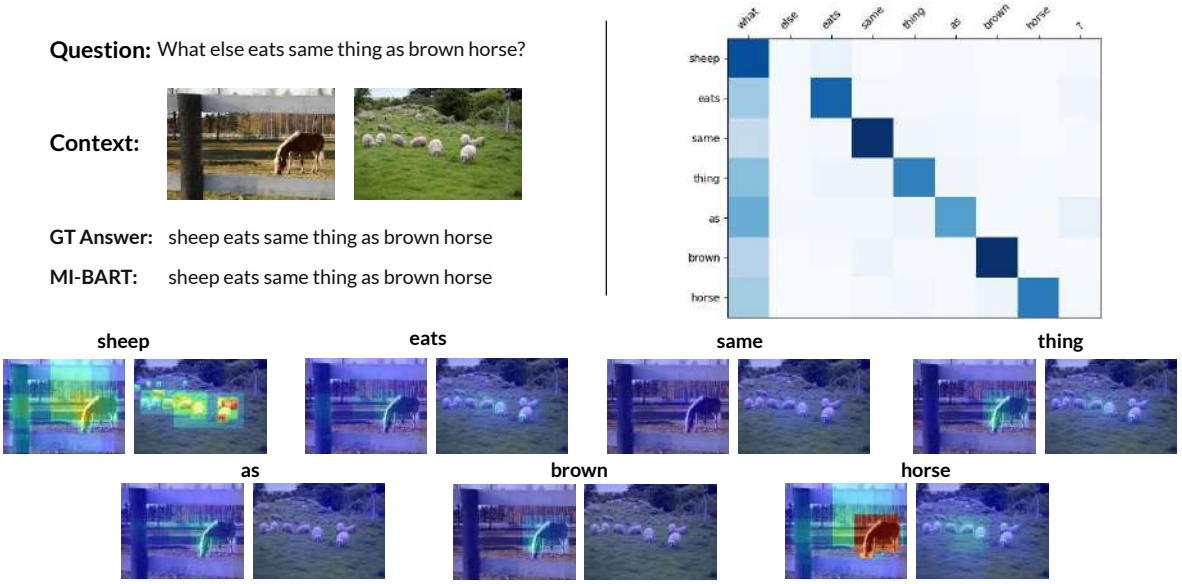


Figure 5.7: Multimodal attention map over the retrieved images and the question during the answer generation. We observe that our proposed MI-BART attends to the relevant regions of both images while generating the main answer word 'sheep'. Further, we see that it attends to brown horse regions of the first image along with the corresponding question parts while generating 'brown horse'. [Special tokens are removed for visualization.] (Best viewed in color).

	Retrieval	QA		
#Irrelevant images	F1	Acc.	F	F×A
25	71.0	76.5	79.3	70.9
50	60.3	73.3	76.9	67.5
100	48.4	69.8	74.2	63.5
200	36.3	66.9	71.8	60.2

Table 5.9: MI-BART performance on RetVQA with the varying number of irrelevant images in the pool.

5.5.7 RESULTS ON PARAPHRASED QUESTIONS

We wished to check if the high accuracies obtained by the proposed model is an outcome of the limited encoding of question templates in our RetVQA dataset. Hence, we evaluate our proposed framework (MI-BART) on 2K paraphrased questions in the test set with our strongest baseline (VLP). A subset of test set questions is paraphrased using the BART [115] paraphrase model (Large)[1]. We show the results in Table 5.10. Our results show that our proposed framework MI-BART is robust to the question templates used to create the data and performs fairly well on the paraphrased questions when compared against another transformer-based baseline VLP. We show a qualitative sample in Figure 5.8, where our model still gives the correct answer to the paraphrased question, whereas VLP entirely generates a wrong answer.

[1] `https://huggingface.co/eugenesiow/bart-paraphrase`

Question: Do pizza and table share the same shape?

Paraphrased Question: Do table and pizza have the same shape?

Context:


Ground truth answer: No, they do not share the same shape.

	MI-BART Answers	VLP Answers
On original question:	No, they do not share the same shape.	No, they do not share the same shape.
On paraphrased question:	No, they do not share the same shape.	Yes , they have same shape.

Figure 5.8: Paraphrased question answering using MI-BART (Ours) and VLP from RetVQA dataset.

	VLP			MI-BART		
Question-type	Acc.	F	F×A	Acc.	F	F×A
Paraphrased	47.8	54.5	37.5	***58.8***	***71***	***51.2***
Non-paraphrased	60.9	65.6	52.7	***76.4***	***75.4***	***67.4***

Table 5.10: Performance on RetVQA with paraphrased questions with the retrieved images.

5.5.8 MORE QUALITATIVE EXAMPLES

We show some examples of predictions using MI-BART (ours) and VLP (best baseline) from RetVQA dataset in Figure 5.9 and Figure 5.10.

5.6 CONCLUSION AND FUTURE SCOPE

In this work, we introduced the RetVQA task. We proposed a unified Multi Image BART model to answer the question from the retrieved images using our relevance encoder. Our proposed framework shows promising improvements over the baselines. We have also performed several ablations to further understand the importance of various modules in the proposed framework. In the future, we would like to explore stronger retrieval models and QA on a large pool of images. We firmly believe RetVQA will pave the way for further research avenues in a broader theme of web image QA.

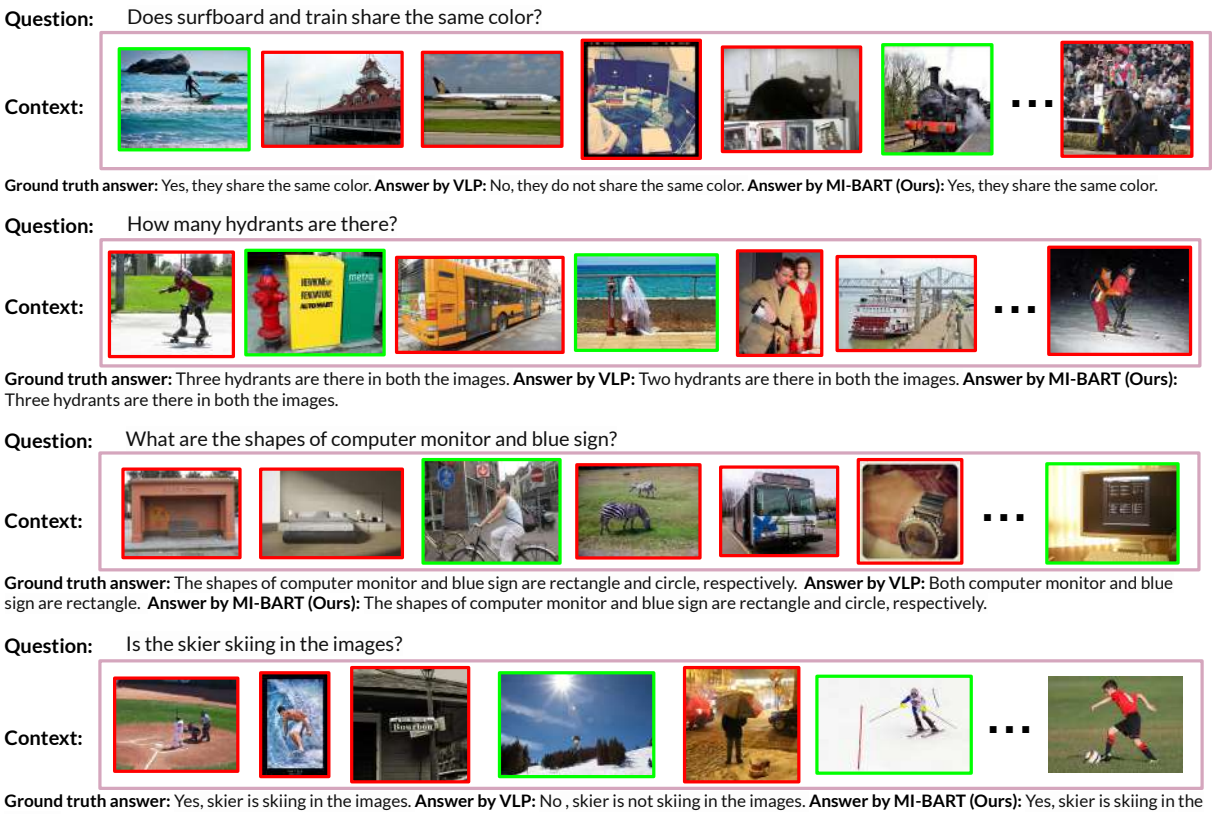


Figure 5.9: Few more selected example predictions using MI-BART (ours) and VLP (best baseline) from RetVQA dataset.

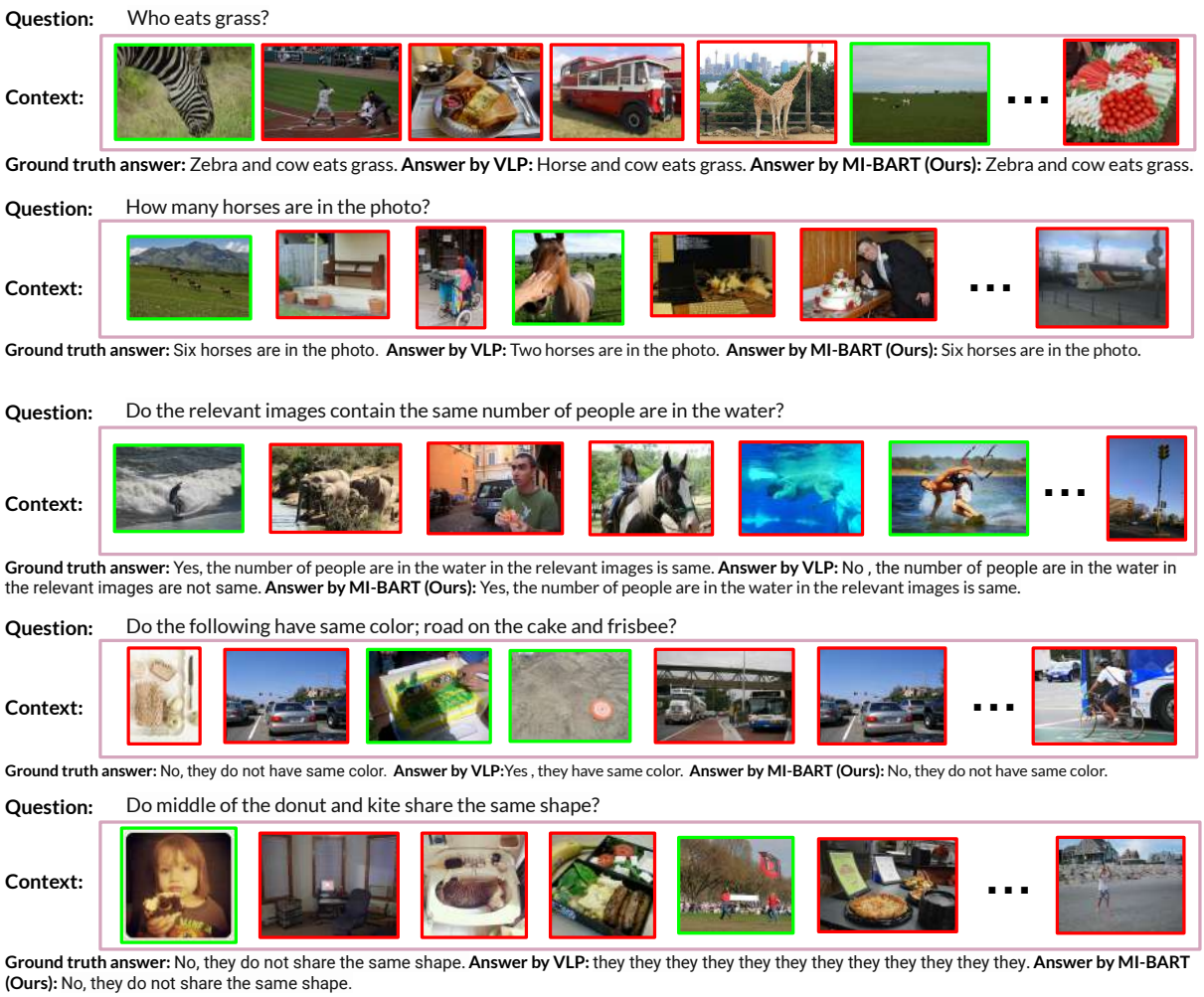


Figure 5.10: Few more selected example predictions using MI-BART (ours) and VLP (best baseline) from RetVQA dataset.

6

Augmenting VLMs with Multimodal Knowledge for Image Captioning

In this chapter, we extend knowledge augmentation to the task of domain-specific image captioning, focusing on fashion. Unlike natural scene captioning, fashion image captioning requires attribute-level precision and domain-specific terminology to describe subtle details such as fabric type, neckline, and silhouette. Supervised solutions are impractical in this domain due to the rapidly evolving nature of fashion inventories, motivating the need for training-free and adaptable approaches. To address this challenge, we propose RA-CoA (Retrieval-Augmented Chain of Attributes), a novel, training-free framework that grounds caption generation in retrieved multimodal knowledge. Specifically, RA-CoA first retrieves visually similar fashion exemplars from a curated product knowledge base, extracts relevant attribute types, and then prompts frozen vision–language models to infer attribute values independently. The final caption is generated by conditioning on the structured attribute-value pairs, ensuring interpretability and faithfulness. Extensive experiments on the FashionGen dataset with both open and closed-source models show that RA-CoA significantly outperforms prior training-free paradigms such as in-context learning and chain-of-thought prompting, while maintaining strong zero-shot adaptability for real-world deployment.

6.1 INTRODUCTION

With the rise of AI-driven personalization in e-commerce, Fashion Image Captioning (FIC) has emerged as a crucial task for enhancing user experience and improving product searchability[1]. Unlike the well-established task of natural scene image captioning [76, 110, 129], fashion image captioning (FIC) [180, 249] necessitates *fine-grained visual reasoning* to accurately capture nuanced product attributes such as stylistic patterns, closure mechanisms, collar and sleeve silhouettes, etc. Furthermore, the rapidly evolving nature of fashion inventories demands *comprehen-*

[1] According to Google consumer research, approximately 85% of shoppers report that accurate product information play an important role in deciding which brand or retailer to buy from, highlighting the need for accurate attribute-level product descriptions in e-commerce [61].

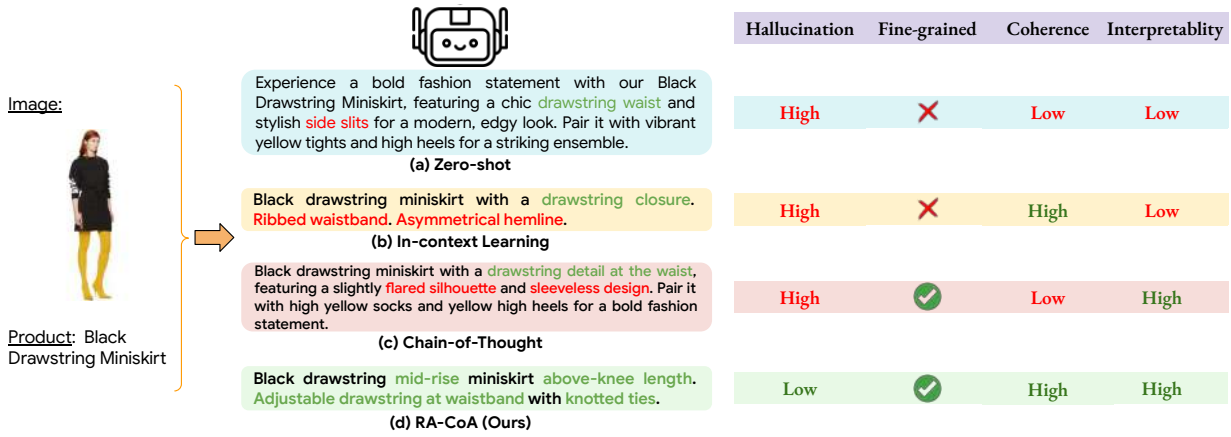


Figure 6.1: Task Overview: Given a fashion image and the product name, existing training-free prompting paradigms may fail to provide attribute-rich, coherent, hallucination-free captions. Captioning with (a) Zero-shot: Produces only a generic description, missing key fine-grained attributes. (b) In-context learning [47]: Improves coherence via exemplars but tends to copy attributes. (c) Chain-of-Thought [235]: Performs stepwise reasoning yet fails in capturing all fine-grained attributes correctly. (d) RA-CoA (Ours): A retrieval-augmented chain-of-attributes framework that aggregates key attributes from similar catalog items and sequentially predicts their values. The predicted attributes, combined with ICL exemplars, guide caption generation toward coherent and attribute-faithful captions.

sive knowledge of domain-specific terminology, including emerging trends, seasonal styles, and an ever-expanding attribute vocabulary. In this dynamic landscape, curating annotated datasets and retraining domain-specific supervised captioning models from time-to-time becomes prohibitively expensive and operationally impractical. This challenge highlights the critical need for training-free frameworks that deliver adaptability, scalability, and efficiency while maintaining high-quality product descriptions.

Instruction-tuned vision-language models (VLMs) [12, 32, 132, 133, 155, 252, 271] present a compelling solution. Their capacity to generate fluent, naturalistic descriptions without task-specific training makes them attractive candidates for fashion image captioning. However, these general-purpose models fundamentally lack the attribute-level precision required for fashion domains, frequently hallucinating or misidentifying fine-grained attributes critical to product identity. Consequently, their zero-shot outputs remain inadequate for e-commerce applications that demand accurate and faithful descriptions for automated product cataloging, fine-grained fashion retrieval, and personalized recommendation systems.

Recent advances demonstrate that vision-language models can benefit substantially from structured prompting strategies such as in-context learning [47, 72] and chain-of-thought reasoning [235]. These techniques have proven effective in eliciting more accurate and coherent responses from VLMs across various tasks. However, in the fashion domain, even these advanced prompting strategies struggle to ground fine-grained attributes reliably, as illustrated in Figure 6.1. *In-context learning*, while effective at improving stylistic coherence through exemplar-based prompting, suffers from attribute hallucination. Models tend to erroneously copy or interpolate attributes from provided examples rather than faithfully grounding them in the target image. *Chain-of-thought*

prompting, despite enabling stepwise reasoning, still struggles with comprehensive fine-grained attribute identification, as VLMs lack the visual acuity to reliably distinguish subtle fashion-specific details (e.g., differentiating between a mandarin collar and a band collar) when reasoning from scratch. The core challenge lies in the fact that directly querying VLMs for fine-grained attributes, often results in hallucinated or imprecise responses due to insufficient visual grounding and the absence of domain-specific knowledge.

To address these limitations, we propose RA-CoA (**R**etrieval **A**ugmented **C**hain **o**f **A**ttributes), a training-free framework that synergistically combines the zero-shot reasoning capabilities of VLMs with explicit retrieval-based grounding and structured attribute reasoning. Unlike direct caption generation or conventional prompting approaches, RA-CoA decomposes the captioning task into two interpretable and complementary stages: (i) *Attribute Set Retrieval*, where we leverage a product knowledge base (ProductKB) to retrieve product-aware visually similar examples and extract candidate attributes; and (ii) *Chain-of-Attribute Reasoning*, where the model performs attribute-by-attribute visual grounding, inferring the value of each candidate attribute from the image, before synthesizing these verified attributes into a coherent, factually grounded caption. By introducing retrieval-augmented attributes prior to reasoning, RA-CoA mitigates the hallucination problems inherent in in-context learning (by providing explicit attribute candidates rather than exemplar captions to copy from) and chain-of-thought (by constraining the reasoning space to visually plausible attributes). This explicit decomposition makes fine-grained visual reasoning interpretable and verifiable, enabling the model to focus its attention on relevant visual regions for each attribute. Importantly, RA-CoA is modular, model-agnostic, and operates with frozen VLMs, requiring no fine-tuning which enables seamless adaptation to evolving fashion inventories and emerging attribute taxonomies. Through this structured retrieval-augmented approach, RA-CoA achieves a balance between zero-shot flexibility and fine-grained precision, producing accurate and interpretable captions with minimal hallucination.

The first contribution of this chapter is RA-CoA, a novel training-free and model-agnostic framework for fashion image captioning that decomposes caption generation into retrieval-augmented attribute identification followed by explicit chain-of-attribute reasoning. This design enhances fine-grained visual grounding while remaining flexibly applicable across both open and closed-source VLMs without any task-specific fine-tuning. The second contribution is a comprehensive evaluation across a range of prompting paradigms, including vanilla zero-shot, in-context learning, and implicit and explicit chain-of-thought prompting. Through automatic metrics and user studies, we demonstrate that RA-CoA consistently improves attribute precision and semantic coherence over all baselines. The third contribution is an assessment of the robustness and real-world applicability of RA-CoA, benchmarked against a state-of-the-art supervised fashion captioning model on held-out, web-curated data. Despite operating without any task-specific training, RA-CoA achieves superior performance, highlighting its ability to generalize under distribution shift and its suitability for scalable deployment in dynamic e-commerce settings.

The remainder of this chapter is organized as follows. In Section 6.2, we review related work on fashion image captioning and large vision–language models. Section 6.3 presents our proposed framework Retrieval-Augmented Chain-of-Attributes (RA-CoA) (Section 6.3.2), including

the construction of the ProductKB (Section 6.3.1). Experimental setup, dataset, evaluation metrics, results, and ablation studies are reported in Section 6.4, followed by comparisons with prior state-of-the-art methods. Finally, Section 6.5 concludes the chapter, Section 6.6 discusses limitations, and Section 6.7 outlines ethical considerations.

6.2 RELATED WORK

6.2.1 FASHION IMAGE CAPTIONING

Fashion image captioning has emerged as a specialized task within the broader field of vision-and-language understanding, and presents unique challenges requiring models to generate accurate, attribute-rich captions that capture both visual appearance and semantic fashion concepts. Early work in this domain focused on fashion retrieval and attribute prediction with datasets such as DeepFashion [135], Fashion200K [70] and FasionIQ [233]. These datasets provide rich annotations for fashion-specific attributes including garment categories, colors, patterns, and styles. However, these datasets are designed for retrieval-centric tasks with relative captions, making them less suited for evaluating free-form captioning. More recent approaches have explored fashion image captioning with FashionGen [180] dataset, which offers comprehensive, attribute-rich product descriptions specifically designed for caption generation. Several models have been developed for this task. [125] pioneered the integration of explicit attribute detection with visual attention mechanisms. The recent rise of large-scale vision-language pre-training has significantly advanced the field with specialized fashion models. FashionBERT [55] adapts masked language modeling for fashion attribute prediction. KaleidoBERT [273] introduces adaptive fashion-text pre-training with dynamic masking strategies. Methods like FashionVLP [59] leverage vision-language pre-training specifically adapted for fashion domains to improve attribute-aware description generation. FashionViL [68] proposes contrastive learning between fashion images and text descriptions. FAME-ViL [69] incorporates multi-modal entity understanding for fashion, and UniFashion [266] presents a unified framework for multiple fashion tasks including captioning and retrieval. However, most existing approaches require substantial task-specific fine-tuning on fashion datasets, creating significant computational overhead and limiting their adaptability to the evolving vocabulary and products in the catalogs. In contrast to these training-intensive paradigms, our proposed RA-CoA is a novel model-agnostic training-free approach that leverages retrieval-augmented structured reasoning to generate attribute-grounded captions in a zero-shot setting.

6.2.2 RETRIEVAL-AUGMENTED IMAGE CAPTIONING

Retrieval-Augmented Generation (RAG) was initially proposed to ground language generation in external knowledge, reducing reliance on parametric memory and improving factuality [117]. This idea has since been adopted in image captioning through retrieval-augmented *training*, where

models learn to condition caption generation on retrieved samples. Early retrieval-augmented image captioning models such as EXTRA [175] jointly encode images and retrieved captions to learn retrieval-conditioned generation. Subsequent analysis has shown that such models are sensitive to retrieval noise and the ordering of retrieved samples [124]. SmallCap [176] explores a lighter alternative by prompting a largely frozen language model with retrieved captions and training only lightweight components to internalize retrieval signals. EVCap [119] further extends this training-based paradigm by introducing an external visual–name memory to support open-world captioning. More recently, retrieval has also been explored for zero-shot captioning. MeaCap [258] addresses open-domain zero-shot image captioning by retrieving textual memory and extracting salient concepts to guide generation. This formulation emphasizes concept completion and semantic plausibility in natural scene images, rather than faithful prediction of structured visual attributes, which is crucial for domains such as fashion. In contrast to these works, RA-CoA targets attribute-faithful captioning in structured domains and is *training-free* by design. It operates in a zero-shot setting using off-the-shelf vision–language models, without any fine-tuning or task-specific supervision. Rather than conditioning generation on retrieved captions or attribute values, RA-CoA retrieves only attribute keys and infers attribute values directly from the VLM's internal knowledge through structured reasoning.

6.2.3 ZERO-SHOT PROMPTING STRATEGIES FOR VISION LANGUAGE MODELS

Large-scale vision-language models (VLMs) [12, 32, 40, 120, 133, 155, 252, 268, 271] have demonstrated strong performance transfer across diverse tasks. Zero-shot learning paradigm with VLMs leverages pre-trained knowledge without task-specific fine-tuning. Beyond basic zero-shot prompting, several advanced prompting paradigms have emerged to elicit optimal performance from VLMs. In-context learning (ICL) [47, 72] extends zero-shot capabilities by providing few-shot examples as demonstrations, enabling models to adapt to new tasks without parameter updates. Chain-of-thought (CoT) [228, 235] prompting decomposes complex reasoning into intermediate steps, initially proposed for language models and later adapted to vision-language tasks. However, when applied to the task of fashion image captioning, current zero-shot prompting strategies struggle with attribute-level precision, often generating generic descriptions that miss intricate design elements of the fashion product. In-context learning frequently suffers from memorization bias [60], where models reproduce content from reference samples rather than adapting to the specific visual attributes of the target image. Chain-of-thought prompting shows limited effectiveness in attribute-centric captioning scenarios where structured domain knowledge and systematic visual analysis are required. These limitations underscore the need for more sophisticated prompting paradigms that can effectively guide VLMs to generate detailed and accurate captions for fashion products. Our work in this chapter addresses this gap by combining retrieval with explicit attribute reasoning to guide zero-shot captioning in VLMs.

6.3 METHODOLOGY

In this section, we present RA-CoA (**R**etrieval **A**ugmented **C**hain **o**f **A**ttributes), a novel training-free approach for fashion image captioning (FIC). RA-CoA works in a step-wise manner by first identifying attributes, then assigning values, and finally composing a coherent description. This stepwise attribute-centric reasoning enables fine-grained understanding prior to generating captions.

Problem Statement: We define fashion image captioning as an attribute-driven process. Given an image of a fashion product **I** and its name **N**, the goal is to generate a caption **C** using a set of attribute-value pairs $\mathcal{A} = \{(a_i, v_i)\}_{i=1}^{n}$ for the specific product, where a_i denotes an attribute type, e.g., sleeve type, neckline and v_i its corresponding value, e.g., half-sleeve, v-neck. RA-CoA models this in three structured stages: ***(i) Identification of attributes***: Determine the relevant attributes $\{a_1, a_2, \cdots, a_n\}$ for the product image **I**, ***(ii) Assignment of values***: For each attribute a_i, reasoning about its appropriate value v_i, and ***(iii) Caption generation***: Synthesizing caption C_t by integrating these attribute-value insights. This reasoning-based decomposition enables fine-grained visual attribute reasoning for fashion image captioning.

6.3.1 PRODUCTKB CONSTRUCTION

We construct a structured Product Knowledge Base (ProductKB) of product images, their captions, and attribute-value tabular data to support retrieval-augmented generation. We leverage the training dataset of FashionGen [180], $\mathcal{D} = \{(I_j, N_j, C_j)\}_{j=1}^{m}$ of m triplets for this purpose. Each triplet consists of a fashion product image (I_j), product name (N_j) and its caption (C_j). Next, we extract structured attribute-value pairs from the captions using the Llama-3.2-3B-Instruct model (M) [63]: $M_{LLM}(C_j) \rightarrow \mathcal{A}_j = \{(a_j, v_j)\}_{i=1}^{m}$, guided by this prompt:

Prompt used for ProductKB construction

You are given features $\{C_j\}$ for an e-commerce product, identify the keys for these features and output it in json format. Keys should be very short, precise and specific, such that if given only the key and product image, its feature value can be predicted. Keys should cover all the features of the product. Break down the features into multiple key-value pairs where appropriate. Do not create sub-keys or sub-values. Both key and value should be str datatype. Strictly output only the json.
Assistant: $\{\mathcal{A}_j\}$.

Popular fashion datasets such as FashionGen [180] often depict models wearing multiple products within a single image, which introduces ambiguity when associating captions and attributes with a specific product. To ensure that each entry in the ProductKB corresponds precisely to the intended product, we leverage the product name N_j to isolate the target product from I_j. Specifically, we use the Florence-2 [238] model (f) to detect and crop the image region corresponding to the product, resulting in a focused crop: $I_j^c = f(I_j, N_j)$. This step ensures that

the stored representation in ProductKB is accurately aligned with the product described in the accompanying caption and attribute annotations.

Our final ProductKB ($\mathcal{P}$) comprises of $\{(I_j^c, N_j, C_j, \mathcal{A}_j)\}_{j=1}^m$, where each entry includes the cropped image focused on the target product I_j^c, the product name N_j, the expert-written caption C_j, and the obtained structured attribute-value pairs $\mathcal{A}_j$. This curated knowledge base supports the efficient retrieval of visually similar products with shared attribute schemas, facilitating accurate and attribute-based caption generation.

6.3.2 RA-COA: RETRIEVAL-AUGMENTED CHAIN-OF-ATTRIBUTES

RA-CoA integrates a retrieval mechanism with a structured reasoning chain for fashion image captioning. The pipeline proceeds in multiple steps: retrieving visually similar products, identifying relevant attribute keys, estimating their values, and finally composing a caption grounded in these attributes. The following subsections elaborate on each stage of the pipeline.

IMAGE EMBEDDING AND RETRIEVAL

The first step involves retrieving relevant attribute knowledge from the ProductKB based on visual similarity. Given a query image $\mathbf{I}$ and its associated product name $\mathbf{N}$, we use the Florence-2 [238] model (f) to isolate the target product and obtain the cropped image $\mathbf{I}^c = f(\mathbf{I}, \mathbf{N})$. We then compute its visual embedding e using a pre-trained CLIP [169] vision encoder (E_V), i.e., $e = E_V(\mathbf{I}^c)$. In a similar manner, we encode all product images in the ProductKB using the same encoder and retrieve the top-K most similar items based on cosine similarity. The resulting set is denoted as $\mathcal{P}_K = \{(I_i^c, N_i, C_i, \mathcal{A}_i)\}_{i=1}^K$. While attributes $\{\mathcal{A}\}_{i=1}^K$ are used to guide value prediction, $\{(I_i^c, N_i, C_i)\}_{i=1}^K$ are used later in the pipeline as in-context exemplars during caption generation.

ATTRIBUTE EXTRACTION

Once top-K similar products are retrieved from the ProductKB, we extract and aggregate the unique attributes (Eq. 6.1).

$$\mathcal{K} = \bigcup_{i=1}^{K} a | (a, v) \in \mathcal{A}_i \tag{6.1}$$

We extract only attribute keys ($\mathcal{K}$), excluding their values, as similar products often share attribute types (e.g., sleeve type, neckline) but differ in specifics. This avoids propagating incorrect values while retaining guidance on which attributes to predict, ensuring value inference remains grounded in the query image.

CHAIN-OF-ATTRIBUTES (COA)

Once the relevant attributes $\mathcal{K} = \{a_1, a_2, ..., a_n\}$ are identified through retrieval, we implement a Chain-of-Attributes reasoning process. Analogous to chain-of-thought reasoning, CoA decomposes fashion understanding into attribute-wise focus steps. For each attribute $a_i \in \mathcal{K}$, we query the VLM to determine its corresponding value:

$$v_i = M_{VLM}(p_{value}(\mathbf{I}, \mathbf{N}, a_i)) \tag{6.2}$$

where $p_{value}(.)$ is a prompt template for attribute value generation:

Prompt used for attribute value generation

<image (**I**) *>*
Given the image of a model wearing the e-commerce product - **N**, how/what is the a_i of the product? Strictly answer as single word or phrase.
Assistant: $\{v_i\}$.

This process results in a structured attribute-value set $\tilde{\mathcal{A}} = \{(a_1, v_1), (a_2, v_2), ..., (a_n, v_n)\}$ where each attribute receives focused attention. By isolating individual attribute queries, CoA reduces ambiguity and cognitive load, leading to more accurate and interpretable value predictions.

CAPTION GENERATION

Once we obtain the comprehensive set of attribute-value pairs $\tilde{\mathcal{A}}$, the final caption is generated by conditioning the vision-language model on the original query image **I**, product name **N**, and the top-K retrieved examples $\mathcal{P}_K = \{(I_i^c, N_i, C_i)\}_{i=1}^K$. These retrieved entries serve as in-context exemplars to guide both the structure and style of the generated description. The caption is produced as:

$$\tilde{\mathbf{C}} = M_{VLM}(p_{caption}(\mathbf{I}, \mathbf{N}, \tilde{\mathcal{A}}, \mathcal{P}_K)). \tag{6.3}$$

where $p_{caption}(.)$ is a prompt template for caption generation as shown below:

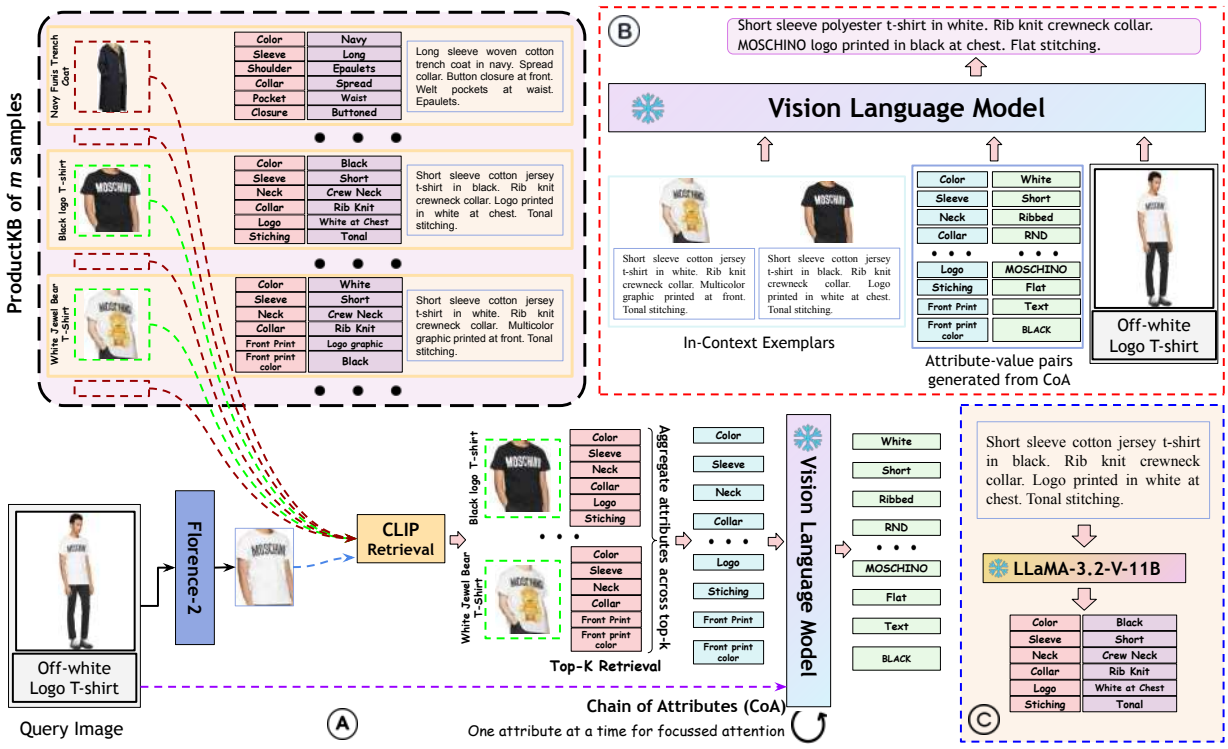


Figure 6.2: Overview of RA-CoA. Ⓐ *Retrieval-augmented Chain-of-Attributes* - For a query fashion image, we crop the target product region using Florence-2, retrieve top-K similar products from ProductKB, and aggregate unique attributes. Each attribute is paired with the image and product name and fed to VLM for value prediction. Ⓑ *Caption Generation* - The resulting attribute-value pairs and in-context exemplars (retrieved products) enable the VLM to generate fine-grained coherent captions. Ⓒ *ProductKB Construction* - We construct our ProductKB by extracting structured attribute-value pairs from FashionGen captions using LLaMA-3.2-V-11B.

Prompt used for caption generation

$<image\ (\mathbf{I})>$
Given the image of a model wearing the e-commerce product $\mathbf{N}$, and its attribute-value pairs $\tilde{\mathcal{A}}$, write a concise and fluent caption for the specific product, incorporating the given attributes.
You may refer to the following in-context examples for style and structure:
Example captions:
$<image\ (I_1^c)>$ Product name: N_1, Caption: C_1 $<image\ (I_2^c)>$Product name: N_2, Caption: C_2. Output only the final caption.
Assistant: $\tilde{\mathbf{C}}$

This final synthesis step integrates the individually analyzed attributes into a holistic description of the fashion item. The complete RA-CoA pipeline is illustrated in Figure 6.2 and summarized in Algorithm 1.

Algorithm 1 RA-CoA Pipeline

Input: Query image $\mathbf{I}$, Query product name $\mathbf{N}$, ProductKB: $\mathcal{P}$ and Vision-Language Model M_{VLM}.

Output: Generated caption $\tilde{\mathbf{C}}$

$\mathbf{I}^c = f(\mathbf{I}, \mathbf{N})$ ▷ Get product-of-interest crop.

$e \leftarrow E_V(\mathbf{I}^c)$ ▷ Generate query embedding.

$\mathcal{P}_K \leftarrow$ Retrieve top-K similar products from $\mathcal{P}$ using $sim(e, e_j)_{j=1}^{m}$.

$\mathcal{P}_K = \{(I_j^c, N_j, C_j, \mathcal{A}_j)\}_{j=1}^{K}$ ▷ Each retrieved product has image, name, caption and attributes.

$\mathcal{K} \leftarrow \bigcup_{(I_j^c, N_j, C_j, \mathcal{A}_j) \in \mathcal{P}_K} \{a \mid (a, v) \in \mathcal{A}_i\}$ ▷ Extract unique keys

$\tilde{\mathcal{A}} \leftarrow \emptyset$ ▷ Initialize attribute set for query image.

for each $a_j \in \mathcal{K}$ ***do***

  $v_j \leftarrow M_{VLM}(p_{value}(\mathbf{I}, \mathbf{N}, a_j))$ ▷ Generate attribute value.

  $\tilde{\mathcal{A}} \leftarrow \tilde{\mathcal{A}} \cup \{(a_j, v_j)\}$ ▷ Add to attribute set.

end for

$\tilde{\mathbf{C}} \leftarrow M_{VLM}(p_{caption}(\mathbf{I}, \mathbf{N}, \tilde{\mathcal{A}}, \mathcal{P}_K))$ ▷ Generate final caption.

return $\tilde{\mathbf{C}}$

6.4 EXPERIMENTS AND RESULTS

6.4.1 DATASET

We conduct our experiments on the FashionGen [180] dataset, which contains roughly 300K high-resolution (1360 × 1360) images of e-Commerce products in the fashion domain, each accompanied by rich metadata and expert-written descriptions that capture fine-grained attributes of the product. In our work, we leverage only three fields: product images, names/titles, and ground-truth captions. FashionGen images often feature a model wearing multiple fashion product - clothing accessorized with bags, shoes, or jewelry. Hence, the target product may not be immediately obvious. Thus, we utilize product names as a guiding cue to focus the VLM's attention towards the intended item. The expert-written descriptions then serve as ground-truth captions for evaluation. From the FashionGen training split, we sample 60K products to build our retrieval knowledge base (KB), which the RA-CoA module uses exclusively for product name-aware image retrieval; these KB samples are never used to train or validate the VLMs. We then assessed the captioning performance on the 7K test samples, comparing the generated captions directly against the FashionGen captions.

6.4.2 VLMS USED

We evaluate a diverse set of state-of-the-art vision–language models to benchmark the effectiveness of our approach across varying model scales and capabilities. To ensure comprehensive coverage, we include (a) ***Small VLMs*** *(models with parameters <4B)*: (i) TinyLLaVA (3B) [268], and (ii) Qwen-2.5VL (3B) [12] and (b) ***Large VLMs*** *(models with parameters >4B)*: (i) Qwen-2.5VL (7B)

and (ii) InternVL2 (8B) [32], and (c) ***Closed-source model***: GPT-4o [155].

VLM PARADIGMS

We compare our proposed framework RA-CoA against standard VLM prompting techniques such as ***(i) Direct prompting*** (Zero-shot): where VLM is prompted with an instruction to generate the description directly given an image. ***(ii) Few-shot prompting*** (In-context Learning - ICL): Here, we follow the standard few-shot in-context prompting technique, where we retrieve K random product images along with their tabular information, and add them to the prompt as in-context exemplars. ***(iii) Implicit Chain-of-thought*** (CoT-i) prompting: We prepend a generic "let's think step by step" cue to the zero□shot prompt, encouraging the VLM to internally structure its reasoning to identify fine-grained attribute-value pairs, without exposing intermediate steps in the output., ***(iv) Explicit Chain-of-Thought*** (CoT-e): Here, we explicitly ask the model to enumerate each reasoning step, itemizing attributes one by one before producing the final caption, thereby surfacing its internal chain of thought. The detailed prompts used for each of these settings are provided below.

6.4.3 PROMPTS USED FOR DIFFERENT PARADIGMS

In this section, we enlist the prompts used for various VLMs under different training-free paradigms used as baselines. These prompts remain consistent across all VLMs including the closed-source GPT-4o.

Prompt used under zero-shot paradigm

<image(**I**)>
Given the image of a model wearing the e-commerce product - **N**, write a concise caption for the specific product, incorporating its fine-grained attributes.
Assistant: $\tilde{\mathbf{C}}$

Prompt used under implicit CoT paradigm

<image(**I**)>
Given the image of a model wearing the e-commerce product - **N**, you have to write a concise caption for the specific product. But think step by step. First, carefully observe the **N** in the image and identify all the fine-grained visual attributes of the specific product. Then, using the identified attributes, write a concise attribute-aware caption for the specific product. Only output the final caption.
Assistant: $\tilde{\mathbf{C}}$

Prompt used under explicit CoT paradigm

<image(**I**)>
Given the image of a model wearing the e-commerce product - **N**, you have to write a concise caption for the specific product. But think step by step. First, carefully observe the **N** in the image and list all the fine-grained visual attributes of the specific product. Then, using the identified attributes, write a concise attribute-aware caption for the specific product. Output both the attributes and the final caption strictly in json.
Assistant: $\tilde{C}$

Prompt used under the ICL paradigm

<image(**I**)>
Given the image of a model wearing the e-commerce product - **N**, write a concise caption for the specific product, incorporating its fine-grained attributes. Take reference from the provided examples.
Example1: <image (I_1^c)> Product name: N_1, Caption: C_1
Example2: <image (I_2^c)>Product name: N_2, Caption: C_2
Output only the final caption.
Assistant: $\tilde{C}$

Under the ICL paradigm, all baselines except TinyLLaVA-3B [268] are fed with two in-context samples, retrieved randomly from the training set. TinyLLaVA cannot accept multi-image inputs, hence we only feed the product name and captions of the retrieved in-context samples in TinyLLaVA.

6.4.4 METRICS

TRADITIONAL METRICS

To measure the captioning performance of these VLMs, we utilize standard image-captioning metrics such as BLEU [159], ROUGE [128] and METEOR [14], where higher values for all the scores are desired.

LLM-AS-JUDGE

While the above metrics provide a general sense of fluency and lexical overlap with references, they often fail to capture whether the caption accurately describes the most relevant visual attributes, particularly in the fashion domain where small attribute errors can significantly mislead. To address this gap, we introduce a task-specific LLM-based evaluation protocol inspired by human verification, which measures how faithfully a caption covers gold-standard attribute-value pairs of a fashion product. We use LLaMA-3.2-V-11B model as a judgment engine to determine

Model (# params)	Paradigm	BLEU-1	Avg. BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR	LLM-Judge
TinyLLaVA (3B)	Zero-shot	12.3	3.5	17.1	2.4	11.4	12.2	17.2
	CoT-i	11.0	3.3	18.6	2.7	12.6	11.0	10.1
	CoT-e	7.8	2.3	12.8	1.8	8.7	8.6	8.1
	ICL	13.1	4.8	24.5	4.9	18.7	13.7	36.5
	RA-CoA	***22.1***	***11.50***	***32.5***	***10.8***	***25.2***	***24.6***	***50.9***
	Δ	+9.8	+8.0	+15.4	+8.4	+13.8	+12.4	+33.7
Qwen-2.5VL (3B)	Zero-shot	8.0	2.2	19.4	2.8	12.7	10.4	31.7
	CoT-i	4.2	1.2	22.8	3.4	14.9	10.0	29.3
	CoT-e	9.1	2.7	20.8	3.4	13.3	10.4	20.9
	ICL	26.2	13.5	29.7	9.7	23.3	26.5	49.8
	RA-CoA	***26.8***	***14.5***	***35.3***	***11.7***	***28.3***	***28.6***	***55.2***
	Δ	+18.8	+12.3	+15.9	+8.9	+15.6	+18.2	+23.5
Qwen-2.5VL (7B)	Zero-shot	8.2	2.2	18.7	2.6	11.6	10.9	33.2
	CoT-i	6.4	1.7	19.0	2.9	12.1	10.2	32.5
	CoT-e	7.5	2.1	20.4	3.4	12.9	10.4	23.0
	ICL	14.5	8.2	29.3	8.1	20.8	19.2	50.0
	RA-CoA	***31.4***	***17.7***	***39.2***	***14.2***	***32.3***	***32.4***	***57.7***
	Δ	+23.2	+15.5	+20.5	+11.6	+20.7	+21.5	+24.5
InternVL2 (8B)	Zero-shot	11.3	3.2	16.2	2.3	11.0	13.8	25.9
	CoT-i	13.7	4.0	16.6	3.0	12.2	15.7	38.4
	CoT-e	13.9	4.2	20.8	3.1	13.9	13.5	26.1
	ICL	30.2	17.5	35.2	13.4	28.7	31.8	58.9
	RA-CoA	***38.6***	***23.7***	***41.0***	***17.4***	***36.0***	***38.1***	***61.1***
	Δ	+27.3	+20.5	+24.8	+15.1	+25.0	+24.3	+35.2
GPT-4o	Zero-shot	10.6	2.9	17.7	2.6	11.4	12.6	34.2
	CoT-e	22.8	14.3	30.2	12.8	23.9	25.4	37.5
	ICL	46.6	32.1	53.8	28.6	46.3	50.3	60.7
	RA-CoA	***63.2***	***48.8***	***69.3***	***44.2***	***62.9***	***67.6***	***76.5***
	Δ	+52.6	+45.9	+51.6	+41.6	+51.5	+55.0	+42.3

Table 6.1: Comparison of RA-CoA against different prompting paradigms across VLMs of varying scales. Δ (gray rows) represents the gain of RA-CoA over Zero-shot.

whether each attribute-value pair of the oracle data (described in section 6.4.7) is correctly captured in the caption. Each evaluation is phrased as a binary Yes/No question. Below is the prompt template used for this metric calculation:

Prompt used for LLM judge

Caption: {$caption(\mathbf{C})$}
Attribute: {$attribute(a)$}
USER: You are given a generated caption, and the ground-truth attributes of an e-commerce fashion product. Does the caption correctly capture the value of the above attribute? Answer 'Yes' or 'No'.
ASSISTANT: {Yes}.

We then count the occurence of 'Yes' and get the average score for number of attributes covered per sample.

6.4.5 IMPLEMENTATION DETAILS

We use PyTorch [162] and the HuggingFace Transformers library [231] for majority of our implementation. For retrieval over the product knowledge base (ProductKB) (Section 6.3.2), we employ FAISS [49] for efficient nearest-neighbor search. Visual embeddings for all ProductKB entries are pre-computed using a CLIP vision encoder [169]. Unless otherwise specified, we use $K{=}1$ for ablation studies in Section 6.4.7. For the VLMs used in our experiments, we rely on the authors' official code repositories or publicly available HuggingFace implementations, prioritizing reproducibility. All VLMs are used in a frozen, inference-only setting without any fine-tuning. Unless stated otherwise, all ablation studies are conducted using InternVL2-8B as the underlying VLM. All experiments are conducted on a single machine equipped with three NVIDIA A6000 GPUs (48GB memory each).

6.4.6 RESULTS AND DISCUSSION

Table 6.1 presents quantitative comparison of our method RA-CoA with other training-free paradigms across varied-scale VLMs. RA-CoA consistently outperforms all baselines, including proprietary GPT-4o, validating its effectiveness and robustness for training-free fashion image captioning. Zero-shot methods yield poor METEOR scores (10.4–13.8), often producing generic descriptions that miss key fashion attributes (see Figure 6.1). Chain-of-Thought reasoning shows limited or even degraded performance, likely due to models describing the entire image rather than the product, leading to attribute omission or hallucination. In-Context Learning improves METEOR by 1.5–18 points over zero-shot, yet still struggles with complete attribute coverage. In contrast, RA-CoA achieves the highest performance across all models. For TinyLLaVA (3B), it doubles the METEOR score (24.6 vs. 12.2), improving 79% over ICL. Gains scale with model size. RA-CoA improves METEOR by 12.4 points for TinyLLaVA (3B) and 24.3 for InternVL2 (8B) over zero-shot. Larger models perform better overall, with InternVL2 (8B) achieving the best open-source results (METEOR: 38.1, LLM Judge: 61.1). GPT-4o, when used with RA-CoA, achieves the highest scores (METEOR: 67.6, LLM Judge: 63.5), though its zero-shot performance is on par with smaller open-source models. These findings affirm that RA-CoA's structured, attribute-focused approach enables more accurate and interpretable captions, with gains increasing alongside model scale.

Qualitative examples in Figure 6.3 further illustrate RA-CoA's superior caption fidelity across diverse fashion categories. The generated descriptions consistently capture fine-grained product attributes with high accuracy across different product categories.

6.4.7 ABLATIONS AND ANALYSIS

We conduct the following ablation studies to validate the design choices and quantify the impact of different components of our method RA-CoA:

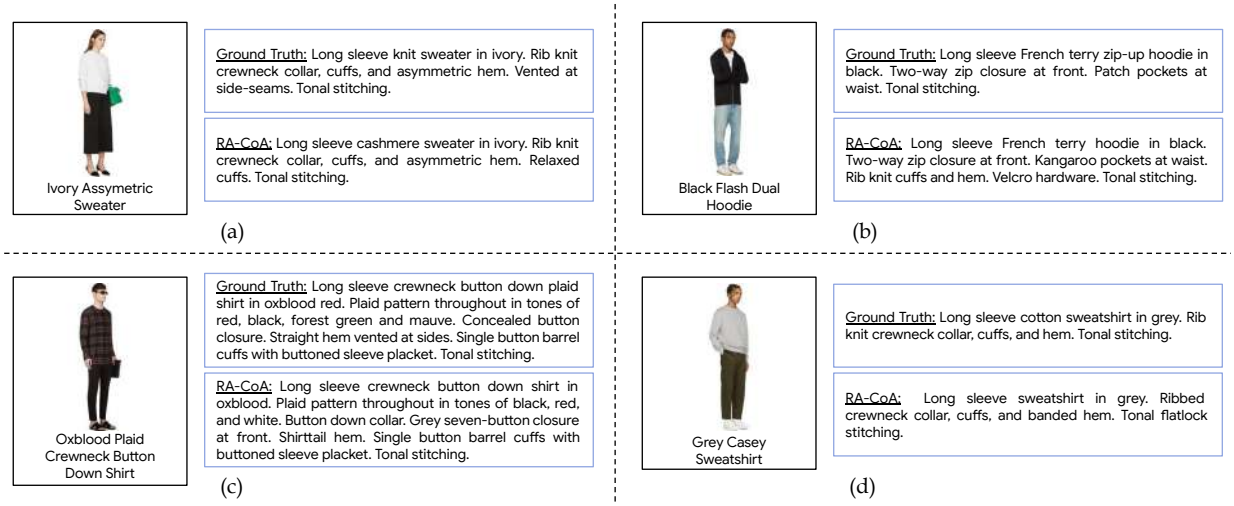


Figure 6.3: A selection of RA-CoA's results. Our proposed approach RA-CoA generates coherent captions with minimal hallucinations that align well with the human-annotated ground truth descriptions.

Model (# params)	Top-K	BLEU-1	Avg. BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR
TinyLLaVA (3B)	1	22.1	11.50	32.5	10.8	25.2	24.6
	2	20.0	9.8	30.5	9.4	23.2	22.1
	3	18.7	8.9	29.2	8.6	21.9	21.0
Qwen-2.5VL (3B)	1	26.8	14.5	35.3	11.7	28.3	28.6
	2	31.3	16.6	37.7	12.6	29.3	31.3
	3	25.6	13.1	33.1	9.8	25.1	28.3
Qwen-2.5VL (7B)	1	31.4	17.7	39.2	14.2	32.3	32.4
	2	31.3	16.6	37.7	12.6	29.4	31.3
	3	29.8	15.4	36.6	11.7	27.7	30.8
InternVL2 (8B)	1	38.6	23.7	41.0	17.4	36.0	38.1
	2	38.3	22.9	40.2	15.8	33.3	38.3
	3	36.5	21.6	38.8	14.9	31.6	37.5

Table 6.2: Ablation study to quantify the impact of top-K retrievals for CoA in our proposed RA-CoA method.

OPTIMAL TOP-K

In RA-CoA, attributes are inferred from retrieved products (Section 6.3.2) and subsequently used to guide caption generation (Section 6.3.2). A key design choice in this process is the number of retrieved samples (top-K) used for attribute aggregation (Eq. 6.1). While increasing K can potentially improve attribute coverage by incorporating more diverse evidence, it may also introduce weakly aligned or irrelevant attributes, particularly in fine-grained fashion domains.

To study this trade-off, we vary K in {1, 2, 3} and report the results in Table 6.2. We observe that captioning performance generally decreases as K increases. This trend is consistent with the increased presence of noise and redundancy at higher K values, which can dilute the relevance

of the aggregated attribute set. In addition, conditioning on larger attribute sets may increase the tendency of VLMs to include unsupported details, affecting caption precision. Overall, these results suggest that a compact set of high-confidence attributes obtained from the single closest retrieved sample (K=1) provides a more favorable balance between attribute relevance and captioning accuracy for most models.

CONTRIBUTION OF ICL EXEMPLARS IN RA-COA

As elaborated in Section 6.3.2, RA-CoA leverages retrieved products as in-context exemplars during caption generation, to guide the VLM for structure and style of the generated caption. In this ablation, to understand the impact of in-context exemplars in the quality of caption generation, we evaluate RA-CoA without exemplars during the caption generation. Table 6.3 presents the quantitative results. Even without exemplars, RA-CoA consistently outperforms zero-shot, CoT-i, and CoT-e approaches, indicating the effectiveness of RA-CoA's attribute-centric reasoning. At the same time, removing exemplars leads to a noticeable but expected performance drop compared to the full RA-CoA. For example, METEOR decreases from 26.8 to 17.7 for Qwen2VL-3B and from 38.1 to 19.8 for InternVL2-8B. This demonstrates the critical role of retrieved exemplars in generating coherent, stylistically appropriate captions that effectively incorporate the identified attributes.

COMPARISON WITH ORACLE VARIANTS

To further study the contribution of different components in RA-CoA, we implement two oracle variants. Since original ground-truth attribute-value pairs are not available in FashionGen dataset, we obtain these pairs from expert-written captions similar to our ProductKB construction (Section 6.3.1) and consider these extracted pairs as ground truth for evaluation purposes.

Caption Generation with Oracle Attributes. Our RA-CoA method relies on retrieval to identify relevant attributes for a fashion item. To understand how close our retrieval mechanism comes to optimal attribute identification, we implement an Attribute-Oracle variant that directly uses ground truth attribute keys. Table 6.3 reveals that while Attribute-Oracle consistently outperforms standard RA-CoA, the gains are relatively modest across most models. TinyLLaVA (3B) improves by just 0.1 METEOR points (24.7 vs 24.6). The gap widens somewhat for larger models, with InternVL2 (8B) showing a 4.4-point improvement (42.5 vs 38.1). These results are encouraging, suggesting our retrieval-based attribute identification approaches the theoretical ceiling, particularly for smaller models. The increased gap for larger models indicates that as VLM capability grows, the limiting factor shifts more toward attribute identification quality rather than the model's ability to leverage those attributes.

Caption Generation with Oracle Attribute-Values. To establish the theoretical upper limit of our approach, we implement an Attribute-value-pair-Oracle variant that directly uses gold-standard attribute-value pairs extracted from expert captions. Table 6.3 shows dramatic performance im-

Model (# params)	Method variant	BLEU-1	Avg.BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR
TinyLLaVA (3B)	RA-CoA	22.1	11.50	32.5	10.8	25.2	24.6
	RA-CoA w/o ICL exemplars	16.8	5.3	19.8	2.5	12.8	17.9
	RA-CoA with Oracle Attributes	22.9	11.2	34.1	10.5	26.4	24.7
	RA-CoA with Oracle Attribute-Values	53.9	44.9	74.0	60.4	71.3	71.6
Qwen-2.5VL (3B)	RA-CoA	26.8	14.5	35.3	11.7	28.3	28.6
	RA-CoA w/o ICL exemplars	17.7	5.7	22.6	2.8	14.2	19.8
	RA-CoA with Oracle Attributes	30.6	16.4	40.2	12.6	31.8	31.8
	RA-CoA with Oracle Attribute-Values	69.4	62.8	87.4	77.2	83.8	82.4
Qwen-2.5VL (7B)	RA-CoA	31.4	17.7	39.2	14.2	32.3	32.4
	RA-CoA w/o ICL exemplars	17.9	5.7	23.8	2.8	14.3	21.7
	RA-CoA with Oracle Attributes	33.4	18.1	44.0	14.6	35.6	34.1
	RA-CoA with Oracle Attribute-Values	86.5	83.3	94.6	91.0	93.4	91.7
InternVL2 (8B)	RA-CoA	38.6	23.7	41.0	17.4	36.0	38.1
	RA-CoA w/o ICL exemplars	12.8	4.1	22.7	2.8	14.6	19.8
	RA-CoA with Oracle Attributes	45.1	27.0	47.2	18.7	40.8	42.5
	RA-CoA with Oracle Attribute-Values	95.7	93.8	98.2	95.5	97.4	97.1

Table 6.3: Ablation study to (1) quantify the importance of retrieval-augmented ICL exemplars, and (2) quantify the gap of RA-CoA with respect to oracle variants.

provements across all models. TinyLLaVA (3B) reaches 71.6 METEOR (vs 24.6 for standard RA-CoA), while InternVL2 (8B) achieves 97.1 METEOR (vs 38.1). Unlike the modest gains from Attribute-Oracle, these substantial improvements reveal that while RA-CoA effectively identifies relevant attributes, the primary performance bottleneck lies in attribute value generation. VLMs struggle significantly with accurately determining attribute values from visual input, despite correctly identifying which attributes to consider. The widening gap with larger models suggests that as VLM capability increases, the limiting factor becomes increasingly centered on value prediction accuracy, pointing to a clear direction for future improvements in the CoA mechanism.

EFFECT OF PRODUCT-AWARE CROPPING ON RETRIEVAL AND CAPTIONING

To isolate the impact of our retrieval backbone and examine the role of product-aware cropping in both retrieval and captioning, we perform a two-part analysis.

Retrieval. We compare two setups: (1) *Global CLIP-only Retrieval* [169], which retrieves similar samples using embeddings of the full image instead of cropped ones, and (2) *Florence–CLIP Retrieval*, which first detects and crops the target product region using Florence-2 [238](as described in Section 6.3.1) before embedding it with CLIP. Figure 6.4 illustrates representative retrieval errors from the CLIP-only approach, where visually unrelated items (e.g., "*Green OG Authentic LX Sneakers*" retrieved for a "*Grey Pocket Logo T-Shirt*") are selected due to background similarity. In contrast, the Florence–CLIP variant consistently retrieves visually and semantically aligned products by focusing on the correct region of interest. This refinement substantially reduces noisy retrievals and establishes a cleaner foundation for downstream attribute aggregation and caption generation.

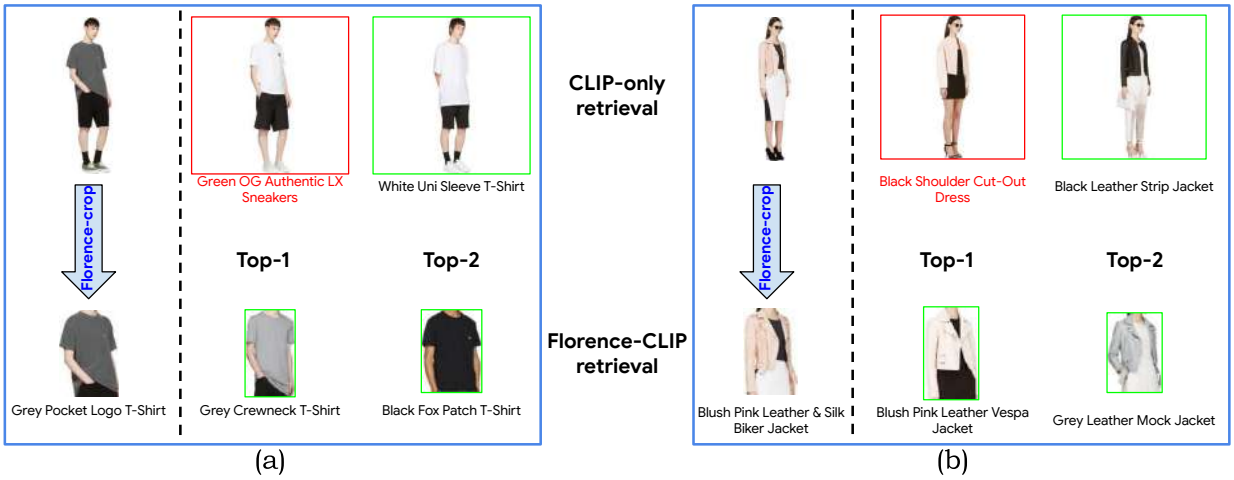


Figure 6.4: Comparison of CLIP-only and Florence-CLIP for retrieval in RA-CoA (Section 6.3.2).

Method	BLEU-1	Avg.BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR
Full Image	38.6	23.7	41.0	17.4	36.0	38.1
Cropped Product	39.4	24.3	41.8	18.0	36.7	38.8

Table 6.4: Comparison of using full image vs. Florence-2–cropped product region as VLM input within the RA-CoA framework. Cropping offers a modest gain while the full image already yields strong results.

Captioning. While product-aware cropping is essential for retrieval, during caption generation we input the full image to the Vision-Language Model (VLM). This choice is motivated by the fact that modern instruction-tuned VLMs exhibit strong coarse-level grounding abilities and can attend to the relevant product when guided by its name in the prompt. To further validate this design, we conducted an experiment where the VLM (InternVL2-8B) received the Florence-2–cropped product region instead of the full image. We observed a modest METEOR gain of +0.8 (38.9 vs. 38.1), suggesting that while explicit cropping can offer a small improvement, the model already performs effectively with full images. Hence, cropping is *necessary for retrieval* but only *optional for caption generation*, providing marginal benefit when used. Results are summarized in Table 6.4.

EFFECT OF PRODUCTKB QUALITY AND SCALE

Since RA-CoA relies on a structured Product Knowledge Base (ProductKB) for retrieval-augmented reasoning, its performance may depend on the quality, completeness, and scale of this knowledge source. To examine this, we conduct a series of controlled experiments analyzing three aspects: (i) varying KB size, (ii) introducing noise into KB entries (Noisy KB), and (iii) simulating sparsity in attribute annotations (Sparse KB).

Varying Knowledge Base Size. We progressively subsample the ProductKB to contain 10K, 20K, 40K, and 60K entries, and evaluate RA-CoA using InternVL2-8B as the VLM. Table 6.5 summarizes the results. The corresponding METEOR scores are 36.9, 35.5, 37.1, 37.9, and 38.1, respec-

ProductKB size	BLEU-1	Avg.BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR
10K	36.2	21.1	38.8	15.1	33.4	35.5
20K	37.6	22.7	40.3	16.6	35.2	37.1
40K	38.5	23.5	41.2	17.2	35.9	37.9
60K (Full)	***38.1***	***23.7***	***41.0***	***17.4***	***36.0***	***38.1***

Table 6.5: Effect of ProductKB size on captioning performance using InternVL2-8B within the RA-CoA framework. Performance improves with KB size, but gains saturate beyond 20K entries, indicating that RA-CoA remains data-efficient even with smaller knowledge bases.

% corrupted	BLEU-1	Avg.BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR
0%	***38.1***	***23.7***	***41.0***	***17.4***	***36.0***	***38.1***
25%	37.0	21.8	39.5	16.1	34.5	36.5
50%	36.9	21.3	39.2	15.3	33.9	35.9
75%	35.7	19.9	37.7	13.8	32.4	34.3

Table 6.6: Effect of noisy ProductKB entries on captioning performance using InternVL2-8B within the RA-CoA framework. Performance decreases with increasing noise in attribute values, suggesting that RA-CoA is relatively robust to corrupted ProductKB entries.

tively. While performance improves with KB size, gains beyond 20K are incremental. This suggests that even moderately sized KBs capture sufficient diversity for most fashion products, and larger KBs mainly help when new or niche SKUs are introduced.

Noisy Knowledge Base. To evaluate robustness to noisy entries in the ProductKB, we introduce noise by shuffling attribute values in 25%, 50%, and 75% of the KB entries (e.g., replacing half-sleeve with full-sleeve or velcro with zip). Table 6.6 summarizes the results. As the noise level increases, METEOR scores decrease slightly from 38.1 (original KB) to 36.5, 35.9, and 34.3, respectively. This limited degradation highlights the robustness of RA-CoA, which stems from a key design choice: post retrieval during *Chain-of-Attributes*, we use only the attribute *keys* from retrieved samples, not their values. This prevents noisy attribute values from being propagated while still guiding the model on which attributes to predict.

Sparse Knowledge Base. To analyze the effect of incomplete attribute annotations, we simulate a sparse ProductKB by randomly dropping 25%, 50%, and 75% of key–value pairs from each KB entry. Results are summarized in Table 6.7. As sparsity increases, METEOR scores decline from 38.1 to 33.9, 29.7, and 25.5, respectively. While moderate sparsity (25%) has limited impact, performance degrades substantially at higher sparsity levels (75%). This degradation can be attributed to two factors. First, fewer available attribute keys reduce explicit guidance on which visual properties the model should reason about. Second, and more critically, sparse KB entries weaken retrieval quality, as retrieved products become less semantically aligned with the query. Since these retrieved samples are also used as in-context exemplars during caption generation, reduced alignment corrupts exemplar quality and leads to weaker captions. Together, these effects highlight the importance of sufficiently informative KB entries for both attribute-centric reasoning and exemplar-guided caption synthesis in RA-CoA.

% key-value pairs dropped	BLEU-1	Avg.BLEU	Rouge-1	Rouge-2	Rouge-L	METEOR
0%	***38.1***	***23.7***	***41.0***	***17.4***	***36.0***	***38.1***
25%	34.4	20.5	39.2	16.4	34.3	33.9
50%	28.4	17.1	37.6	16.4	33.1	29.7
75%	21.1	13.1	35.9	16.7	31.6	25.5

Table 6.7: Effect of sparse ProductKB entries on captioning performance using InternVL2-8B within the RA-CoA framework. Performance degrades as sparsity in the ProductKB increases, as it affects retrieval quality and attribute guidance.

Method	Fine-tuning	BLEU-1	Avg.BLEU	Rouge-1	Rouge-2	ROUGE-L	METEOR
UniFashion	✓	9.6	3.0	16.4	1.5	12.6	11.9
RA-CoA (Ours)	✗	***17.2***	***6.1***	***22.9***	***35.6***	***16.4***	***19.2***

Table 6.8: Quantitative comparison of RA-CoA with UniFashion [266] (prior SOTA supervised method) showcasing real-world generalization of RA-CoA despite being completely training-free.

COMPARISON WITH PRIOR SOTA AND GENERALIZATION OF RA-COA BEYOND FASHIONGEN

Table 6.8 shows quantitative comparison of RA-CoA with prior state-of-the-art approach, UniFashion [266]. Notably, UniFashion is a fully supervised method, explicitly trained on large fashion datasets including FashionGen. Thus, direct evaluation of UniFashion on the FashionGen test set would be an unfair comparison with our training-free RA-CoA. To enable a fair and unbiased evaluation, we curated 50 fashion product images with their corresponding captions from the web[2] and evaluated both approaches on this held-out real-world set. In this setting, our training-free RA-CoA (with InternVL2-8B) consistently outperforms the supervised UniFashion (7B) baseline across all evaluation metrics, achieving gains of 7.6 points in BLEU-1, 3.1 points in average BLEU, 6.5 points in ROUGE-1, 34.1 points in ROUGE-2, 3.8 points in ROUGE-L, and 7.3 points in METEOR. The overall low scores can be attributed to the difference in style of ground truth captions and the captions in ProductKB (derived from FashionGen dataset). This comparison highlights the performance gap that arises between training-intensive and training-free paradigms under real-world distribution shift, demonstrating the robustness and scalability of RA-CoA in practical deployment scenarios.

USER PREFERENCE STUDY

To complement automatic evaluation, we conducted a user preference study to assess the qualitative usefulness of captions generated by different paradigms. We recruited six human evaluators, each of whom rated 50 randomly sampled, disjoint test images, resulting in 300 unique evaluation instances. For each image, users were shown anonymized captions generated by InternVL2-8B under all evaluated VLM paradigms. The order of captions was randomized to mitigate positional bias. User preferences are summarized in Table 6.9. Zero-shot captions are rarely

[2]www.amazon.com

Method	Preference (%)
Zero-shot	3.7
CoT-i (Implicit CoT)	6.7
CoT-e (Explicit CoT)	7.0
In-Context Learning (ICL)	14.0
RA-CoA (Ours)	***68.6***

Table 6.9: User preference study over 300 unique test images using InternVL2-8B as the backbone. RA-CoA is preferred in 68.6% of cases, reflecting higher perceived coherence, completeness, and usefulness.

preferred, highlighting the challenge of generating complete and user-aligned descriptions without additional guidance. Both implicit and explicit CoT variants provide only marginal gains over Zero-shot, suggesting that reasoning traces alone, without explicit grounding in retrieved attribute knowledge, do not substantially improve caption quality. Although ICL outperforms Zero-shot and CoT-based methods, it remains considerably less preferred than RA-CoA, indicating that exemplar-based prompting is insufficient for consistently high-quality caption generation. Overall, these results emphasize the importance of retrieval-augmented, attribute-aware reasoning for producing captions that better align with human expectations of coherence, completeness, and usefulness.

ATTRIBUTE-WISE SENSITIVITY ANALYSIS

To better understand attribute-level behavior beyond holistic caption quality, we conduct an explicit attribute-wise sensitivity analysis that evaluates how accurately different attributes are captured in the generated captions. Specifically, we extract the attribute values from the generated captions using LLaMA-3.2-8B-Instruct [63], and compare them against ground-truth attribute labels. Exact string matching is insufficient in this setting due to frequent lexical variability among semantically equivalent attributes (e.g., crew-neck vs. round-neck, zip vs. zipper, or slip-on vs. no-closure). To address this, we adopt a hybrid evaluation protocol: attributes that exactly match the labels are assigned a score of 5, while non-exact matches are evaluated using a LLaMA-based semantic similarity score on a 0–5 scale [143].

From this experiment, we observe the following: *(i) High-confidence or less sensitive attributes*: RA-CoA achieves higher accuracy (avg. score > 3.5) for the attributes that are visually prominent such as color, closure, sleeve type, front print, and hood. *(ii) Low-confidence or sensitive attributes*: RA-CoA achieves lower accuracy (average score < 2.5) for attributes that are often occluded, subtle, or viewpoint-sensitive, such as, hemline, fly type, and pockets.

ERROR ANALYSIS

We conducted an error analysis of the captions generated by RA-CoA to understand their shortcomings. While RA-CoA produces coherent captions with reduced hallucination, some limita-

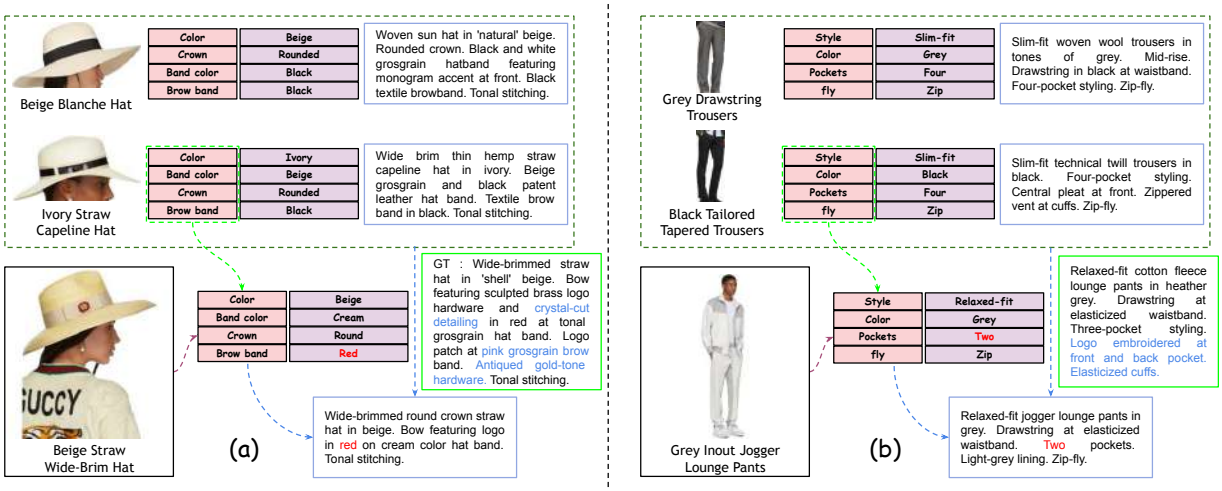


Figure 6.5: Error analysis into RA-CoA's generations. The text in red denotes incorrect/hallucinated features. The text in blue in the GT caption indicates features missed in the generated caption.

Step	Time (seconds)
ROI crop using Florence-2	0.30
CLIP embedding extraction for ROI	0.04
Retrieval from ProductKB (Section 6.3.2)	0.01
Chain-of-attributes (Section 6.3.2)	1.40
Attribute-value aware captioning (Section 6.3.2)	1.80
Average inference time per sample (Total)	***3.5***

Table 6.10: Computational latency analysis of RA-CoA's components.

tions remain and can be broadly categorized into two types: (i) missing attributes and (ii) hallucinated values. An illustration of these error types is shown in Figure 6.5. These issues arise from gaps in attribute coverage during retrieval and the VLM's limited grounding between visual cues and attribute semantics. These challenges could be mitigated by (i) expanding the ProductKB to include a more diverse and comprehensive set of product samples, and (ii) performing one-time pretraining of VLMs on structured attribute–value annotations to improve semantic grounding of product attributes, which we leave for future work.

LATENCY ANALYSIS

To assess the practical feasibility of RA-CoA for real-world e-commerce applications, we analyze the per-sample inference time of our framework. Table 6.10 reports the computational breakdown averaged across 1,000 samples on a single NVIDIA A6000 GPU. The total average inference time per sample is approximately 3.5 seconds, with the primary computational bottleneck being the two-stage VLM prompting: chain-of-attribute guided value generation (1.4s) and attribute-value aware caption generation (1.8s). In contrast, the retrieval is highly efficient. ROI extraction using Florence-2 takes 0.3s, CLIP embedding extraction requires only 0.04s, and retrieval from a Produc-

tKB of 60K samples via Faiss indexing [49] completes in 0.01s. This latency profile demonstrates that RA-CoA achieves practical inference speeds suitable for e-commerce applications.

6.5 CONCLUSION

In this chapter, we proposed RA-CoA, a training-free, model-agnostic framework for fashion image captioning that enhances attribute-level precision and interpretability by disentangling captioning into chain-of-attribute reasoning followed by generation. Unlike supervised methods that require frequent retraining to adapt to evolving fashion trends and vocabularies, RA-CoA leverages a product knowledge base and prompts frozen VLMs to infer relevant attributes and synthesize coherent captions. Extensive evaluations demonstrate that our approach reduces hallucination, improves caption faithfulness, and supports scalable, real-world deployment in dynamic fashion environments.

6.6 LIMITATIONS

Despite RA-CoA's improvements in fashion image captioning, there are a few limitations: (i) performance depends on the quality and diversity of the product knowledge base; biases or gaps in the retrieval database may propagate to generated captions. (ii) as revealed by our oracle experiments, there exists a substantial gap between attribute identification and accurate value prediction, VLMs still struggle with inferring fine-grained attribute values from visual input, particularly for subtle characteristics like fabric composition. (iii) The current implementation also focuses on Western fashion vocabularies and may inadequately capture cultural-specific fashion elements.

6.7 ETHICAL CONSIDERATIONS

While RA-CoA offers accessibility benefits in fashion e-commerce, we acknowledge several ethical considerations. FashionGen contains human models wearing fashion products, raising privacy concerns and potential demographic biases. We mitigate these by using the dataset under its original license, and avoiding identity analysis. Our retrieval mechanism may still propagate biases if the ProductKB lacks diversity in represented styles or cultural demographics. The approach could theoretically be misused for misleading product descriptions, though our attribute-grounded design provides inherent safeguards. Future work should explore dataset de-identification, demographic fairness measures, and further reduction of computational requirements while improving attribute inference capabilities.

7
Making VLMs Efficient

In this chapter, we address the second objective of this thesis: making vision-language models more efficient. While Large Vision-Language Models (L-VLMs) have demonstrated remarkable performance across various vision and language tasks, including visual question answering (VQA), their high computational cost renders them impractical for resource-constrained settings and inference-heavy applications. In contrast, Small Vision-Language Models (S-VLMs) offer computational efficiency but suffer from a significant performance gap compared to their larger counterparts. To bridge this gap, we introduce the Model Parity Aligner (MPA), a novel framework designed to systematically improve S-VLMs by leveraging unlabeled images and effective knowledge transfer from L-VLMs. Unlike traditional knowledge distillation methods that rely on labeled training data, MPA employs a strategic parity-based approach that precisely identifies knowledge disparities between S-VLMs and L-VLMs, then optimizes training by targeting only these specific disparities. We conduct extensive experiments on four diverse VQA benchmarks: TextVQA, ST-VQA, ChartQA, and OKVQA. Each benchmark requires specialized reasoning capabilities, including text recognition, chart interpretation, and commonsense and factual understanding. Our results demonstrate that MPA consistently enhances S-VLM performance across all benchmarks, significantly reducing the performance gap while maintaining computational efficiency.

7.1 INTRODUCTION

Large vision and language models (L-VLMs) have recently made remarkable progress on various vision and language tasks, including visual question answering (VQA) [32, 40, 120, 133, 221, 252, 271]. This makes them a de facto first choice for the VQA task on a new data set that does not have labeled training samples. However, L-VLMs may not be the most practical choice in resource-constrained settings and especially for inference-heavy tasks such as VQA, due to their high computational requirements and latency. In contrast, smaller vision and language models (S-VLMs) are more efficient but fall significantly short in performance, as shown in Figure 7.1. This raises a critical question: *Can we improve* S-VLM*s by a relevant and effective knowledge transfer*

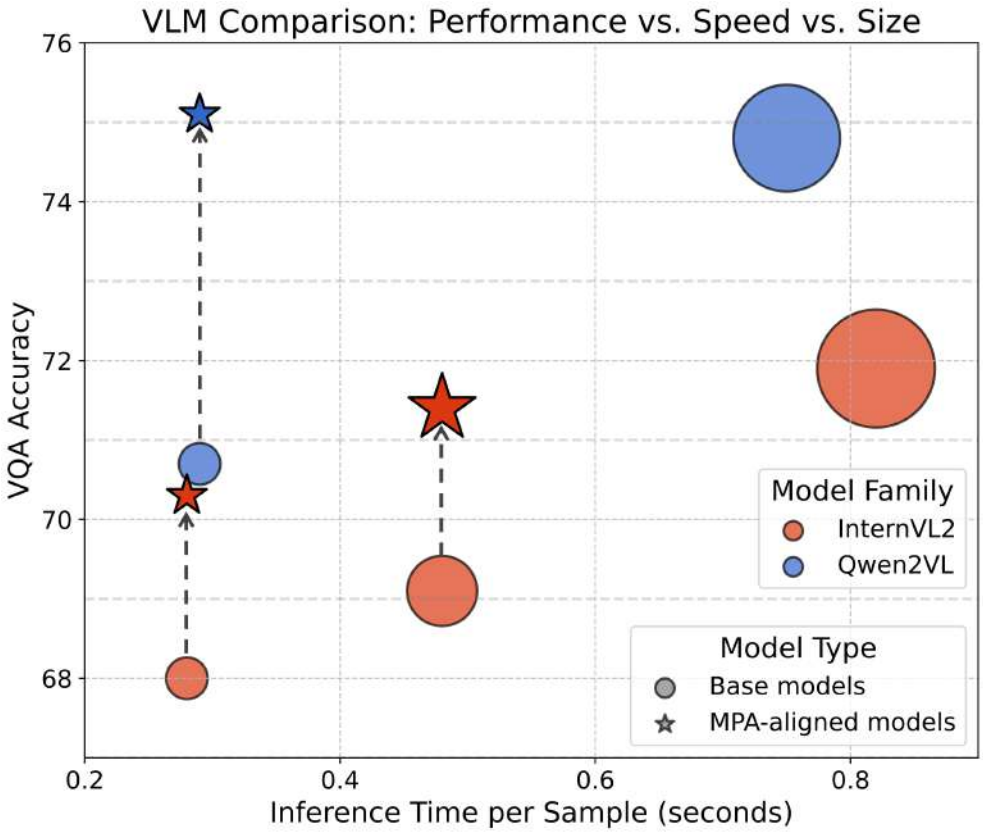


Figure 7.1: Small models often struggle to match the performance of their larger counterparts. We show model sizes using circle with radius proportional to the parameter count, and their respective inference time and VQA accuracies in X and Y-axis, respectively on one of the datasets used in this work [196]. Proposed MPA significantly enhances VQA accuracy for five S-VLMs across four datasets. ***(Best viewed in color)***.

from L-VLMs*?*

Several techniques have been explored to transfer knowledge from large neural models to smaller ones such as: (i) knowledge distillation (KD) [22, 64, 74, 108, 182, 192, 246] trains a small model (student) to mimic a large model (teacher) by learning from its soft labels or intermediate representations. However, KD typically relies on labeled training data, which may not always be available, and effectively distilling multimodal knowledge remains a challenge due to the complex interplay between vision and language features. (ii) Adapter-based methods [45, 77, 81, 131] introduce lightweight trainable layers into large models to enable efficient fine-tuning. Although these methods reduce training costs, they still require access to large models during inference, limiting their practical advantages in resource-constrained environments. (iii) Self-supervised learning and pseudo-labeling [29, 103, 172, 215, 241] provide an alternative by leveraging unlabeled data to generate training signals. However, naïve pseudo-labeling often propagates noisy predictions, reducing overall effectiveness. Moreover, the challenge of systematically transferring knowledge from large to small vision-language models using pseudo-labeling remains largely under explored. Addressing this gap is crucial for making smaller models more capable without the high computational cost of large models for inference.

To fill the aforementioned gaps, we introduce the Model Parity Aligner (MPA) – a framework that enables effective knowledge transfer from L-VLM to S-VLM using only unlabeled images. Instead of relying on traditional knowledge distillation or fine-tuning, MPA utilizes large model-guided pseudo-labeling with quality assessment. MPA accurately identifies and addresses the knowledge gaps between S-VLM and L-VLM, ensuring that small models learn from high-confidence predictions while minimizing error propagation. By leveraging the strong reasoning capabilities of large VLMs to create high-quality supervision signals through systematic parity assessment, MPA efficiently addresses performance gaps while maintaining computational efficiency.

We conducted extensive experiments and ablation studies to evaluate the effectiveness of the MPA. Specifically, we used four public datasets – TextVQA [196], ST-VQA [20], ChartQA [147], and OKVQA [146]. These datasets require additional capabilities such as visual text understanding, chart interpretation, and world knowledge integration, making them well-suited to test the robustness of MPA. We experimented with ten combinations of L-VLM and S-VLM pairs, demonstrating that MPA consistently improves S-VLM performance across all benchmarks, highlighting its effectiveness in knowledge transfer.

The first contribution of this work is the Model Parity Aligner (MPA), an effective approach that empowers small VLMs and improves their visual question answering performance using only unlabeled images, thereby eliminating the need for expensive labeled datasets. The second contribution is a novel parity-based training paradigm employed by MPA. In this paradigm, the L-VLM generates pseudo-labels for unlabeled images while also identifying and targeting specific knowledge gaps between the S-VLM and the L-VLM. This strategy ensures reliable supervision, minimizes noise, and maximizes relevant knowledge transfer. The third contribution is comprehensive experimental validation across four diverse VQA benchmarks. MPA achieves consistent improvement over strong baselines, and our findings further show that it not only boosts VQA performance but also enables the S-VLM to benefit from closed-source L-VLMs. In addition, MPA enhances the core capabilities of S-VLMs beyond VQA, including text recognition and text-aware captioning.

The remainder of this chapter is organized as follows. In Section 7.2, we review related literature on small and large VLMs, knowledge distillation, and data augmentation for VQA. Section 7.3 presents our proposed framework, the Model Parity Aligner (MPA), including its three modules: Pseudo Annotator (Section 7.3.1), Parity Identifier (Section 7.3.2), and Parity Leveler (Section 7.3.3). Experimental setup, datasets, evaluation protocols, and results are detailed in Section 7.4, along with ablations and qualitative analyses. Finally, Section 7.5 concludes the chapter with key findings and future work, followed by a discussion of limitations and ethical considerations and broader impact.

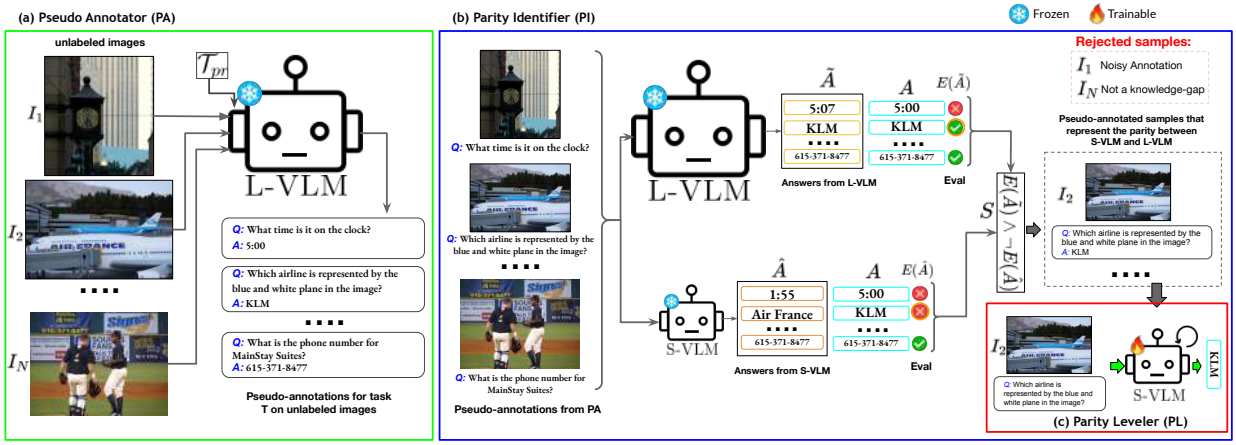


Figure 7.2: Overview of the proposed MPA framework. It consists of three modules, namely (a) Pseudo Annotator (Section 7.3.1), (b) Parity Identifier (Section 7.3.2), and (c) Parity Leveler (Section 7.3.3). Given a set of unlabeled images $\mathcal{I}$ and task $\mathcal{T}$, MPA begins with automatically annotating the unlabeled images, followed by strategic data selection that targets knowledge gaps of S-VLM with the L-VLM, while accounting for annotation quality. This selection process identifies parity, capturing instances where the L-VLM answers correctly while the S-VLM fails. Finally, PL updates the S-VLM's parameters on the obtained parity subset. ***(Best viewed in color)***.

7.2 RELATED WORK

Small and Large VLMs: Following the success of large language models (LLMs) [21, 46, 50, 166, 211, 247] across NLP tasks, vision and language models (VLMs) [32, 40, 133, 144, 221, 252, 268, 271] have been developed that process both visual and textual data. Although state-of-the-art VLMs achieve impressive zero-shot performance, their growing parameter count impose significant constraints on computational efficiency, accessibility, and deployment costs. This trade-off between efficiency and capacity requires the development of smaller VLMs [144, 187, 268] that maintain competitive performance with reduced computational demands [140]. The key approach to developing S-VLMs from L-VLMs involves substituting the internal LLM with lightweight alternatives [1, 88, 209, 260]. Inspired by the literature on LLM [140], we follow a parameter-based taxonomy where VLMs with $\leq$ 5B parameters are classified as S-VLMs, while those that exceed this threshold are L-VLMs. For context, a small 4B-parameter VLM constitutes just 0.2% of the estimated 1.8T parameters of GPT-4.

Knowledge Distillation: It transfers knowledge from large teacher models to smaller student models using KL-divergence over soft logits [74] or feature representations [182, 217, 245]. With LLMs adhering to scaling laws, their distillation has gained significant interest. Recent methods for LLMs [64, 108] and L-VLMs [22, 192, 246] explore KL-Divergence variants, while others [78, 177, 210] distill reasoning via LLM-generated Chain-of-Thought rationales. In contrast to standard KD, which distills over the labeled dataset, our method identifies and supervises only the samples that represent knowledge gaps between the student and teacher. This targeted strategy enables efficient, model-agnostic training using only input-output access to the teacher—including closed-source L-VLMs.

Algorithm 2 Model Parity Aligner (MPA)

Input: Large Vision-Language Model (L-VLM) parameterized by ϕ; Small Vision-Language Model (S-VLM) parameterized by θ; unlabeled images: $\mathcal{I} = \{I_1, I_2, \cdots, I_N\}$; ***task***: $\mathcal{T}$.
Output: Enhanced S-VLM with updated parameters ($\hat{\theta}$).

$\mathcal{D}_{PA}^{\mathcal{T}} \leftarrow$ ***PA***(L-VLM$_\phi$, $\mathcal{I}$, $\mathcal{T}$) ▷ PA - Pseudo Annotator, $\mathcal{D}_{PA}^{\mathcal{T}}$: pseudo-annotated data
$\mathcal{D}_{PI}^{\mathcal{T}} \leftarrow$ ***PI***(L-VLM$_\phi$, S-VLM$_\theta$,$\mathcal{D}_{PA}^{\mathcal{T}}$) ▷ PI - Parity Identifier, $\mathcal{D}_{PI}^{\mathcal{T}}$: parity dataset
S-VLM$_{\hat{\theta}} \leftarrow$ ***PL***(S-VLM$_\theta$, $\mathcal{D}_{PI}^{\mathcal{T}}$) ▷ PL: Parity Leveler
return S-VLM$_{\hat{\theta}}$

Data Augmentation for VQA: Vision and language tasks such as VQA have traditionally been benefited by data augmentation, and visual question generation becomes a natural choice to generate augmented data [53, 89, 90, 109, 151, 214, 218, 262]. Although few methods [28, 97, 103, 104] augment the data-scarce VQA datasets to improve performance, other methods [13, 25] leverage large-scale image-caption datasets to generate noisy VQA labels and use them as VQA foundational data. Distinctively different from these lines of work, we employ L-VLMs to pseudo-label unlabeled images with a quality check to discard noise, ensuring minimal yet effective annotations for targeted improvements of S-VLMs.

7.3 MODEL PARITY ALIGNER (MPA)

Given a task $\mathcal{T}$ and a set of unlabeled images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, our goal is to empower small vision-language models (S-VLMs) with task-specific capabilities and improve their performance on the task $\mathcal{T}$. In this work, we restrict ourselves to the VQA task and experiment with various variants of VQA that require interpretation of visual text, chart, and external knowledge. Inspired by the standard machine learning lifecycle [198], our proposed Model Parity Aligner (MPA) framework follows a systematic approach to achieve this goal. The process begins with automatically annotating unlabeled images $\mathcal{I}$ for task $\mathcal{T}$ using the Pseudo Annotator module discussed in Section 7.3.1, followed by strategic data selection with automatic quality assessment of the annotations using the Parity Identification module discussed in Section 7.3.2. This automatically curated and cleaned data is then utilized to fine-tune the S-VLM model using the Parity Leveler module discussed in Section 7.3.3. The workflow of our proposed MPA framework, which includes its three interconnected modules, is illustrated in Figure 7.2 and described in Algorithm 2.

The proposed Model Parity Aligner (MPA) consists of three main modules: (a) Pseudo Annotator (PA), (b) Parity Identifier (PI), (c) Parity Leveler (PL). These modules work together to systematically enrich S-VLMs. The MPA takes S-VLM$_\theta$, L-VLM$_\phi$, a set of unlabeled images $\mathcal{I}$ and the task $\mathcal{T}$ as inputs and returns an enhanced S-VLM$_{\hat{\theta}}$ where ϕ, θ, $\hat{\theta}$ are the parameters of L-VLM, S-VLM, updated S-VLM, respectively. It should be noted here that $|\hat{\theta}| = |\theta| \ll |\phi|$ where $|\cdot|$ denotes the size of the model. Next, we provide an in-depth overview of each module.

Algorithm 3 Pseudo Annotator (PA)

Input: Large Vision-Language Model (L-VLM) parameterized by ϕ; unlabeled images: $\mathcal{I} = \{I_i\}_{i=1}^{N}$; task prompt: $\mathcal{T}_{pr}$.
Output: pseudo-annotated images for the ***task*** $\mathcal{T}$: $\mathcal{D}_{PA}^{\mathcal{T}} = \{(I_i, Q_i, A_i)\}_{i=1}^{N}$.

$\mathcal{D}_{PA}^{\mathcal{T}} \leftarrow [\ ]$
for I_i in $\mathcal{I}$ ***do***
  $(Q, A)_i \leftarrow \text{L-VLM}_{\phi}(\mathcal{T}_{pr}, I_i)$
  $\mathcal{D}_{PA}^{\mathcal{T}}$.append$((I_i, Q_i, A_i))$ ▷ Triplet: (I_i, Q_i, A_i) is considered as one pseudo-annotated sample for task $\mathcal{T}$.
end for
return $\mathcal{D}_{PA}^{\mathcal{T}}$

7.3.1 PSEUDO ANNOTATOR (PA)

This module which is described in Algorithm 3 is responsible for obtaining pseudo-annotation for unlabeled images $\mathcal{I}$. We employ an L-VLM to generate annotations for the unlabeled images for the task $\mathcal{T}$. In this work, we experimented with two L-VLMs. Since we have only access to unlabeled images, we ask L-VLM to generate task-specific visual question and answer pairs. The generation of visual questions (VQG) has been shown to improve the visio-lingual abilities of a vision and language model [28, 96]. In this work, we additionally ask L-VLM to generate the corresponding answer. To be precise, L-VLM is prompted with a task-specific prompt $\mathcal{T}_{pr}$ to create task-specific question-answer pairs[1] $(Q, A)_i$ for each image I_i within $\mathcal{I}$, where $i \in \{1, \cdots, N\}$. The module produces the pseudo-annotated dataset $\mathcal{D}_{PA}^{\mathcal{T}}$ for task $\mathcal{T}$: $\{(I_i, Q_i, A_i)\}_{i=1}^{N}$, with each triplet (I_i, Q_i, A_i) representing an annotated sample for task $\mathcal{T}$. The L-VLM-driven automated annotation presents challenges, e.g., (i) noisy annotations and (ii) hallucinated content necessitating careful quality validations. Our proposed PI module, described next, inherently accounts for quality validations and minimizes such noisy annotations, while sampling for parity samples.

7.3.2 PARITY IDENTIFIER (PI)

This module capitalizes on the existing capabilities of S-VLM while isolating its knowledge gaps relative to L-VLM. Rather than following conventional approaches [25, 28, 103] of using all pseudo-annotated data for training, we implement a more targeted methodology to identify specific knowledge disparities between models. We evaluated both L-VLM and S-VLM in zero-shot settings by presenting each model with image-question pairs (I_i, Q_i) from the PA-annotated dataset $\mathcal{D}_{PA}^{\mathcal{T}}$. The respective answers - $\tilde{A}_i$ from L-VLM and $\hat{A}_i$ from S-VLM—are then compared against

[1]For example, in the case of ChartQA, $\mathcal{T}_{pr}$ instructs the model to focus on reasoning over charts, including trend analysis and numerical interpretation. Similarly, for TextVQA, the prompt emphasizes reading and comprehending scene text to formulate relevant questions and answers. This ensures that the generated QA pairs align with the specific reasoning challenges posed by each task.

Algorithm 4 Parity Identifier (PI)

Input: Large Vision-Language Model (L-VLM) parameterized by ϕ; Small Vision-Language Model (S-VLM) parameterized by θ; pseudo-annotated data: $\mathcal{D}_{PA}^{\mathcal{T}}$.
Output: Parity (Knowledge gap) between L-VLM and S-VLM: $\mathcal{D}_{PI}^{\mathcal{T}} = \{(I_i, Q_i, A_i)\}_{i=1}^{K}, K \ll N$.

$\mathcal{D}_{PI}^{\mathcal{T}} = [\ ]$
for (I_i, Q_i, A_i) in $\mathcal{D}_{PA}^{\mathcal{T}}$ ***do***
  $\tilde{A}_i \leftarrow \text{L-VLM}_{\phi}(I_i, Q_i)$
  $\hat{A}_i \leftarrow \text{S-VLM}_{\theta}(I_i, Q_i)$
  if $\tilde{A}_i == A_i$ ***and*** $\hat{A}_i \neq A_i$ ***then*** ▷ Eq. 7.1 & Eq. 7.2
    $\mathcal{D}_{PI}^{\mathcal{T}}$.append$((I_i, Q_i, A_i))$ ▷ Satisfies Eq. 7.2 criteria
  else
    continue
  end if
end for
return $\mathcal{D}_{PI}^{\mathcal{T}}$

the pseudo annotation A_i using the following expression.

$$E(X) = \begin{cases} 1, & \text{if } X = A, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } X \in \{\tilde{A}, \hat{A}\}. \tag{7.1}$$

Further, we select samples that satisfy the following Boolean condition S.

$$S\big((I, Q, A)\big) = \begin{cases} 1, & \text{if } E(\tilde{A}) \wedge \neg E(\hat{A}), \\ 0, & \text{otherwise.} \end{cases} \tag{7.2}$$

Here, Boolean condition S selects an annotated triplet (I_i, Q_i, A_i) if $\tilde{A}_i$ correctly matches A_i while $\hat{A}_i$ does not, thereby precisely identifying the knowledge gap between the models where S-VLM requires improvement. In other words, S selects those samples where L-VLM answers correctly, while S-VLM answer is incorrect, assuming the pseudo-annotated answer as ground truth. This methodology inherently performs quality verification by leveraging L-VLM's superior answering capabilities, as these models are primarily instruction-tuned for answering rather than annotating. By selecting only instances where L-VLM demonstrates consistency between its annotation and answering phases, PI module effectively filters out noisy or hallucinated annotations. The resulting parity subset $\mathcal{D}_{PI}^{\mathcal{T}} : \{(I_i, Q_i, A_i)\}_{i=1}^{K}$ with $K \ll N$, constitutes highly efficient samples focused exclusively on the specific knowledge deficiencies of S-VLM. This targeted approach eliminates the need to train on potentially problematic samples or the entire annotation set, optimizing both training efficiency and model performance. This module is detailed in Algorithm 4.

Algorithm 5 Parity Leveler (PL)

Input: Small Vision-Language Model (S-VLM) parameterized by θ; parity set: $\mathcal{D}_{PI}^{\mathcal{T}} = \{(I_i, Q_i, A_i)\}_{i=1}^{K}$

Output: Enhanced S-VLM with updated parameters $\hat{\theta}$

for iter $= 1$ to L ***do*** ▷ L: total no. of iterations
  for $\{(I_i, Q_i, A_i)\}_{i=1}^{b}$ in $\mathcal{D}_{PI}^{\mathcal{T}}$ ***do*** ▷ b: batch size
    $\{\hat{A}_i\}_{i=1}^{b} \leftarrow$ S-VLM$_\theta(\{(I_i, Q_i)\}_{i=1}^{b})$
    Compute $\mathcal{L}_{gen}(\{\hat{A}_i, A_i\}_{i=1}^{b})$ ▷ Answer generation loss
    Update θ using $\mathcal{L}_{gen}$ ▷ Gradient descent
  end for
end for
return S-VLM$_{\hat{\theta}}$

7.3.3 PARITY LEVELER (PL)

This module fine-tunes S-VLM on the parity (knowledge gap) samples identified by the PI module for L number of iterations. We feed each sample $\{I, Q\}_i$ from $\mathcal{D}_{PI}^{\mathcal{T}}$, within an instruction prompt template to S-VLM to generate the accurate answer A_i to the visual question Q_i on the image I_i. S-VLM learns $P(A_i|Q_i, I_i)$ by modeling the task as a text generation problem, auto-regressively generating the tokens in the answer.

$$\mathcal{L}_{gen}(\theta) = -\frac{1}{b}\sum_{i=1}^{b}\left[\sum_{t=1}^{m}\log P_\theta(A_{i_t}|A_{i<t}, \{I_i, Q_i\})\right] \tag{7.3}$$

Once all answer tokens $A_{i_{1:m}}$ are obtained, we optimize the model using the generation loss $\mathcal{L}_{gen}$, defined over the minibatches of size b samples (Eq.7.3) which is minimized via stochastic gradient descent. Note that L-VLM parameters ϕ remain frozen throughout MPA. For an algorithmic description of this module, refer to Algorithm 5.

7.4 EXPERIMENTS AND RESULTS

Datasets. We evaluate our approach on four widely-used public VQA benchmarks, namely, TextVQA [196], ST-VQA [20], ChartQA [147], and OKVQA [146]. These datasets are relevant to MPA because they introduce diverse reasoning challenges, such as text, chart, external world understanding beyond traditional VQA [7], making them strong benchmarks for evaluating gains in S-VLM. TextVQA consists of 28K images with 45K manually annotated question-answer pairs. It is split into 21K images with 35K questions for training, 3K images with 3.7K questions for validation, and a private test set. Since the testset is private, for this dataset, we report all the result on validation set. ST-VQA contains 23K images and 31K questions, with 16K images and 22K questions for training, and 2.8K images with 4K questions for testing. ChartQA includes 21.6K charts with 32.3K question-answer pairs, split into 19K charts with 28K questions for training, 1K charts with 1.8K questions for validation, and 1.6K charts with 2.5K questions for testing. OKVQA

S-VLM	Method	L-VLM: Qwen2VL-7B [221] TextVQA	ST-VQA	ChartQA	OKVQA	L-VLM: InternVL2-8B [32] TextVQA	ST-VQA	ChartQA	OKVQA	Gains: Max	Average
SmolVLM-500M	ZS	55.3	78.5	56.5	38.2	55.3	78.5	56.5	38.2	3.4	2.4
	MPA	***57.6***$_{(+2.3)}$	***80.3***$_{(+1.8)}$	***59.9***$_{(+3.4)}$	***40.7***$_{(+2.5)}$	***57.7***$_{(+2.4)}$	***80.7***$_{(+2.2)}$	***59.3***$_{(+2.8)}$	***39.9***$_{(+1.7)}$		
TinyLLaVA-2B	ZS	47.1	44.7	12.0	43.6	47.1	44.7	12.0	43.6	15.2	6.8
	MPA	***53.5***$_{(+6.4)}$	***48.7***$_{(+4.0)}$	***24.0***$_{(+12.0)}$	***46.6***$_{(+3.0)}$	***51.9***$_{(+4.8)}$	***49.8***$_{(+5.1)}$	***27.2***$_{(+15.2)}$	***47.2***$_{(+3.6)}$		
InternVL2-2B	ZS	68.0	63.0	63.2	42.7	68.0	63.0	63.2	42.7	5.1	3.0
	MPA	***70.3***$_{(+2.3)}$	***65.5***$_{(+2.5)}$	***68.3***$_{(+5.1)}$	***45.6***$_{(+2.9)}$	***69.5***$_{(+1.5)}$	***65.7***$_{(+2.7)}$	***68.2***$_{(+5.0)}$	***44.6***$_{(+1.9)}$		
InternVL2-4B	ZS	69.1	63.2	73.1	50.5	69.1	63.2	73.1	50.5	4.7	2.1
	MPA	***71.4***$_{(+2.3)}$	***66.6***$_{(+3.4)}$	***73.8***$_{(+0.7)}$	***52.3***$_{(+1.8)}$	***70.3***$_{(+1.2)}$	***67.9***$_{(+4.7)}$	***74.0***$_{(+0.9)}$	***52.0***$_{(+1.5)}$		
Qwen2VL-2B	ZS	70.6	62.5	65.9	47.1	70.7	62.5	65.9	47.1	4.7	2.6
	MPA	***75.1***$_{(+4.5)}$	***67.2***$_{(+4.7)}$	***67.6***$_{(+1.7)}$	***48.9***$_{(+1.8)}$	***72.3***$_{(+1.6)}$	***66.6***$_{(+4.1)}$	***66.9***$_{(+1.0)}$	***48.9***$_{(+1.8)}$		

Table 7.1: Comparison of our proposed MPA framework performance with the baselines on TextVQA, ST-VQA, ChartQA and OKVQA. The parenthesis (+x) denotes the improvement of +x% over the zero-shot S-VLM by our proposed MPA. The max and average columns show the overall performance gains across all tests for each S-VLM.

consists of 14K images with 14K questions, divided into 9K questions for training, 5K for testing. Further, as MPA is primarily designed for label-free training, we exclude all question-answer annotations from the training splits of each dataset during evaluation.

S-VLMs and L-VLMs used. Following the parameter-based taxonomy defined for Vision-Language Models (VLMs) in Section 7.2, where models with $\leq$ 5B parameters are classified as small VLMs (S-VLMs), while those exceeding 5B parameters are large VLMs (L-VLMs) [140], we chose five models that range from 500M to 4B parameters as S-VLM, namely SmolVLM-500M [144], TinyLLaVA-2B [268], InternVL2-2B [32], Qwen2VL-2B [221], and InternVL2-4B [32]; and two open-source models viz. Qwen2VL-7B [221] and InternVL2-8B [32] and one closed-source model, i.e., GPT-4o [156] as L-VLM.

7.4.1 RESULTS AND DISCUSSION

We present the quantitative results of our MPA framework across four datasets evaluated in ten combinations of two L-VLMs and five S-VLMs in Table 7.1. The results show that MPA consistently improves the performance of all S-VLMs in all datasets with 15.2% maximum and 3.4% average gain in an absolute scale. Here, we analyze the results from the following three key perspectives.

(i) S-VLM family-specific analysis The most noticeable gains are as follows (refer Table 7.1): TinyLLaVA-2B achieves 27.2% accuracy on ChartQA with our MPA framework, guided by InternVL2-8B, marking an absolute improvement of +15.2% over its original zero-shot performance. Similarly, Qwen2VL-2B, guided by Qwen2VL-7B and InternVL2-4B, guided by InternVL2-8B in our MPA framework achieve +4.7% and +4.7% improvements, respectively, on ST-VQA. On ChartVQA, SmolVLM-500M, guided by Qwen2VL-7B in our MPA framework, improves by +3.4%, while InternVL2-2B, guided by Qwen2VL-7B, gains +5.1%. These results highlight effectiveness of MPA in enhancing the performance of S-VLMs across diverse VQA tasks.

S-VLM	GPT-4o as L-VLM
TinyLLaVA-2B	47.1
TinyLLaVA-2B + MPA	**55.4**$_{(+8.3)}$
Qwen2VL-2B	70.6
Qwen2VL-2B + MPA	**75.4**$_{(+4.8)}$

Table 7.2: Comparison of MPA-aligned S-VLMs against baseline S-VLMs on TextVQA, with GPT-4o as L-VLM.

(ii) VQA Task-specific analysis We observe that TinyLLaVA-2B+MPA aligned with InternVL2-8B achieves a notable +15.2% gain on ChartQA, highlighting our MPA's strength as a knowledge alignment module. In this scenario, it effectively identifies and bridges the knowledge gap between L-VLM and S-VLM for '*complex visual reasoning that involves interpreting charts and graphs*. Improvements on TextVQA (+6.4%) and ST-VQA (+5.1%) further demonstrate MPA's ability to transfer '*visual text understanding*' from larger to smaller models. The modest gain on OKVQA reflects its reliance on external knowledge, which S-VLM inherently lack. While MPA enhances internal knowledge utilization, it cannot fully address such gaps without RAG or fine-tuning on knowledge-rich data. The results validate the effectiveness of MPA within its scope, while highlighting the challenges of knowledge-intensive visual question answering.

(iii) Model size-specific Analysis: MPA improves performance on all model scales, from SmolVLM-500M to InternVL2-4B, demonstrating its versatility. In particular, TinyLLaVA-2B achieves the highest average gain of +6.8 across all tasks, whereas InternVL2-4B shows a comparatively modest improvement of +2.1. We attribute this contrast to two factors: (i) Pretraining data gaps: smaller models like TinyLLaVA-2B benefit more from MPA as it effectively fills missing capabilities through targeted alignment; (ii) Diminishing returns with scale: it is inherently harder to align larger models (4B in this case) that already possess stronger capabilities, in line with scaling laws.

(iv) L-VLM-Specific Analysis: We analyze the effectiveness of different guiding L-VLMs within MPA by computing average gains across five S-VLMs and four VQA datasets. Qwen2VL-7B achieves the highest average improvement of +3.5 points, followed closely by InternVL2-8B with +3.2 points. This suggests that while both models are effective guides, Qwen2VL-7B offers a slightly stronger alignment signal, potentially due to differences in their pretraining objectives or representations. These results highlight that MPA is robust to the choice of L-VLM, yet benefits from stronger or more task-aligned guides.

ABLATIONS AND ANALYSIS

We conduct the following ablations and analysis:

(i) How effective is MPA in aligning S-VLMs with closed-source models?: MPA can also leverage powerful closed-source L-VLMs to improve S-VLMs. To assess this, we performed experiments using GPT-4o [156] as the guiding L-VLM. As shown in Table 7.2, MPA consistently improves

Task	Dataset	Metric	S-VLM	S-VLM+MPA
OCR	ICDAR2015 [98]	WRR	31.9	***36.4***(↑ 4.5)
TC	TextCaps [193]	BLEU-1	7.9	***15.3***(↑ 7.4)
		ROUGE-L	17.4	***20.6***(↑ 3.2)
		CIDEr	8.7	***38.1***(↑ 29.4)

Table 7.3: MPA transfers the fundamental capabilities beyond VQA. In our MPA framework, we use S-VLM: TinyLLaVA-2B, L-VLM: Qwen2VL-7B. Here, ***OCR:*** visual-text recognition, ***TC:*** text-aware image captioning. ***WRR:*** word recognition rate.

L-VLM	Status	A ↑	AC ↑	TR ↑	HLS ↑
Qwen2VL-7B	Pre-PI	0.76	0.68	0.8	58
	Post-PI	***0.92***	***0.84***	***0.92***	***74***
InternVL2-8B	Pre-PI	0.74	0.65	0.78	56
	Post-PI	***0.87***	***0.78***	***0.88***	***73***

Table 7.4: User study on the pseudo-annotations quality: Pre-PI and Post-PI in MPA. ***A***: answerability, ***AC***: answer correctness, ***TR***: task relevancy, ***HLS***: Human Likeness Score. Refer Section 7.4.1 for more details.

performance across all aligned S-VLMs, despite having no access to the guiding model's logits or weights. This demonstrates MPA's unique advantage over standard distillation methods, which require full model access. With the expected rise in powerful closed-source models [156, 208], such alignment strategies become increasingly valuable. In fact, our results show that integrating powerful L-VLM, e.g. GPT-4o through MPA brings S-VLMs closer or even better in performance to significantly larger models, e.g., MPA-aligned Qwen2VL-2B (75.4%) outperforms Qwen2VL-7B (74.7%).

(ii) Does MPA transfers the fundamental capabilities beyond VQA?: MPA is designed to enhance the VQA performance of S-VLMs by aligning them with L-VLMs, and our results confirm its effectiveness. To examine whether MPA also transfers broader fundamental capabilities such as visual text understanding, we evaluate zero-shot TinyLLaVA-2B and its MPA-aligned counterpart on two different tasks: visual text recognition on ICDAR 2015 [98] and text-aware image captioning on TextCaps [193], using Qwen2VL-7B as the guiding L-VLM in MPA. As shown in Table 7.3, the MPA-aligned model improved text recognition accuracy by 4.5% on an absolute scale and yields notable improvements in captioning metrics such as ROUGE-L and CIDEr. These results suggest that MPA transfers fundamental text understanding capabilities from L-VLMs to S-VLMs beyond the VQA.

(iii) How effective is the role of PI in pseudo-annotation quality correction?: Incorrect annotations may cause models to learn spurious patterns, exhibit biased behavior, and suffer from degraded performance and reliability in downstream tasks. To assess the impact of the PI module on generated annotation quality, we conducted a user study in which three annotators evaluated 500 randomly sampled pseudo-annotations prior and post PI processing. The evaluation used the following metrics: (a) *Answerability (A)*: 1 if the question is answerable from the image, 0 otherwise; (b) *Answer Correctness (AC)*: 1 if the answer is correct, assuming the question is valid; (c) *Task Relevance (TR)*: 1 if the question aligns with the task, 0 otherwise; and (d) *Human-Likeness Score*

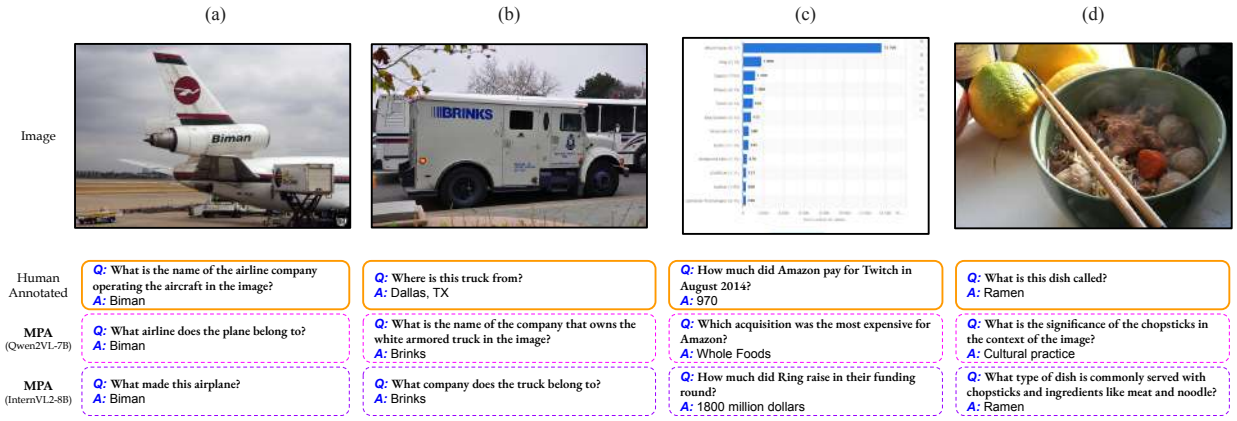


Figure 7.3: A selection of few pseudo annotations generated by our framework. We further show human annotations from their respective original dataset train splits. ***(Best viewed in color)***.

Method	TextVQA	ST-VQA	ChartQA	OKVQA
LoRA SFT	71.9	63.4	66.1	47.9
Full SFT	71.8	61.7	65.7	47.7
MPA	***75.1***	***67.2***	***67.6***	***48.9***

Table 7.5: Comparison of few-shot methods vs MPA-aligned Qwen2VL-2B with Qwen2VL-7B as L-VLM. Please note that MPA operates without any human-labeled samples, whereas the other two baselines each use 100 human-labeled samples.

(HLS): percentage of PI-sampled annotations mistaken for human-annotated ones in a mixed set. As shown in Table 7.4, post-PI annotations exhibited higher quality across all metrics, with more being identified as human-annotated. Figure 7.3 provides visual evidence by illustrating the high correlation between MPA-generated annotations and human annotated samples. These results validate that PI effectively filters noise and corrects errors, enhancing the overall reliability of MPA-generated annotations.

(iv) How does MPA compare to few-shot supervised baselines? While MPA is designed for a setting where human-labeled training data is unavailable, obtaining a small labeled set (e.g., 100 samples) is often feasible. In such scenarios, commonly adopted few-shot supervised methods like LoRA-based SFT and full SFT can be applied directly to the S-VLM. To benchmark MPA against these methods, we fine-tune Qwen2VL-2B using both approaches and compare them with MPA-aligned Qwen2VL-2B (using Qwen2VL-7B as L-VLM). As shown in Table 7.5, MPA consistently outperforms both baselines without labeled supervision, demonstrating high-quality label generation and effective knowledge transfer.

(v) Does PI filtering improve over raw pseudo-labels or full human-labeled data? While our primary focus is on label-free training using MPA, we further investigate the quality of supervision introduced by PI filtering. Specifically, we compare three settings for training Qwen2VL-2B: (i) full human-labeled data (HL), (ii) pseudo-labeled data (PL) from MPA without PI filtering, and (iii) high-quality subset selected by PI that targets the knowledge gap. As shown in Table 7.6, the PI-selected subset achieves the highest accuracy across all tasks–TextVQA (75.1%), ST-VQA

Data	Labels	TextVQA		ST-VQA		ChartQA	
		#Samples ↓	Acc. ↑	#Samples ↓	Acc. ↑	#Samples ↓	Acc. ↑
Original	HL	35K	72.7	22K	65.5	28K	66.9
MPA (w/o PI)*	PL	21K	73.3	15K	65.5	19K	67.4
MPA (w/o PI)	PL	21K	73.6	15K	65.8	19K	67.4
MPA	PL	***2K***	***75.1***	1.5K	***67.2***	1.6K	***67.6***

Table 7.6: Ablation result of using samples from MPA v/s MPA without PI filtering, with Qwen2VL-7B as L-VLM and Qwen2VL-2B as S-VLM inside MPA. HL: Human Labeled. PL: Pseudo Labeled.

S-VLM	Samples	TextVQA	ST-VQA	ChartQA	OKVQA
TinyLLaVA-2B	MPA (w/o PI)	56.4	79.1	57.5	39.2
	MPA	***57.6***	***80.3***	***59.9***	***40.7***
TinyLLaVA-2B	MPA (w/o PI)	52.1	46.3	23.3	44.6
	MPA	***53.5***	***48.7***	***24.0***	***46.6***
InternVL2-2B	MPA (w/o PI)	69.0	64.5	66.7	44.0
	MPA	***70.3***	***65.5***	***68.3***	***45.6***
InternVL2-4B	MPA (w/o PI)	69.8	65.1	72.9	51.2
	MPA	***71.4***	***66.1***	***73.8***	***52.3***
Qwen2VL-2B	MPA (w/o PI)	73.6	65.8	67.4	47.2
	MPA	***75.1***	***67.2***	***67.6***	***48.9***

Table 7.7: Additional results for MPA vs. MPA (w/o PI) across all S-VLMs, using Qwen2VL-7B as the L-VLM inside MPA.

(67.2%), and ChartQA (67.6%), despite using far fewer samples. Interestingly, the performance gain from full human-labeled data over zero-shot baselines is relatively limited. Prior work [103] suggests that excessive labeled data can introduce redundancy or noise, reducing the marginal benefit of supervision. This highlights the value of PI filtering in identifying high-utility samples that yield more efficient and effective learning.

(vi) Cross-Domain applicability (Medical VQA): To evaluate MPA's utility beyond standard VQA tasks, we assess its performance in the medical domain using the PathVQA dataset [73]. We compare zero-shot TinyLLaVA-2B with its MPA-aligned counterpart, guided by Qwen2VL-7B. We focus on the binary (yes/no) subset of PathVQA, as the open-ended questions often contain highly specialized medical terminology that poses challenges even for large models and may not reflect generalizable reasoning capabilities. As shown in Table 7.8, MPA yields a gain of +2.4%, demonstrating effective knowledge transfer even in diverse domain-specific settings. These results highlight MPA's ability to generalize across domains without requiring task-specific data or fine-tuning.

Model	Method	Acc. (%)
TinyLLaVA-2B	ZS	51.2
TinyLLaVA-2B	MPA	**53.6**$_{(+2.4)}$

Table 7.8: Performance on Medical VQA (PathVQA). MPA-aligned TinyLLaVA-2B (with Qwen2VL-7B as L-VLM) shows improved cross-domain generalization.

Task	Dataset	Metric	S-VLM (Zero-shot)	S-VLM (HL)	S-VLM (MPA)
OCR	ICDAR2015	WRR	31.9	33.2	**36.4** (↑ 4.5)
TC	TextCaps	BLEU-1	7.9	13.4	**15.3** (↑ 7.4)
		ROUGE-L	17.4	18.3	**20.6** (↑ 3.2)
		CIDEr	8.7	34.6	**38.1** (↑ 29.4)

Table 7.9: Comparison of OCR and text-aware captioning performance. Despite using no ground-truth labels, MPA outperforms both the zero-shot baseline and models trained on human-labeled data (HL).

7.4.2 ADDITIONAL ANALYSIS

(i) Additional comparisons: utility of PI filtering over raw pseudo-labels. We extend the analysis of PI filtering by reporting the results of MPA (w/o PI) across all S-VLMs with Qwen2VL-7B as the guiding L-VLM within MPA. As shown in Table 7.7, MPA consistently outperforms MPA (w/o PI) across all tasks and S-VLMs, despite using far fewer training samples (for instance, ∼2K vs. ∼21K for TextVQA). These results reinforce the utility of PI filtering in isolating knowledge-gap samples that provide more efficient and targeted supervision.

(ii) Expanded comparison on OCR and text-aware captioning tasks. In Table 7.3, we examined whether MPA-trained models can improve fundamental capabilities such as OCR and text-aware image captioning, even without direct supervision. We further evaluate this setting by comparing against models fine-tuned on the original human-labeled training splits of TextVQA; the results are presented in Table 7.9. As shown, MPA not only improves over the zero-shot baseline but also surpasses models trained with human-labeled annotations. This highlights that the gains stem from the effectiveness of the MPA pipeline, rather than from overlap between benchmarks, and demonstrates that MPA successfully transfers core visual-linguistic capabilities in a label-free manner.

(iii) Computational and API cost of PA and PI: MPA is a one-time pipeline where each image is processed by the L-VLM during the PA phase, and each generated (image, question) pair is passed once through the L-VLM and S-VLM during the PI phase. For open-source L-VLMs like Qwen2VL-7B deployed locally, this is computationally lightweight: on a machine with 3 A6000 (48GB) GPUs, generating approximately 21K pseudo-annotations (e.g., for TextVQA) takes around 4-6 hours end-to-end. Further, the PI step takes another 2-3 hours to identify the samples that represent the knowledge gaps. Alternatively, while using GPT-4o via API, we estimate the total cost of PA + PI for a single S-VLM–task pair to be around $11, making MPA a highly cost-effective label-free alternative to supervised training.

S-VLM	Batch Size	LR
Qwen2VL-2B	16	1e-5
InternVL2-2B	16	4e-5
InternVL2-4B	6	4e-5
SmolVLM-500M	16	1e-4
TinyLLaVA-2B	16	1e-4

Table 7.10: Hyperparameters used in the parity leveler module (Section 7.3.3) for each S-VLM.

7.4.3 IMPLEMENTATION DETAILS

We implement our method using PyTorch. Majority of the chosen S-VLMs and L-VLMs employed in our proposed method MPA, we use their original code-base repositories and/or their Huggingface implementations depending on the ease of reproducibility. Parity leveler (Section 7.3.3) module trains the entire S-VLM on the samples obtained from the PI module (Section 7.3.2) for one epoch, for all the benchmark datasets. Hyperparameters used by the PL module for different S-VLMs are summarized in Table 7.10. All our experiments are conducted on a machine with three Nvidia A6000 GPUs (48 GB each). For every L-VLM and S-VLM combination, it took approximately, 5-12 GPU hours for entire MPA, for one dataset. We use `gpt-4o-2024-11-20` [156] for our closed-source L-VLM ablation.

7.4.4 PROMPTS USED

In this section, we provide the VLM prompts used in the PA module (Section 7.3.1) to generate pseudo-annotations for all four datasets:

(i) Prompt for TextVQA [196] and STVQA [20]:

L-VLM prompt used in the PA module of MPA: TextVQA and ST-VQA

<image(**I**)>
The objective is to generate a question-answer pair for a Textual Visual Question Answering (Text-VQA) task. Your task is to create a contextually relevant question that directly relates to the image's content, incorporating reasoning or direct references to the text, and its correct answer.
Output:
- Question: A natural language question grounded in the image's content and text.
- Answer: A concise response (single word, phrase, or Yes/No) derived from the text or reasoning based on it.
Assistant: Question: $\tilde{\mathbf{Q}}$, Answer: $\tilde{\mathbf{A}}$

(ii) ChartQA [147]:

L-VLM prompt used in the PA module of MPA: ChartQA [147]

<chart image (**I**)>
The objective is to generate a question-answer pair for a Chart Visual Question Answering (ChartVQA) task. Your task is to create a contextually relevant question that directly relates to the content of a given chart, incorporating reasoning based on the visualized data.
Output Requirements:
- Question: A natural language question grounded in the chart's content, requiring numerical reasoning, trend analysis, or data lookup.
- Answer: A concise response (single word, number, phrase, or Yes/No) derived from the chart's data.
Guidelines for Question Generation:
1. Direct Lookup Questions – extracting specific values from the chart.
2. Comparison Questions – comparing values between different categories.
3. Trend & Pattern Recognition – identifying increases, decreases, or correlations in the data.
4. Inference-Based Questions – requiring reasoning beyond direct value lookup.
Ensure the question is meaningful and the answer is accurate based on the chart data.
Assistant: Question: $\tilde{\mathbf{Q}}$, Answer: $\tilde{\mathbf{A}}$

(iii) OKVQA [146]:

L-VLM prompt used in the PA module of MPA: OKVQA [146]

<image(**I**)>
The objective is to generate a question-answer pair for a Knowledge-based Visual Question Answering (K-VQA) task. Your task is to create a contextually relevant question that directly relates to the image's content while requiring external world knowledge to answer correctly, and its correct answer.
Output Requirements:
- Question: A natural language question grounded in the image's content but requiring reasoning beyond direct perception, incorporating real-world knowledge.
- Answer: A single-word response based on general world knowledge.
Guidelines for Question Generation:
1. Object & Scene Understanding – identifying objects or actions in the image and connecting them to broader knowledge.
2. Commonsense Reasoning – requiring logical deductions about the scene.
3. Cultural & Historical Context – related to well-known historical events, traditions, or cultural references.
4. Scientific & Factual Knowledge – involving basic physics, biology, geography, or general knowledge.
5. Everyday Life & Social Understanding – questions about daily activities, professions, or human behaviors.
Ensure that the generated question is meaningful and requires external knowledge beyond just the image's visual content.
Assistant: Question: $\tilde{\mathbf{Q}}$, Answer: $\tilde{\mathbf{A}}$

Note that, to ensure fair comparison, the pseudo-annotation prompts are same for all variants of L-VLMs used. Further, the prompt we used for QA is '`Answer the following question in a single word or phrase`', which is common for all datasets across all S-VLMs.

7.4.5 QUALITATIVE RESULTS

Figure 7.4 presents a selection of examples where MPA alignment enables S-VLM to correct errors made by the original zero-shot S-VLM. From a rigorous examination of the results, we find that MPA significantly improves performance in visual text reasoning, plot interpretation, and

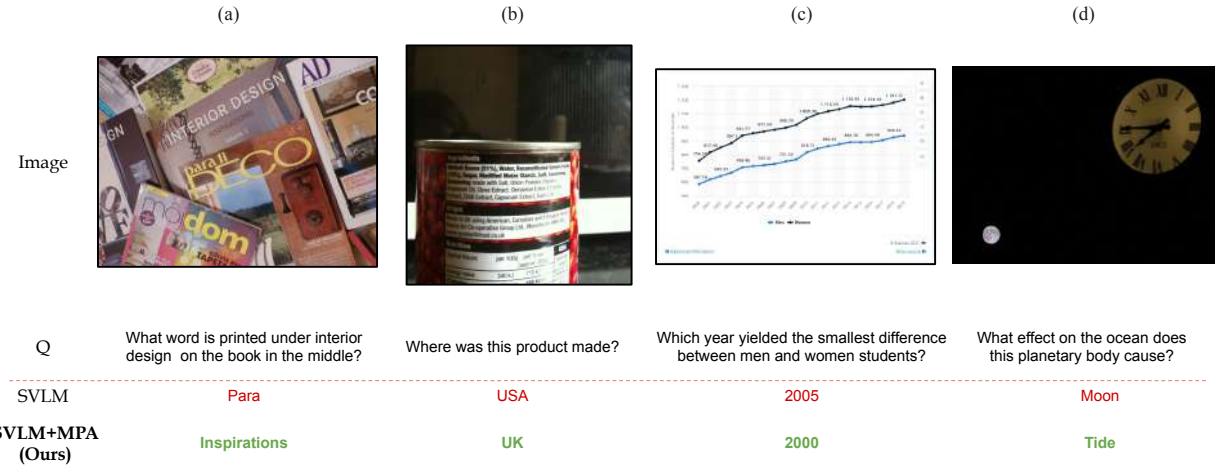


Figure 7.4: A selection of results showing zero-shot S-VLM versus MPA-aligned S-VLM. MPA config: S-VLM: Qwen2VL-2B, L-VLM: Qwen2VL-7B. Green and red text correspond to correct and incorrect answers, respectively. ***(Best viewed in color)***

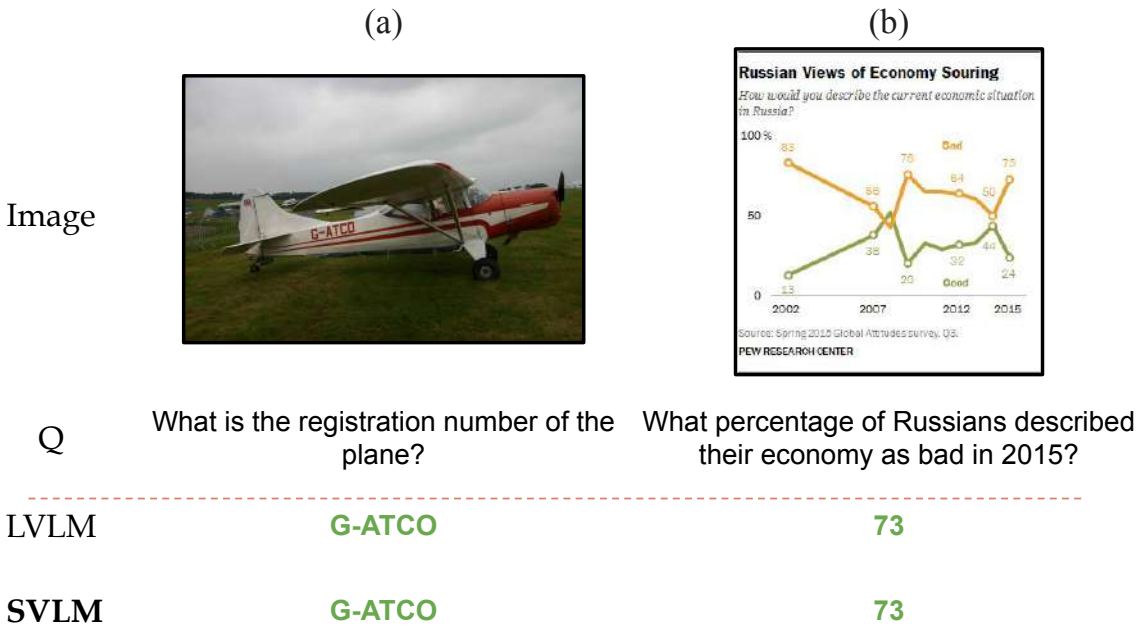


Figure 7.5: Pseudo-annotations discarded by PI module as they do not constitute knowledge-gap.

knowledge-based question answering. Further, we show additional qualitative samples for showing zero-shot S-VLM versus MPA-aligned S-VLM across all four datasets: TextVQA in Figure 7.7, ST-VQA in Figure 7.8, ChartQA in Figure 7.9 and OKVQA in Figure 7.10.

Further, we show a few selected examples of noisy annotations and another set of examples which do not represent a disparity between S-VLM and L-VLM, that are discarded by the PI module in Figure 7.6 and Figure 7.5 respectively.

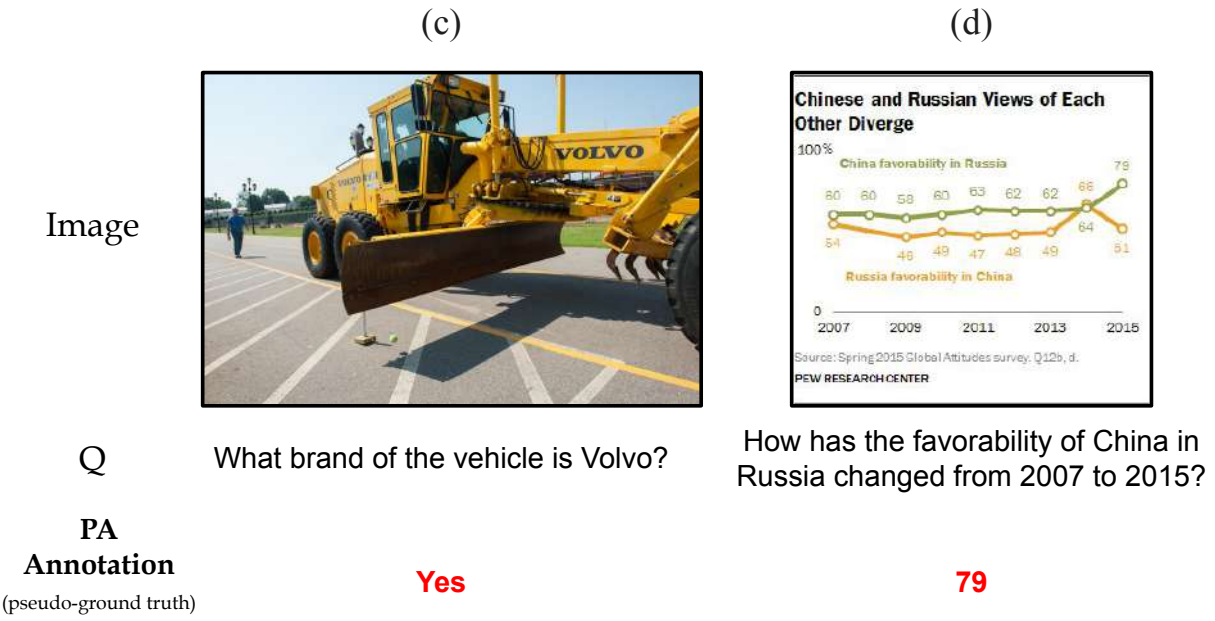


Figure 7.6: Pseudo-annotations discarded by PI module as they are noisy annotations.

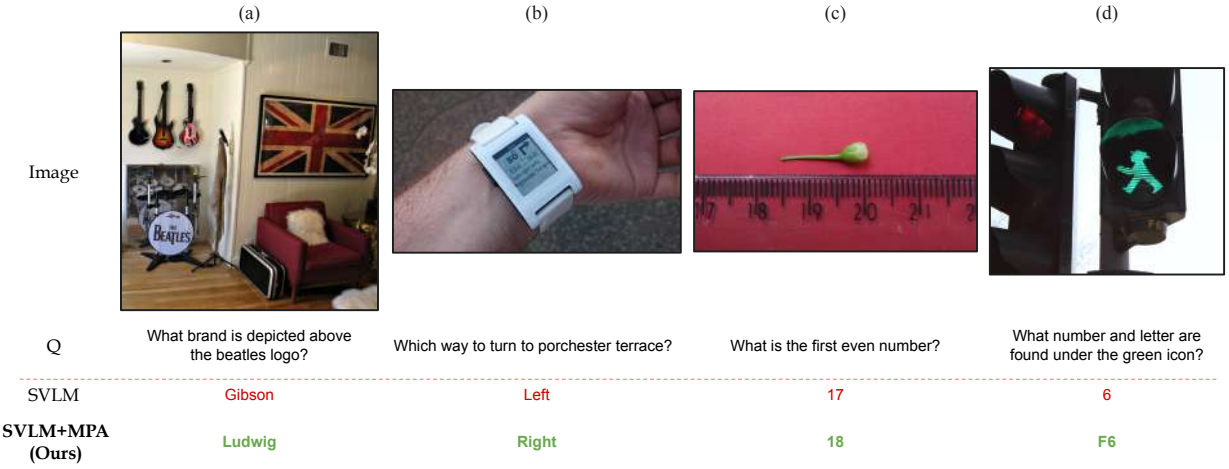


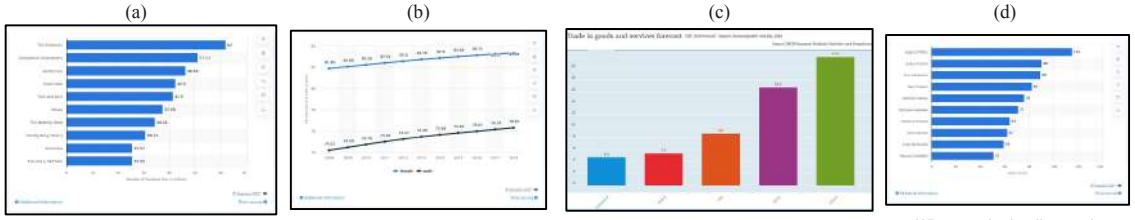
Figure 7.7: Few more results from TextVQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM.

7.5 CONCLUSION AND FUTURE WORK

In this work, we introduced the Model Parity Aligner (MPA), a novel framework that enhances small vision-language models (S-VLMs) by leveraging unlabeled images and effective knowledge transfer from large vision-language models (L-VLMs). Unlike traditional knowledge distillation techniques that rely on labeled data and access to large model logits, MPA employs pseudo-labeling with quality assessment, ensuring that small models learn from high-confidence supervision while avoiding error propagation. Our experiments across four diverse VQA benchmarks, viz. TextVQA, ST-VQA, ChartQA and OKVQA demonstrate that MPA consistently improves S-

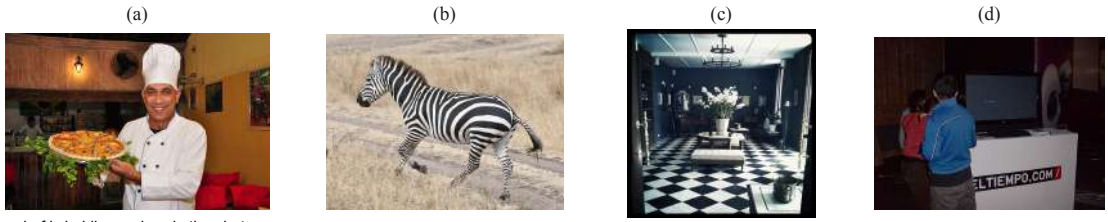
Q	What is the punishment for honking?	What letters come after the letters ATV/ on the same button?	What is printed on the right side of the clock?	What company's logo is in the black box in the upper left?
SVLM	$150 penalty	DVD	Letters	Burberry
SVLM+MPA (Ours)	**$350 penalty**	**DTV**	**1240 KC**	**Gucci**

Figure 7.8: Few more results from STVQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM. Green and red text correspond to correct and incorrect answers, respectively. ***(Best viewed in color)***



Q	How many facebook fans did the simpsons have as of June 2011?	Does the life expectancy decrease over the years?	How many places are mentioned in the graph?	Who was the leading goal scorer for Celtic FC as of September 2020?
SVLM	26.1 million	Yes	10	Scott Sinclair
SVLM+MPA (Ours)	**62 million**	**No**	**5**	**Leigh Griffiths**

Figure 7.9: Few more results from ChartQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM. Green and red text correspond to correct and incorrect answers, respectively. ***(Best viewed in color)***



Q	The chef is holding a pizza in the photo so what type of restaurant does this suggest he may be cooking at?	At what speed does this animal run?	What is the name of the floor pattern?	Can you guess the model of tv shown in this picture?
SVLM	Pizza	10 mph	Diamond	No
SVLM+MPA (Ours)	**Italian**	**30 mph**	**Checkered**	**LG**

Figure 7.10: Few more results from OKVQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM. Green and red text correspond to correct and incorrect answers, respectively. ***(Best viewed in color)***

VLM performance, making them more viable for real-world applications with limited resources.

Despite these improvements, there still remains a gap between S-VLMs and L-VLMs that

highlights the need for further advancements. As future work, we aim to explore more robust knowledge alignment strategies, including iterative refinement of pseudo-labels, leveraging diverse sources of unlabeled data, and integrating multi-step reasoning from L-VLMs into S-VLMs training. Additionally, extending MPA to tasks beyond visual question answering, and exploring teacher-student pairs at smaller scales, could further broaden its applicability across diverse tasks and compute budgets. We view MPA as a first step toward achieving model parity in vision and language models via targeted knowledge alignment, and firmly believe that it shall open up future research avenues for more efficient and capable small models for vision-language tasks.

7.6 LIMITATIONS

Our proposed MPA framework depends on access to a large vision-language model (L-VLM) for generating and validating pseudo-annotations. In even stricter resource-constrained settings, this may limit applicability of MPA. Further, when leveraging proprietary closed-source models via commercial APIs, reproducibility and transparency may be compromised due to limited insight into model behavior and potential changes in API responses over time. Our experiments also focus primarily on English-language datasets and VQA-related tasks; generalization to multilingual, or more complex reasoning tasks remains an open direction.

7.7 ETHICAL CONSIDERATIONS AND BROADER IMPACT

In this work, we used open-source datasets which may contain social or cultural biases. The proposed framework also depends on outputs from large-scale vision-language models (L-VLMs), which are known to occasionally generate hallucinated or biased content. Although the Parity Identifier (PI) module is designed to filter out low-quality or incorrect annotations, it cannot entirely eliminate inherited biases from the underlying L-VLM. Further, this work involves a human evaluation study in which three annotators were employed to assess the quality of pseudo-annotations generated by our MPA framework. All annotators were compensated fairly in accordance with local wage norms. They were not exposed to harmful, offensive, or sensitive content, and no personally identifiable information was collected at any stage of the study.

Broader Impact: The proposed MPA framework enables efficient training of small vision-language models (S-VLMs) using only unlabeled data, reducing reliance on expensive human annotations. By transferring capabilities from large vision-language models (L-VLMs) to compact models, MPA makes high-performing multimodal systems more accessible in low-resource settings. This democratization of vision-language technology can benefit real-world applications in healthcare, agriculture, and accessibility, particularly in regions with limited compute or labeled data. Furthermore, the proposed approach encourages the development of scalable alignment strategies that can generalize to diverse, resource-constrained communities.

8
Conclusions and Future Scope

8.1 CONCLUSION AND FUTURE WORK

This thesis set out to address the challenges of *knowledge-intensive visual tasks*, where perception alone is insufficient and external knowledge must be retrieved, grounded, and reasoned over. While large vision–language models (LVLMs) have advanced rapidly, they continue to face two central challenges. First, they are often not *effective* at reasoning with external knowledge: their reliance on parametric memory leads to hallucinations, limited grounding, and difficulties in multi-image reasoning. Second, they are not *efficient* to train or deploy: scaling up comes at prohibitive computational cost, limiting their accessibility and practicality. This thesis positioned itself within this landscape by tackling these two challenges in parallel, aiming to make VLMs both more effective and more efficient.

To improve effectiveness, we first introduced retrieval-augmented methods that integrate textual knowledge through visual entity linking. This was demonstrated in the context of image retrieval (KRAMT on COFAR) and knowledge-intensive visual question answering (VisTEL and KaLMA for Text-KVQA), showing that grounding visual entities in external knowledge improves factual accuracy. The scope of retrieval was then extended to visual knowledge, where we proposed a multi-image VQA framework (MI-BART on RetVQA) that aggregates supporting evidence across images. Subsequently, we leveraged multimodal retrieval for domain-specific captioning, where our retrieval-augmented framework generated faithful and interpretable descriptions in the fashion domain.

To improve efficiency, we proposed supervision strategies that exploit knowledge disparities between large and small models. This enabled compact VLMs to approximate the performance of much larger models without relying on costly labeled datasets, offering a scalable path toward practical deployment.

Together, these contributions establish a unified research trajectory that advances the state of vision–language modeling along the two complementary dimensions of effectiveness and efficiency. By developing retrieval-augmented methods, this thesis demonstrates how VLMs can

move beyond perception into reliable and factual multimodal reasoning. By proposing efficient training paradigms, it shows how such models can be scaled down without sacrificing performance, making them more practical for real-world use. In doing so, the thesis contributes to the broader goal of building VLMs that are both knowledge-aware and scalable.

Impact of this thesis. The resources and ideas introduced in this thesis are already influencing ongoing research in multimodal AI. The RetVQA [164] benchmark has been adopted by various research groups [123, 236] for developing retrieval-based vision–language systems, and has also enabled further progress: Li et al. [123] reported a new state of the art with 80.3% accuracy, improving upon the 76.5% accuracy achieved by our proposed MI-BART. Beyond this benchmark, ideas from RetVQA [164] and VisTEL–KaLMA [165] have been adapted to knowledge-intensive tasks in the audio domain [163], showing that the approaches developed here generalize beyond vision to other modalities that require external knowledge grounding.

By framing knowledge augmentation and efficiency as the two central challenges for LVLMs, this thesis provides a foundation for future research on scalable and knowledge-aware multimodal reasoning. The methods and datasets proposed here are already contributing to ongoing research and will continue to serve as valuable resources for the community in the years ahead.

8.2 FUTURE WORK

This thesis established retrieval augmentation and targeted supervision as effective directions for improving vision–language models on knowledge-intensive visual tasks. Building on these contributions, we identify four focused directions for future research.

1. ***Resolving the retrieval dilemma.*** A central open challenge is deciding *when* a model should retrieve external knowledge and *when* it should rely on its internal parametric memory. Current retrieval-augmented VLMs often retrieve by default, which increases inference cost and can introduce irrelevant or distracting evidence. Future work should therefore develop retrieval policies that are selective and uncertainty-aware. Such policies could use confidence estimation, answer consistency to determine whether retrieval is necessary, what to retrieve, and how much retrieved context should be trusted. Solving this retrieval dilemma is important for building systems that are not only accurate, but also efficient and robust.
2. ***Uncertainty-aware model parity alignment.*** While MPA showed that parity-guided supervision can effectively transfer knowledge from large models to small ones, the current formulation can be extended further. A promising direction is to make MPA uncertainty-aware, so that pseudo-supervision is guided not only by the knowledge-gap between models, but also by the confidence and reliability of the generated labels from the large VLMs. This would help distinguish samples that represent knowledge-gaps from noisy ones and could improve knowledge transfer across diverse tasks, including open-ended generation, reasoning. Such an extension would make MPA more general, reliable, and better suited for training compact VLMs in low-supervision settings.

3. ***Broader retrieval sources.*** Existing retrieval pipelines in VLMs are mostly restricted to static image–text corpora. An important next step is to expand retrieval to richer sources such as videos, temporal visual streams, structured databases, and specialized knowledge repositories. Access to such heterogeneous retrieval sources would allow VLMs to reason over temporal events, combine structured and unstructured evidence, and better support domain-specific applications. Realizing this direction will require new retrieval and fusion mechanisms that can align information across modalities, time, and knowledge formats.

4. ***Advanced knowledge-intensive tasks.*** The methods in this thesis have primarily been validated on retrieval, question answering, and captioning. However, many real-world applications involve more complex forms of multimodal reasoning where factual grounding is critical. For instance, embodied tasks such as robot navigation under factual constraints or human–robot collaboration require integrating perception, language understanding, and external knowledge in real time. Similarly, tasks in domains like scientific discovery or legal document understanding demand reasoning with highly specialized knowledge. Extending retrieval-augmented frameworks to such advanced knowledge-intensive tasks would push the boundaries of what VLMs can achieve and demonstrate their utility in interactive, high-stakes environments.

Together, these directions offer pathways toward building VLMs that are not only more reliable and knowledge-aware but also more adaptive, scalable, and efficient across a wide range of multimodal settings.

8.3 ETHICAL CONSIDERATIONS AND BROADER IMPACTS

This thesis studies retrieval-augmented and scalable vision-language models using publicly available datasets, with some of them adapted for the specific tasks considered in this work. While these approaches improve performance on knowledge-intensive visual tasks, they also raise important ethical considerations.

First, the datasets and benchmarks used in this thesis may contain geographical, linguistic, and cultural biases. Publicly available image sources are often unevenly distributed across regions and communities, which can limit representation. In addition, parts of this thesis, especially those involving scene text and associated knowledge, are predominantly centered on English-language data, which may reduce generalization to multilingual and culturally diverse settings. Second, several methods in this thesis build on large pretrained vision-language or multimodal foundation models. Such models may inherit and sometimes amplify biases present in their large-scale pretraining corpora, including social, cultural, and demographic biases. They may also produce hallucinated or uneven outputs across domains, languages, and user groups. Third, although the proposed methods are intended for research in retrieval, question answering, and captioning, their real-world deployment should be approached with care. In particular, systems used in public-facing or decision-support settings should undergo additional evaluation for fairness,

robustness, cultural sensitivity, and reliability before use.

Publications

Publications and Manuscripts which are forming part of this thesis

1. Prajwal Gatti, ***Abhirama Subramanyam Penamakuri***, Revant Teotia, Anand Mishra, Shubhashis Sengupta, and Roshni Ramnani. COFAR: Commonsense and Factual Reasoning in Image Search. In *Proceedings of Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (AACL-IJCNLP 2022)*, pages 1185–1199.

2. ***Abhirama Subramanyam Penamakuri***, Manish Gupta, Mithun Das Gupta, and Anand Mishra. Answer Mining from a Pool of Images: Towards Retrieval-Based Visual Question Answering. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pages 1312-1321.

3. ***Abhirama Subramanyam Penamakuri*** and Anand Mishra. Visual Text Matters: Improving Text-KVQA with Visual Text Entity Knowledge-aware Large Multimodal Assistant. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 20675–20688.

4. ***Abhirama Subramanyam Penamakuri****, Navlika Singh*, Piyush Arora*, and Anand Mishra. When Big Models Train Small Ones: Label-Free Model Parity Alignment for Efficient Visual Question Answering using Small VLMs. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2025)*, pages 31644–31661. (*: equal contribution)

5. ***Abhirama Subramanyam Penamakuri****, Shreya Shukla*, and Anand Mishra. RA-CoA: Training-free Fashion Image Captioning via Retrieval-Augmented Chain-of-Attributes, [Under peer review]. (*: equal contribution)

Other Publications and Manuscripts during PhD (not part of this thesis)

6. K Lokesh*, ***Abhirama Subramanyam Penamakuri****, Uday Agarwal, Apoorva Challa, Shreya K Gowda, Somesh Gupta, Anand Mishra. PatientVLM Meets DocVLM: Pre-Consultation Dialogue Between Vision-Language Models for Efficient Diagnosis. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI 2026)*, pages 7485-7493. (*: equal contribution)

7. ***Abhirama Subramanyam Penamakuri****, Kiran Chhatre*, and Akshat Jain. Audiopedia: Audio QA with Knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, pages 1–5. (*: equal contribution)

8. Nakul Sharma, ***Abhirama Subramanyam Penamakuri***, Anand Mishra. Contrastive Multi-View Textual-Visual Encoding: Towards One Hundred Thousand-Scale One-Shot Logo Identification. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2022)*.

9. Somraj Gautam*, ***Abhirama Subramanyam Penamakuri****, Abhishek Bhandari, Gaurav Harit. Mind the (Language) Gap: Towards Probing Numerical and Cross-Lingual Limits of LVLMs. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL) at Empirical Methods in Natural Language Processing (EMNLP 2025)*, pages 568–584. (*: equal contribution)

10. Anik De, ***Abhirama Subramanyam Penamakuri***, Harshiv Shah, Aditya Rathore, Devesh Sharma, Rajeev Yadav, Sagar Agarwal, Pravin Kumar, Anand Mishra. Bharat Scene Text Dataset: A Novel Dataset and Benchmark for Indian Language Scene Text Understanding. Accepted for publication in *International Journal on Document Analysis and Recognition (IJDAR) (ICDAR-IJDAR journal track)*.

Bibliography

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. Reqa: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[4] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140, 2019.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[6] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[8] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[9] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[11] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[12] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[13] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weaqa: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, 2021.

[14] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005.

[15] Ankan Bansal, Yuting Zhang, and Rama Chellappa. Visual question answering on image sets. In *ECCV*, 2020.

[16] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022.

[17] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

[18] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[19] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE, 2019.

[20] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, 2019.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[22] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024.

[23] Mathilde Caron, Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. A generative approach for wikipedia-scale visual entity recognition. In *CVPR*, 2024.

[24] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *CVPR*, 2022.

[25] Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1947–1963, 2022.

[26] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[27] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.

[28] Long Chen, Yuhang Zheng, and Jun Xiao. Rethinking data augmentation for robust visual question answering. In *European conference on computer vision*, pages 95–112. Springer, 2022.

[29] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE international conference on computer vision*, pages 1409–1416, 2013.

[30] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *EMNLP*, 2023.

[31] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[32] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[33] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, 2016.

[34] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[35] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.

[36] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[37] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[38] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.

[39] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[40] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2024.

[41] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[42] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[43] DemandSage. How many google searches per day [2025 data], 2025. Accessed: 2025-06-25.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[45] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

[47] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[48] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[49] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 2025.

[50] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[51] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018.

[52] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[53] Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054, 2018.

[54] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[55] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260, 2020.

[56] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015.

[57] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, 2020.

[58] Prajwal Gatti, Abhirama Subramanyam Penamakuri, Revant Teotia, Anand Mishra, Shubhashis Sengupta, and Roshni Ramnani. Cofar: Commonsense and factual reasoning in image search. In *AACL-IJCNLP*, 2022.

[59] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14105–14115, June 2022.

[60] Shahriar Golchin, Mihai Surdeanu, Steven Bethard, Eduardo Blanco, and Ellen Riloff. Memorization in in-context learning. *arXiv preprint arXiv:2408.11546*, 2024.

[61] Google. Product information preference statistics. `https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/product-information-preference-statistics/`, 2019. Accessed: 2026-01-17.

[62] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[63] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[64] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[65] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, 2022.

[66] Dalu Guo, Chang Xu, and Dacheng Tao. Bilinear graph networks for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[67] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[68] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *European conference on computer vision*, pages 634–651. Springer, 2022.

[69] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680, 2023.

[70] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[72] Pengfei He, Yingqian Cui, Han Xu, Hui Liu, Makoto Yamada, Jiliang Tang, and Yue Xing. Towards the effect of examples on in-context learning: A theoretical case study. *Stat*, 14(1):e70045, 2025.

[73] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[74] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[75] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[76] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[77] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[78] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023.

[79] Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. Answer generation for retrieval-based question answering systems. In *ACL/IJCNLP*, pages 4276–4282, 2021.

[80] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[81] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[82] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *ICCV*, 2023.

[83] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 804–813, 2017.

[84] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379, 2023.

[85] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[86] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[87] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[88] HuggingFaceTB. smolllm: A collection of small language models. `https://huggingface.co/collections/HuggingFaceTB/smollm-6695016cad7167254ce15966`, 2023. Accessed: 2025-03-06.

[89] Soumya Jahagirdar, Shankar Gangisetty, and Anand Mishra. Look, read and ask: Learning to ask questions by reading text in images. In *International Conference on Document Analysis and Recognition*, pages 335–349, 2021.

[90] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6485–6494, 2017.

[91] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[92] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017.

[93] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015.

[94] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[95] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA, 2009.

[96] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of the 10th international conference on natural language generation*, pages 198–202, 2017.

[97] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613, 2021.

[98] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukás Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 1156–1160, 2015.

[99] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[100] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[101] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.

[102] Mahmoud Khademi, Ziyi Yang, Felipe Vieira Frujeri, and Chenguang Zhu. Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[103] Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, and Manmohan Chandraker. Q: How to specialize large vision-language models to data-scarce vqa tasks? a: Self-train on unlabeled images! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15005–15015, 2023.

[104] Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6361, 2021.

[105] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, pages 1571–1581, 2018.

[106] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, pages 5583–5594, 2021.

[107] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[108] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 24872–24895, 2024.

[109] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2008–2018, 2019.

[110] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[111] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[112] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

[113] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[114] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *ACL*, pages 8211–8225, 2020.

[115] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[116] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 2020.

[117] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[118] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[119] Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2024.

[120] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[121] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems NeurIPS*, pages 9694–9705, 2021.

[122] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, 2019.

[123] Peize Li, Qingyi Si, Peng Fu, Zheng Lin, and Yan Wang. Multimodal hypothetical summary for retrieval-based multi-image question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4851–4859, 2025.

[124] Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. Understanding retrieval robustness for retrieval-augmented image captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9299, 2024.

[125] Xianrui Li, Zhiling Ye, Zhao Zhang, and Mingbo Zhao. Clothes image caption generation with attribute detection and visual attention model. *Pattern Recognition Letters*, 141:68–74, 2021.

[126] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[127] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481, 2020.

[128] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[129] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[130] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571, 2022.

[131] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[132] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[133] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[134] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[135] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[136] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[137] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[138] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

[139] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[140] Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

[141] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[142] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[143] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179, 2024.

[144] Andres Marafioti. Smolvlm - small yet mighty vision language model, 2024.

[145] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021.

[146] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[147] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022.

[148] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[149] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 2013.

[150] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[151] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL*, 2016.

[152] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31, 2018.

[153] Medhini Narasimhan and Alexander G Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 451–468, 2018.

[154] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 30–38, 2016.

[155] OpenAI. Gpt-4 api documentation. OpenAI API Documentation, 2024. Accessed: 2024-02-16.

[156] OpenAI. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

[157] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[158] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[159] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[160] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. VisualCOMET: Reasoning about the dynamic context of a still image. In *ECCV*, 2020.

[161] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[162] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[163] Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. Audiopedia: Audio qa with knowledge. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[164] Abhirama Subramanyam Penamakuri, Manish Gupta, Mithun Das Gupta, and Anand Mishra. Answer mining from a pool of images: Towards retrieval-based visual question answering. In *IJCAI*. ijcai.org, 2023.

[165] Abhirama Subramanyam Penamakuri and Anand Mishra. Visual text matters: Improving text-KVQA with visual text entity knowledge-aware large multimodal assistant. In *EMNLP*, 2024.

[166] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36, 2024.

[167] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[168] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[169] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[170] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[171] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[172] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018.

[173] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[174] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *ACL*, 2018.

[175] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, 2023.

[176] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023.

[177] Leonardo Ranaldi and Andre Freitas. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, 2024.

[178] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

[179] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[180] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

[181] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[182] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[183] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[184] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[185] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162, 2022.

[186] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.

[187] Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, and Jiajun Ding. Imp: Highly capable large multimodal models for mobile devices. *arXiv preprint arXiv:2405.12107*, 2024.

[188] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[189] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[190] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016.

[191] Ilya Shnayderman, Liat Ein-Dor, Yosi Mass, Alon Halfon, Benjamin Sznajder, Artem Spector, Yoav Katz, Dafna Sheinwald, Ranit Aharonov, and Noam Slonim. Fast end-to-end wikification. *CoRR*, abs/1908.06785, 2019.

[192] Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024.

[193] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758, 2020.

[194] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[195] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *ICCV*, pages 4602–4612, 2019.

[196] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[197] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. Mimoqa: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, 2021.

[198] Victoria Stodden. The data science life cycle: a disciplined approach to advancing data science as a science. *Commun. ACM*, 63(7):58–66, 2020.

[199] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[200] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. In *CVPR*, pages 7736–7745, 2018.

[201] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

[202] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[203] Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. Visual named entity linking: A new dataset and A baseline. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *EMNLP (Findings)*, 2022.

[204] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[205] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.

[206] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, 2019.

[207] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016.

[208] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[209] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[210] Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation. *arXiv preprint arXiv:2402.04616*, 2024.

[211] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[212] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[213] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[214] Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. Guiding visual question generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1640–1654, 2022.

[215] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.

[216] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[217] Guo-Hua Wang, Yifan Ge, and Jianxin Wu. Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8183–8195, 2021.

[218] Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph F. JáJá, and Larry Davis. TAG: boosting text-vqa via text-aware visual question-answer generation. In *BMVC*, 2022.

[219] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[220] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.

[221] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[222] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017.

[223] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1290–1296, 2017.

[224] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020.

[225] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[226] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[227] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV*, 2024.

[228] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[229] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[230] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020.

[231] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[232] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[233] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021.

[234] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630, 2016.

[235] Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025.

[236] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

[237] Wansen Wu, Tao Chang, and Xinmeng Li. Vision-language navigation: A survey and taxonomy. *arXiv preprint arXiv:2108.11544*, 2021.

[238] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[239] Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. Grounding language models for visual entity recognition. *arXiv preprint arXiv:2402.18695*, 2024.

[240] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[241] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[242] Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[243] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR, 2016.

[244] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[245] Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. Feature normalized knowledge distillation for image classification. In *European conference on computer vision*, pages 664–680. Springer, 2020.

[246] Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

[247] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[248] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. *arXiv preprint arXiv:2202.10401*, 2022.

[249] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 1–17. Springer, 2020.

[250] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.

[251] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[252] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[253] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[254] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. In *EMNLP*, pages 2115–2129, 2021.

[255] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[256] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *CoRR*, 2021.

[257] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[258] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14100–14110, 2024.

[259] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[260] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[261] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

[262] Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. Automatic generation of grounded visual questions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4235–4243, 2017.

[263] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[264] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Enhanced visual instruction tuning for text-rich image understanding. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[265] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[266] Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. Unifashion: A unified vision-language model for multimodal fashion retrieval and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1490–1507, 2024.

[267] Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. Visual entity linking via multimodal learning. *Data Intell.*, 4(1):1–19, 2022.

[268] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

[269] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, 2020.

[270] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.

[271] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[272] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.

[273] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12647–12657, 2021.