

UCLA

UCLA Electronic Theses and Dissertations

Title

Advancing Large Vision-Language Models with Efficiency, Reliability and Visual Knowledge

Permalink

<https://escholarship.org/uc/item/41n8q0hf>

Author

Hu, Wenbo

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Advancing Large Vision-Language Models with
Efficiency, Reliability and Visual Knowledge

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Wenbo Hu

2024

© Copyright by

Wenbo Hu

2024

ABSTRACT OF THE THESIS

Advancing Large Vision-Language Models with
Efficiency, Reliability and Visual Knowledge

by

Wenbo Hu

Master of Science in Computer Science

University of California, Los Angeles, 2024

Professor Nanyun Peng, Chair

Humans perceive and understand the world primarily through the complementary modalities of vision and language. The development of Large Vision-Language Models (LVLMs) marks a significant step toward realizing general artificial intelligence.

This thesis advances LVLMs through three critical aspects: (1) improving the efficiency of LVLMs by controlling visual input representations, (2) evaluating hallucination and informativeness in LVLMs, and (3) assessing the integration of visually augmented knowledge. We introduce MQT-LLaVA, a model capable of elastically encoding an image into a variable number of visual tokens, enabling dynamic visual processing. Additionally, we present VALOR-EVAL, a two-stage evaluation system leveraging human-annotated data to address not only object hallucination but also nuanced issues of attribute and relational hallucinations, as well as informativeness. Furthermore, MRAG-Bench systematically identifies and categorizes scenarios where visually augmented knowledge outperforms textual knowledge, providing valuable insights into retrieval-augmented LVLMs. The thesis concludes by analyzing the latest advances in LVLMs, highlighting their current challenges, and exploring future directions for research and development.

The thesis of Wenbo Hu is approved.

Kai-Wei Chang

Yuchen Cui

Nanyun Peng, Committee Chair

University of California, Los Angeles

2024

*To my family and my love
who taught me to
dream,
persevere, and
be grateful.*

Contents

Abstract	ii
List of Figures	vii
List of Tables	viii
Acknowledgements	xii
1 Introduction	1
1.1 Motivation	1
1.2 Thesis Contribution	2
1.3 Thesis Overview	2
2 Dynamic Visual Tokens Advances LVLM Efficiency	5
2.1 Introduction	5
2.2 Matryoshka Query Transformer	7
2.3 Experiments	10
2.4 Analysis	13
2.5 Conclusion	18
3 Advance Hallucination and Informativeness Evaluation of LVLM	19
3.1 Introduction	19
3.2 VALOR-BENCH	21
3.3 VALOR-EVAL	26
3.4 Experiment	28
3.5 Conclusion	34
4 Advance LVLM with Visual Retrieval-Augmented Knowledge	35
4.1 Introduction	35
4.2 MRAG-BENCH	38
4.3 Experiments	43
4.4 Analysis	46
4.5 Conclusion	50
5 Future Work and Conclusion	51
5.1 Current Challenges and Future Prospects	51
5.2 Conclusion	53

Appendices	54
A Additional Results from Chapter 2	55
A.1 Additional Results	55
B Additional Results from Chapter 3	57
B.1 Large Vision-Language Models	57
B.2 Conditional Probabilities	58
B.3 Captions Generation Prompts	59
B.4 Features Extraction Prompts	59
B.5 Features Matching Prompts	59
B.6 Qualitative Results	60
C Additional Results from Chapter 4	72
C.1 MRAG-BENCH Details	72
C.2 Experiment Setting Details	79
C.3 More Results	82

List of Figures

2.1	Our model, MQT-LLAVA, matches LLaVA-1.5 performance on 11 benchmarks using only 256 visual tokens instead of 576. We achieve a 2x speed-up with 256 tokens and 8X speed-up in TFLOPs using 16 tokens with only a 2.4 performance drop compared to LLaVA-1.5 on MMBench.	6
2.2	Our model employs a query transformer to encode images as visual tokens. We randomly select the first m tokens during training, and enable flexible choice of <i>any</i> m number under M during inference, where M is the maximum number of initialized tokens.	8
2.3	With only 2 visual tokens, MQT-LLAVA outperforms InstructBLIP (which uses 32 visual tokens) on all 8 benchmarks it is evaluated on.	12
2.4	Grad-CAM visualization of 1 randomly picked token from using 8, 16, 64, 256 visual tokens, respectively, to encode an image. The model effectively concentrates on high-level concepts using fewer tokens and delves into low-level details with more tokens. The complete input to the third image is “List all the objects on the desk. The objects on the desk include a computer monitor, a keyboard, a mouse, a cell phone, and a pair of headphones”.	13
2.5	The number of visual tokens impact different tasks differently (x-axis is in log-scale). Our model’s performance on ScienceQA, MME-Cognition and MMMU is remarkably robust to token reduction.	14
2.6	Examples from MME Cognition. Grad-CAM results are from using 16 tokens which answered all the questions correctly.	16
2.7	Comparison of correct and failure cases in 16 vs 144 visual tokens on Science-QA (test-set).	17
3.1	Overview of VALOR-EVAL evaluation framework: (1) Firstly, LVLMS generate captions from VALOR-BENCH benchmark images. (2) Following this, LLMs are employed to <i>extract</i> pivotal features that encapsulate from the generated descriptions. (3) Subsequently, these features are <i>aligned</i> with a pre-defined list of ground-truth features using LLMs, facilitating the creation of two essential outputs: a dictionary of matched features and a more extensive dictionary encompassing broader conceptual matches. (4) Finally, we calculate two key metrics: <i>faithfulness</i> and <i>coverage</i> . These metrics measure the LVLMS’ comprehension by evaluating how well the generated captions encapsulate the salient features of the images and the breadth of concepts they cover, respectively.	22

4.1	Example scenarios from MRAG-BENCH. Previous benchmarks [Chang et al., 2022, Chen et al., 2023c, Mensink et al., 2023] mainly focused on retrieving from textual knowledge. However, there are scenarios where retrieving correct textual knowledge is hard and sometimes not as useful as visual knowledge.	36
4.2	Scenarios distribution of MRAG-BENCH.	38
4.3	Qualitative examples on MRAG-BENCH. For each scenario, we show the result of GPT-4o [OpenAI, 2023], Gemini Pro [Team et al., 2023], LLaVA-Next-Interleave [Li et al., 2024b] and Mantis-8B-Siglip [Jiang et al., 2024a]. The ground-truth answer is in blue.	40
4.4	Qualitative Example of Proprietary model (Gemini Pro) identifies and utilizes correct examples, while open-source model (LLaVA-Next-Interleave) is misled by noisy retrieved information, resulting in incorrect answers.	47
4.5	Left: LLaVA-Next-Interleave results with 4 different multimodal retrievers. Its performance using retrieved images correlates 95% with retriever’s Recall@5 scores. Right: Average results of three random seed runs. Improve the number of ground-truth RAG examples shows steady increase of model’s performance, reaches the maximum with 10 examples.	49
A.1	Grad-CAM visualization from all the tokens in our model when inference with 8 tokens. Input: “How many cats are there in the image? Answer: 2”.	56
B.1	Object existence evaluation example from three representative models in our benchmark VALOR-BENCH. Text in red indicating models’ hallucinations.	69
B.2	Positional relation evaluation example from three representative models in our benchmark VALOR-BENCH. Text in red indicating models’ hallucinations.	70
B.3	Comparative relation evaluation example from three representative models in our benchmark VALOR-BENCH. Text in red indicating models’ hallucinations.	71
C.1	Human evaluation interface without RAG examples	80
C.2	Human evaluation interface with ground-truth RAG examples	81

List of Tables

2.1	Comparison with state-of-the-art methods on 11 vision-language benchmarks. Our model (MQT-LLAVA) with up to 256 tokens achieves on par or better than LLaVA-1.5 performance across 11 benchmarks, outperforming it on 6 of 11 benchmarks. We mark the best performance in bold and the second-best <u>underlined</u> . #Tokens is the number of visual tokens used during inference. Avg is the normalized average across 11 benchmarks, out of 100. Benchmark names are abbreviated for brevity: SQA ^I : ScienceQA-IMG, MME ^P : MME Perception, MME ^C : MME Cognition, MMB: MMBench, LLaVA ^W : LLaVA-Bench (In-the-Wild). *The training images of the datasets are observed during training.	11
2.2	For simplicity in ablation studies, we evaluate all the models with 256 visual tokens. All models are trained with the same hyperparameters.	18
3.1	Comparison of existing hallucination evaluation benchmarks for LVLMs, including POPE Li et al. [2023c], HaELM Wang et al. [2023a], HallusionBench Guan et al. [2023], Halle-Switch Zhai et al. [2023], NOPE Lovenia et al. [2023], Bingo Cui et al. [2023], FaithScore Jing et al. [2023], AMBER Wang et al. [2023b], MERLIM Villa et al. [2023]. ? refers to features not explicitly mentioned in the paper. Open Vocab represents evaluating free-form generated captions without constraints to pre-defined vocabulary.	20
3.2	In the VALOR-BENCH benchmark, we categorize images into three main areas: object existence, attributes, and relations, as outlined in Section 3.2.1 and Section 3.2.1. Attributes are further split into <i>object</i> (focusing on color and count of each item not related to people) and <i>people</i> (emphasizing the attire colors and the total number of individuals. For relations, we examine both <i>positional</i> relations between objects and <i>comparative</i> sizes.	26
3.3	The overall evaluation results of object existence, attribute, and relation hallucination in VALOR-BENCH using GPT-4 as the LLM Agent within VALOR-EVAL. The highest is highlighted in blue , while the worst performance is highlighted in yellow . Faithfulness and coverage scores are in percentage (%). For images that contain people, GPT-4V refrains from generating comments, and we marked this score with an asterisk (*).	29
3.4	Pearson correlation (ρ) between our GPT-4-based evaluation framework VALOR-EVAL and human judgements.	31
3.5	Model performance comparison on our data selection method against random selection. Faithfulness and coverage scores are in percentage (%).	32

3.6	Comparison of LLM-augmented CHAIR with original CHAIR metric. Here, F and C denote faithfulness and coverage scores in percentage (%). Acc (F) represents the average percentage of hallucinated objects detected by the metric. Acc (C) denotes the average percentage of objects detected by metric.	33
4.1	Compared with previous works, MRAG-BENCH focuses on evaluating LVLMs in utilizing vision-centric retrieval-augmented multimodal knowledge. “Diverse scenarios” refers to whether a benchmark categorized different scenarios during evaluation. 🌐: Web, 🗺️: ImageNet [Russakovsky et al., 2015], 🌻: Flowers102 [Nilsback and Zisserman, 2008], 🚗: StanfordCars [Krause et al., 2013].	37
4.2	Key statistics of MRAG-BENCH.	38
4.3	Accuracy scores on MRAG-BENCH. The highest scores for open-source models in each section and proprietary models are highlighted in blue and red, respectively. Both Retrieved RAG and GT RAG employ top-5 image examples (except for the incomplete scenario, where a single example is intuitively sufficient). The relative difference in performance compared to the score without RAG is shown in subscript, with blue indicating performance drops and red indicating improvements.	44
4.4	LVLMs performance on MRAG-BENCH with textual knowledge v.s visual knowledge. Both the open-source and proprietary model benefit more from image knowledge.	48
A.1	Results of MQT-LLAVA with different numbers of visual tokens. To demonstrate our flexibility in selecting any number of tokens up to 256, we chose a random number of visual tokens during inference, 77, which was not seen during training.	55
B.1	Architectures of mainstream LVLMs evaluated in our benchmark. InstructBLIP Dai et al. [2023], LLaVA-1.5 Liu et al. [2023a], MiniGPT-v2 Chen et al. [2023a], mPLUG-Owl2 Ye et al. [2023b], BLIVA Hu et al. [2024c], CogVLM Wang et al. [2024], InternLM-XComposer2 Dong et al. [2024], Qwen-VL Bai et al. [2023], Emu2 Sun et al. [2024] and GPT-4V OpenAI [2023].	57
B.2	Prompt template for extracting objects . {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.	60
B.3	Prompt template for extracting attributes (object) . {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.	61
B.4	Prompt template for extracting attributes (people) . {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.	61
B.5	Prompt template for extracting positional relations . {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.	62
B.6	Prompt template for extracting comparative relations . {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.	63
B.7	Prompt template for matching objects in image caption and reference caption. {In-context examples} are in-context examples. {Input Ground Truth Objects} are the ground truth objects list {Input Generated Objects} are the extracted objects list from the extraction step which are originally captions generated by evaluated models.	64

B.8	Prompt template for matching attributes (object) in image caption and reference caption.	65
B.9	Prompt template for matching attributes (people) in image caption and reference caption.	66
B.10	Prompt template for matching positional relations in image caption and reference caption. {In-context examples} are in-context examples. {Input Ground Truth Relations} are the ground truth relation list {Input Generated Relations} are the extracted relation list from the extraction step which are originally captions generated by evaluated models.	67
B.11	Prompt template for matching comparative relations in image caption and reference caption.	68
C.1	Prompt template to extract multiple choice answer from model’s response. {In-context examples} are in-context examples.	82
C.2	Recall@5 scores with 4 retriever models on MRAG-BENCH.	82
C.3	LLaVA-Next-Interleave accuracy scores on MRAG-BENCH with 4 different retrievers.	82

Acknowledgements

I would like to extend heartfelt gratitude to Professor Nanyun Peng and Professor Kai-Wei Chang for their unwavering support of my work and for the invaluable mentorship and guidance they have provided throughout my Master's program. I am also deeply appreciative of the many enriching interactions with my colleagues in the UCLA NLP Lab and PlusLab. Additionally, this thesis has greatly benefited from two published papers and one manuscript under review, co-authored with the outstanding team of Amita Kamath, Haoyi Qiu, Jia-Chen Gu, Liunian Harold Li, Mohsen Fayyaz, Pan Lu, and Zi-Yi Dou, listed alphabetically.

Chapter 1

Introduction

1.1 Motivation

The quest to develop AI systems that emulate human-like intelligence remains one of the most profound challenges in artificial intelligence. Humans excel at integrating sensory perceptions with linguistic and cognitive understanding, enabling them to perceive, reason, and interact seamlessly with their environment. However, even as machine learning models achieve remarkable milestones, they often fall short of replicating this holistic understanding. My thesis is motivated by the promise of bridging this gap through the development of Large Vision-Language Models (LVLMs), which aim to combine visual and linguistic modalities to address complex, real-world tasks.

Despite their transformative potential, LVLMs face three critical challenges that must be overcome to advance the field: improving computational efficiency, ensuring trustworthiness through robust evaluation, and enabling seamless integration of visual knowledge into reasoning processes. Addressing these challenges not only holds the potential to enhance the reliability and scalability of LVLMs but also paves the way for creating AI systems that can better align with human cognitive capabilities. This thesis seeks to tackle these challenges through a series of innovations, contributing to the evolution of multimodal AI and its applicability in real-world scenarios.

The interplay of vision and language is central to human cognition. From interpreting a bustling

cityscape to reasoning about an abstract painting, humans effortlessly combine visual perception with linguistic and conceptual understanding to navigate the world. Emulating this ability has been a long-standing goal in AI research, with the emergence of Large Vision-Language Models (LVLMs) marking a significant step forward. By integrating visual and textual data at scale, LVLMs hold the promise of tackling complex tasks that require multimodal reasoning.

1.2 Thesis Contribution

This thesis advances the development of LVLMs across three interrelated dimensions. First, it addresses the need for computational efficiency by introducing MQT-LLaVA, a novel model capable of elastically encoding images into variable numbers of visual tokens. This enables adaptive visual processing, reducing computational costs without compromising accuracy. Second, it examines the reliability of LVLM outputs by presenting VALOR-EVAL, a robust evaluation framework designed to assess hallucination and informativeness, providing nuanced insights into model trustworthiness. Finally, the thesis explores the integration of visual knowledge with MRAG-Bench, a benchmark that systematically evaluates how visually augmented knowledge can enhance reasoning compared to purely textual approaches.

Through these contributions, this work not only advances the technical foundations of LVLMs but also provides actionable frameworks for evaluating their performance and guiding future research. By addressing efficiency, reliability, and knowledge integration, this thesis aims to move LVLMs closer to achieving the human-like intelligence needed to solve real-world challenges.

1.3 Thesis Overview

The remainder of this thesis is organized as follows:

Chapter 2 introduces Matryoshka Query Transformer (MQT), which allows for a flexible choice of the number of visual tokens and accommodates varying computational constraints in different tasks. Leveraging MQT, we build MQT-LLaVA, a vision-language model that matches

the performance of LLaVA-1.5 using less than half the number of visual tokens, and outperforms it in 6 out of 11 benchmarks. We further explore the performance and computation trade-offs across 11 tasks and demonstrate that a significant speed-up can be achieved with minimal performance drop by reducing the number of visual tokens (e.g., 8X fewer TFLOPs with 2.4 points drop on MMBench)

Chapter 3 introduces VALOR-BENCH, a comprehensive human-annotated dataset covering relation, attribute, object with challenging images selected based on associative bias. We propose an LLM-based two-stage evaluation framework VALOR-EVAL that generalizes previous methods to consider the precision and informativeness trade-off and handle object, attribute, and relation evaluation in open vocabulary settings. We evaluate 10 mainstream LVLMs on VALOR-BENCH, focusing on the balance between faithfulness and coverage score. We notice that even GPT-4V(ision) still suffers from hallucination, achieving a relatively low faithfulness score despite covering more information within an image compared to other models.

Chapter 4 introduces the first visual centric RAG benchmark that focuses on utilization of visual information, unlike previous benchmarks focusing on retrieving and utilizing external textual knowledge for question answering. MRAG-Bench consists of 16,130 images and 1,353 human-annotated multiple-choice questions across 9 distinct real-world scenarios, evaluated on 14 large vision-language models (LVLMs). Extensive experiments demonstrated that visual augmentation provides greater utility than textual information in our benchmark, offering valuable insights such as 1) how retrieved visual knowledge benefits LVLMs, 2) GPT-4o lags significantly behind human performance in visual information utilization, and 3) Open-source models are more susceptible to noisy retrieved examples than proprietary models.

Chapter 5 concludes the thesis by discussing future work and the broader impact of the research, emphasizing the potential of AI systems to achieve human-like multimodal intelligence. It outlines key challenges and promising directions for advancing Large Vision-Language Models (LVLMs), including handling long-context scenarios, improving efficiency through lightweight architectures, expanding into new domains such as 3D and embodied AI, and integrating external tools for real-

world applications. By addressing these challenges, this research lays the groundwork for the next generation of LVLMs, with the potential to transform AI's role in solving complex, real-world problems.

Chapter 2

Dynamic Visual Tokens Advances LVLM Efficiency

2.1 Introduction

Recent work in Large Vision-Language Models (LVLMs) [Bai et al., 2023, Liu et al., 2023b, OpenAI, 2023] has shown remarkable performance across a broad range of vision-language tasks [Cai et al., 2024, Chen et al., 2023b, Huang et al., 2023b, Li et al., 2023b]. These LVLMs typically consist of a vision encoder to embed images into grid features, which are fed into a Large Language Model (LLM) [Chiang et al., 2023, Touvron et al., 2023b] for processing and reasoning alongside a text input.

A key research question is how to transform these raw visual embeddings into the visual tokens that are fed into the LLM. Prior work either directly projects the grid features with a multi-layer perceptron (MLP) [Liu et al., 2023b] or compresses the grid features into fewer tokens with a query transformer or resampler [Alayrac et al., 2022, Bai et al., 2023, Dai et al., 2023, Li et al., 2023a, Ye et al., 2023a]. However, these models all need to pre-determine how many tokens an image is worth,

The contents of this chapter appeared in paper Hu et al. [2024a]

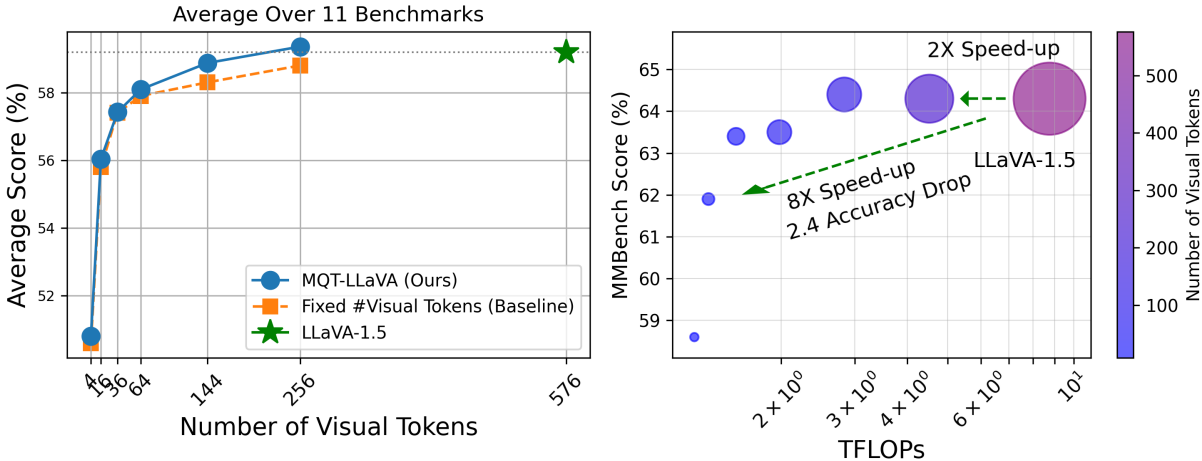


Figure 2.1: Our model, MQT-LLaVA, matches LLaVA-1.5 performance on 11 benchmarks using only 256 visual tokens instead of 576. We achieve a 2x speed-up with 256 tokens and 8X speed-up in TFLOPs using 16 tokens with only a 2.4 performance drop compared to LLaVA-1.5 on MMBench.

and set a fixed number for all images. Finding a *flexible number* that adaptively strikes a balance between efficiency and performance is difficult. More visual tokens encode more information, but come at a higher inference cost, as the complexity of the transformers used in these LVLMs scales quadratically with the number of input tokens. Additionally, not all applications require or allow the same token budget: some applications have limited computational resources, necessitating a lower token budget to ensure real-time processing. In practice, most best-performing LVLMs choose a fixed, large number of visual tokens per image (e.g., 576 for LLaVA-1.5) without the ability to adaptively adjust the visual token allocation at deployment time.

In this work, inspired by Matryoshka Representation Learning (MRL) [Kudugunta et al., 2023, Kusupati et al., 2022], we introduce Matryoshka Query Transformer (MQT), a simple way to train a single LVLm that supports adaptively changing the number of visual tokens at inference time. We use a query transformer [Alayrac et al., 2022, Li et al., 2022] with M latent query tokens to transform grid features into visual tokens. Crucially, during each training step, we train the model using only the first m latent query tokens while dropping the rest, where m is randomly selected within the range of M . With such a tail-token dropping strategy, the query tokens form a Matryoshka structure. Intuitively, the significance of each token correlates with its placement within this nested

structure. During inference, we have the flexibility to selectively utilize solely the initial m visual tokens.

We combine MQT with LLaVA-1.5: the resulting model, MQT-LLAVA, is able to match LLaVA-1.5 performance across 11 benchmarks using only a maximum of 256 tokens, instead of LLaVA’s fixed 576. When the maximum number of tokens is dropped drastically to only 2 tokens, MQT-LLAVA performance drops by only 3% on ScienceQA and 6% on MMMU. Finally, we study the performance of 2, 4, 8, 16, 36, 64, 144, and 256 visual tokens during inference across 11 benchmarks, and offer a trade-off in the selection of visual tokens that balances achieving the highest accuracy with minimizing computational costs on different tasks. Interestingly, we find that changing the number of visual tokens impacts different tasks very differently. For instance, tasks involving language-based reasoning and subject-level scientific knowledge can achieve excellent performance with only a few tokens, whereas complex open-ended visual question tasks that involve rich local information details require a larger number of tokens.

2.2 Matryoshka Query Transformer

2.2.1 Preliminary: Matryoshka Representation Learning (MRL).

MRL [Kudugunta et al., 2023, Kusupati et al., 2022] involves training models with nested dimensions to learn representations at multiple granularities, enabling adaptive deployment per computational constraints. MRL defines a series of models f_1, f_2, \dots, f_M with the same input and output space but growing hidden dimensions.

The name “Matryoshka” comes from the fact that the parameters of f_m are contained by f_{m+1} . For example, in Kudugunta et al. [2023], $\{f_m\}$ are a series of Transformers with the same depth but different widths. Consider a specific Feed Forward Network (FFN) block in f_M that has d_M neurons in the hidden layer. Then, the FFN block in f_m will contain the first d_m neurons, and

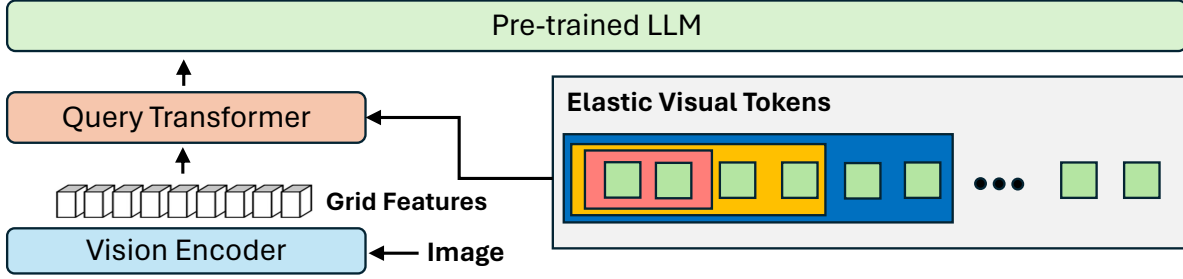


Figure 2.2: Our model employs a query transformer to encode images as visual tokens. We randomly select the first m tokens during training, and enable flexible choice of *any* m number under M during inference, where M is the maximum number of initialized tokens.

$d_1 \leq d_2 \leq \dots \leq d_M$. MRL then trains these models jointly with the following loss:

$$\sum_m c_m \cdot \mathcal{L}(f_m(x); y), \quad (2.1)$$

where \mathcal{L} is the loss function and y is the ground truth label. Note that for each training step, MRL performs forward and backward passes for all M models, inducing significant training overhead compared to training one model. After training, MRL can perform inference with any hidden dimension $d_{i \leq M}$, enabling flexible deployment based on specific needs. MRL is our motivation to train LVLMs that can perform inference with a flexibly selected number of visual tokens.

2.2.2 MQT-LLAVA

We first explain how we encode images with a query transformer, then discuss our training paradigm.

Encoding images with a Query Transformer. We employ a query transformer-based architecture to extract visual tokens from images following previous work [Bai et al., 2023, Li et al., 2022]. Specifically, an input image x is first processed by an image encoder and are then flattened into $H \times W$ grid features $\mathbf{G} = [\mathbf{g}_{11}, \dots, \mathbf{g}_{1W}, \dots, \mathbf{g}_{H1}, \dots, \mathbf{g}_{HW}]$. Then, a query transformer Q is applied to compress the grid features to M visual tokens. Specifically, Q assumes a set of latent *query tokens* $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_M]$ as input, where M is usually smaller than $H \times W$. The query tokens cross-attend to the grid features and compress the information into the query tokens. The final-layer

query tokens become the visual tokens \mathbf{V} that are fed to a large language model together with the input text tokens. I.e., $\mathbf{V} = Q(\mathbf{Z}, \mathbf{G})$. A linear projection layer is added in the end to match the hidden size of the language model.¹

Matryoshka Query Transformer. To enable elastic inference, given the M latent query tokens $\mathbf{Z} = [z_1, \dots, z_M]$, at each training step, we feed only the first m query tokens to the query transformer Q . Subsequently, we obtain only m visual tokens from the query transformer. m can be any number equal to or smaller than the maximal token number M . In practice, we choose m from a linear set of maximum dimensions, in increments of 2, e.g. m can be any number in $\{2, 4, 6, \dots, 252, 254, 256\}$ when $M = 256$. From a training efficiency perspective, our approach uses, on average, half of the visual tokens compared to the original query transformer-based models.

Formally, given an input image with its corresponding text question q and answer y , at each training step, we randomly select a m and feed the first m latent tokens $\mathbf{Z}_{1:m}$ and the text question q to the model. We compare the model output and y and minimize

$$c_m \cdot \mathcal{L}(\text{LM}(\mathbf{V}, q); y), \text{ where } \mathbf{V} = Q(\mathbf{Z}_{1:m}, \mathbf{G}), \quad (2.2)$$

LM is the language model, \mathcal{L} is the language modeling loss function, and c_m is a constant coefficient to control the weight of different numbers of visual tokens, which is always set to 1 in our setting.

Discussion. Here we discuss several interesting properties of MQT. (1) Unlike the original matryoshka representation learning that maintains a nested structure in the parameter space, we specifically target LVLMs and make the visual tokens Matryoshka-like. (2) Despite discarding the tail $M - m$ tokens during each training step, models trained with this token-dropping strategy perform comparably to those trained consistently with all M tokens, as long as we utilize the entire M tokens during inference for both models. (3) Unlike the original MRL, which performs forward and backward passes for all M configurations in each step, we now select just one model

¹Unlike previous work [Bai et al., 2023, Ye et al., 2023a] that first applies projection followed by attention, we empirically find that our “attention then projection” architecture performs better (c.f. §2.4.3).

configuration per training step, significantly cutting training costs. (4) Our cost reduction enables training across a broader spectrum of m values, facilitating the training of models with a more diverse range of choices compared to the original MRL’s limited scope.

2.3 Experiments

We first introduce the implementation details of our query transformer architecture (§2.3.1). We then show the empirical performance of our approach compared to state-of-the-art models across 11 benchmarks (§2.3.2). Finally, we further study the performance-efficiency trade-off (§2.3.3).

2.3.1 Experimental Setup

MQT-LLAVA Implementation Details. We implement our models based on LLaVA-1.5 [Liu et al., 2023a], except that we use our Matryoshka Query Transformer instead of an MLP to obtain the visual tokens. The MQT is a single-layer Transformer with cross-attention. Following Liu et al. [2023a], we select CLIP ViT-L/14 [Radford et al., 2021] as our vision encoder, supporting 336x336 image resolution, and Vicuna-v1.5 [Chiang et al., 2023] as our LLM. As studied in Hu et al. [2024c], Liu et al. [2023b], Zhu et al. [2023], we adopt a two-stage training approach. We train only the query transformer in the first-stage alignment, using LLaVA-558K for 1 epoch with a batch size of 256 and a learning rate of 1e-3. We then fine-tune both the query transformer and LLM using LLaVA-665K for 2 epochs with a batch size of 128 and a learning rate of 2e-5. All training is on 8xA6000s, with 4 and 30 hours per stage, respectively. We apply MQT during the second stage (c.f. §2.4.3).

Baselines. As shown in Table 2.1, we compare our model with LLaVA-1.5 [Liu et al., 2023a] and our model’s baseline LLaVA query transformer (QT-LLaVA), which is trained with a fixed number of 256 visual tokens across all training stages. We also list other models’ results for comparison, including BLIP-2 [Li et al., 2023a], InstructBLIP [Dai et al., 2023], Shikra [Chen et al., 2023b],

Method	LLM	Res.	#Tokens	VizWiz	SQA ^I	VQA ^{v2}	GQA	POPE	MME ^P	MME ^C	MMM	MMB	LLaVA ^W	MM-Vet	Avg
BLIP-2	Vicuna-13B	224	32	19.6	61	41.0	41	85.3	1293.8	–	–	–	38.1	22.4	–
InstructBLIP	Vicuna-7B	224	32	34.5	60.5	–	49.2	–	1084	229	30.6	–	60.9	26.2	–
InstructBLIP	Vicuna-13B	224	32	33.4	63.1	–	49.5	78.9	1212.8	243	33.8	–	58.2	25.6	–
Shikra	Vicuna-13B	224	256	–	–	77.4*	–	–	–	–	–	58.8	–	–	–
IDEFICS-9B	LLaMA-7B	224	64	35.5	–	50.9	38.4	–	–	–	–	48.2	–	–	–
IDEFICS-80B	LLaMA-65B	224	64	36.0	–	60.0	45.2	–	–	–	–	54.5	–	–	–
Qwen-VL	Qwen-7B	448	256	35.2	67.1	78.8*	59.3*	–	–	–	–	38.2	–	–	–
Qwen-VL-Chat	Qwen-7B	448	256	38.9	68.2	78.2*	57.5*	–	1487.5	–	–	60.6	–	–	–
LLaVA-1.5	Vicuna-1.5-7B	336	576	50.0	66.8	78.5*	62.0*	85.9	1510.7	316.1	34.7	64.3	63.4	30.5	59.2
QT-LLaVA	Vicuna-1.5-7B	336	256	51.1	68.1	76.8*	61.5*	84.1	1431.2	348.2	34.3	64.0	63.9	27.9	58.8
MQT-LLaVA	Vicuna-1.5-7B	336	256	53.1	67.6	76.8*	61.6*	84.4	1434.5	353.6	34.8	64.3	64.6	29.8	59.4
MQT-LLaVA	Vicuna-1.5-7B	336	144	<u>52.0</u>	67.5	76.4*	61.4*	83.9	1446.4	351.8	34.4	64.4	61.4	<u>29.9</u>	58.9
MQT-LLaVA	Vicuna-1.5-7B	336	64	51.5	67.0	75.3*	60.0*	83.6	1464.3	<u>352.9</u>	34.4	63.5	59.4	28.9	58.3
MQT-LLaVA	Vicuna-1.5-7B	336	36	51.0	66.8	73.7*	58.8*	81.9	1416.3	349.3	34.4	63.4	59.6	27.8	57.4
MQT-LLaVA	Vicuna-1.5-7B	336	16	49.8	67.5	71.1*	57.6*	80.8	1408.5	349.3	33.6	61.9	55.2	25.3	56.1
MQT-LLaVA	Vicuna-1.5-7B	336	8	49.4	66.2	67.2*	55.5*	79.4	1282.2	323.6	33.1	58.6	51.4	21.3	53.3
MQT-LLaVA	Vicuna-1.5-7B	336	4	49.4	65.1	64.1*	53.0*	77.6	1176.1	296.8	32.8	56.5	44.3	20.2	50.8
MQT-LLaVA	Vicuna-1.5-7B	336	2	48.5	65.0	61.0*	50.8*	74.5	1144.0	268.9	32.5	54.4	41.7	19.5	49.0

Table 2.1: Comparison with state-of-the-art methods on 11 vision-language benchmarks. Our model (MQT-LLaVA) with up to 256 tokens achieves on par or better than LLaVA-1.5 performance across 11 benchmarks, outperforming it on 6 of 11 benchmarks. We mark the best performance in **bold** and the second-best underlined. #Tokens is the number of visual tokens used during inference. Avg is the normalized average across 11 benchmarks, out of 100. Benchmark names are abbreviated for brevity: SQA^I: ScienceQA-IMG, MME^P: MME Perception, MME^C: MME Cognition, MMB: MMBench, LLaVA^W: LLaVA-Bench (In-the-Wild). *The training images of the datasets are observed during training.

IDEFICS [IDEFICS, 2023], and Qwen-VL [Bai et al., 2023].

Evaluation Benchmarks. We evaluate our model across 11 mainstream benchmarks, including VizWiz [Gurari et al., 2018], ScienceQA-IMG [Lu et al., 2022a], VQA-v2 [Goyal et al., 2017], GQA [Hudson and Manning, 2019b], POPE [Li et al., 2023c], MME Perception [Fu et al., 2023], MME Cognition [Fu et al., 2023], MMBench [Liu et al., 2023c], LLaVA-Bench (In-the-Wild) [Liu et al., 2023b], and MM-Vet [Yu et al., 2024].

2.3.2 Main Results

Table 2.1 presents the results of MQT-LLaVA with inference visual token budgets of 2, 4, 8, 16, 36, 64, 144, and 256. We refer to the baseline approach, where the model is trained with a fixed number of visual tokens across all training stages, as LLaVA Query Transformer (QT-LLaVA). MQT-LLaVA outperforms the baseline QT-LLaVA with 256 tokens in 9 out of 11 benchmarks.

One possible explanation is that by enforcing our model to only see fewer tokens during training, the stricter constraint helps the model generalize better to unseen tasks. This is especially evident in the higher performance on VizWiz. When compared to open-source state-of-the-art models, our model with 256 tokens achieves on par or better than LLaVA-1.5 performance with 576 tokens across 11 benchmarks, outperforming it in 6 out of 11 benchmarks. Even with 64 tokens, our model falls short of LLaVA-1.5 by only 0.9 points on average. When drastically drop to only 2 tokens, our score falls by only 3% on ScienceQA and 6% on MMMU. While directly adding a query transformer to LLaVA degrades the performance, our strategy can achieve comparable or better performance than LLaVA-1.5.

We explore performing inference using a variety of numbers of visual tokens, including 1) an extremely low number of tokens; 2) a number of visual tokens unseen during training. As shown in Figure 2.3, MQT-LLAVA with only 2 visual tokens outperforms InstructBLIP (Vicuna-7B), which is based on Q-Former [Li et al., 2023a] using 32 visual tokens. This demonstrates the effectiveness of our model in compressing visual information, pointing to its potential use for applications in computation-heavy tasks. For an unseen number of visual tokens, we pick a random number of visual tokens: 77, and include its results in Appendix A.1. Despite never being explicitly trained for this number of tokens, our model can generalize to any number within 256 during inference, demonstrating a further benefit of our elastic approach.

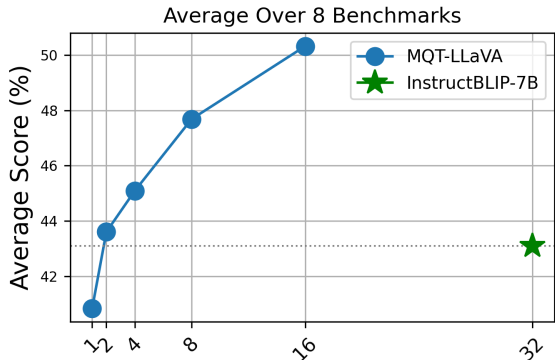


Figure 2.3: With only 2 visual tokens, MQT-LLAVA outperforms InstructBLIP (which uses 32 visual tokens) on all 8 benchmarks it is evaluated on.

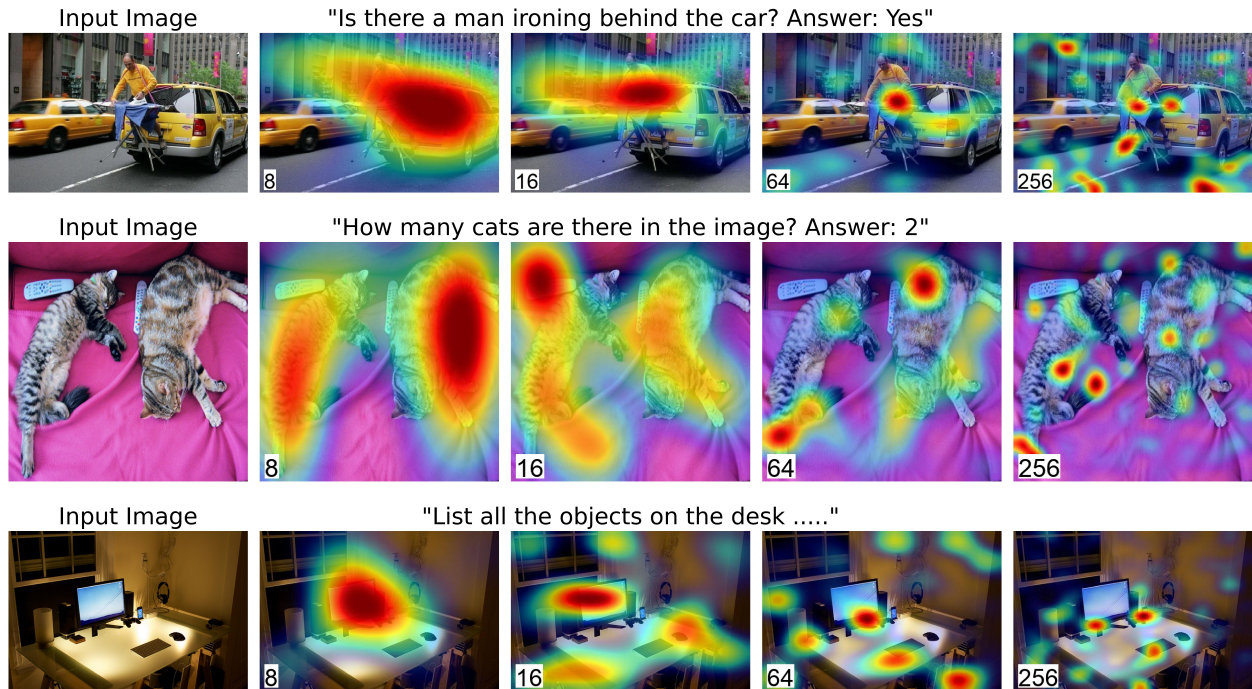


Figure 2.4: Grad-CAM visualization of 1 randomly picked token from using 8, 16, 64, 256 visual tokens, respectively, to encode an image. The model effectively concentrates on high-level concepts using fewer tokens and delves into low-level details with more tokens. The complete input to the third image is “List all the objects on the desk. The objects on the desk include a computer monitor, a keyboard, a mouse, a cell phone, and a pair of headphones”.

2.3.3 Computational Efficiency

To demonstrate our computational efficiency, we compute TFLOPs when running MQT-LLAVA on MMBench with 8, 16, 36, 64, 144, and 256 visual tokens, compared to LLaVA with 576 tokens. As shown in Figure 2.1, we are able to achieve significant speed-ups with little-to-no performance loss: our model with 256 and 144 tokens respectively achieve a 2x and 3x speed-up compared to LLaVA-1.5 while maintaining the same or even better performance; and when using 16 tokens, we achieve an 8x speed-up with a performance drop of only 2.4 points.

2.4 Analysis

To better understand the meaning of visual tokens and to systematically study the number of tokens required by different vision-language tasks, we investigate two key questions: (1) How does the

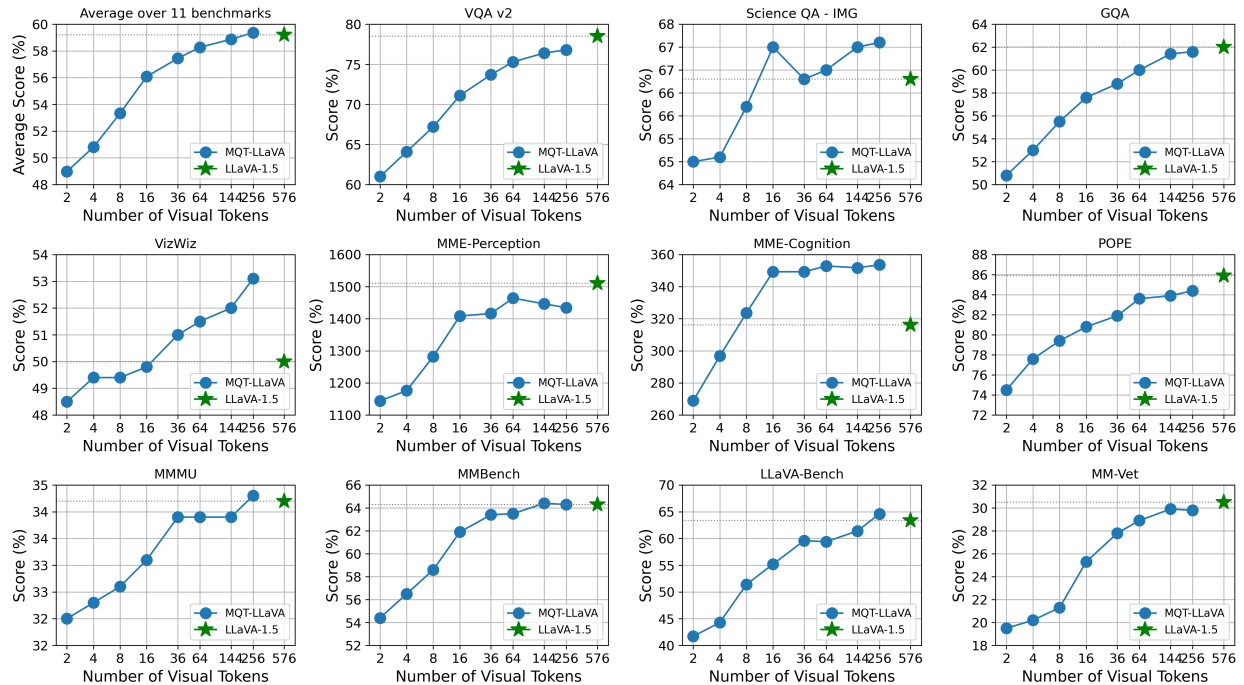


Figure 2.5: The number of visual tokens impact different tasks differently (x-axis is in log-scale). Our model’s performance on ScienceQA, MME-Cognition and MMMU is remarkably robust to token reduction.

focus of the model change with varying numbers of visual tokens? (§2.4.1); and (2) How do different numbers of visual tokens impact various tasks? (§2.4.2)

2.4.1 How does the focus of the model change with varying numbers of visual tokens?

To explore what visual information each token encodes, we utilize Grad-CAM [Selvaraju et al., 2017] to visualize the focus of visual tokens. As illustrated in Figure 2.4, we qualitatively analyze the results of using 8, 16, 64, and 256 tokens.

We observe that the model’s focus changes with the number of tokens used. When using a few tokens (e.g., 8), the model accurately concentrates on global visual concepts related to the question. As the number of tokens increases (e.g., 256), the model not only attends to the relevant objects but also delves into localized details. For example, in the third image, with 8 tokens, the model focuses on the monitor. With 16 tokens, it includes both the monitor and the mouse. With 64 tokens,

it highlights the monitor and keyboard. Finally, with 256 tokens, the model encompasses several objects, including the monitor, keyboard, and cell phone. In the examples from the first and second images, our model effectively focuses on the man ironing behind the car and the two cats, even with only 8 tokens. The impressive qualitative results, especially those using only a few tokens, demonstrate the effectiveness of our approach and the strong capabilities obtained despite using a minimal number of tokens.

2.4.2 How do different numbers of visual tokens impact different tasks?

When using varying numbers of visual tokens during inference, we observe that the model’s performance change varies across different tasks.

Tasks requiring a large number of visual tokens. Tasks that require fine-grained visual understanding and deep reasoning across multiple areas of the image naturally demand a higher number of visual tokens for optimal performance. When the number of visual tokens decreases, the encoded image information is reduced, leading to performance degradation. This trend is evident in tasks such as VQAv2, GQA, VizWiz, MMBench, LLaVA-Bench, and MM-Vet. As illustrated in Figure 2.5, the performance on these tasks gradually declines as the number of visual tokens decreases from 256, with a more rapid decline observed when the tokens are further reduced.

Tasks robust to visual token reduction. In contrast, for several benchmarks primarily targeting the visual perception skills of models, performance remains consistent when gradually reducing the number of visual tokens until a threshold is reached. Beyond this threshold, performance drops significantly (see Figure 2.5). This “turning point” is observed in benchmarks such as MME Cognition, MME Perception, POPE, and MMMU.

For instance, in MME-Cognition (see Figure 2.6), tasks involving commonsense reasoning, code reasoning, and numerical calculation can be performed effectively with as few as 16 visual tokens, allowing the model to focus on the relevant image sections. Similar results are seen in other

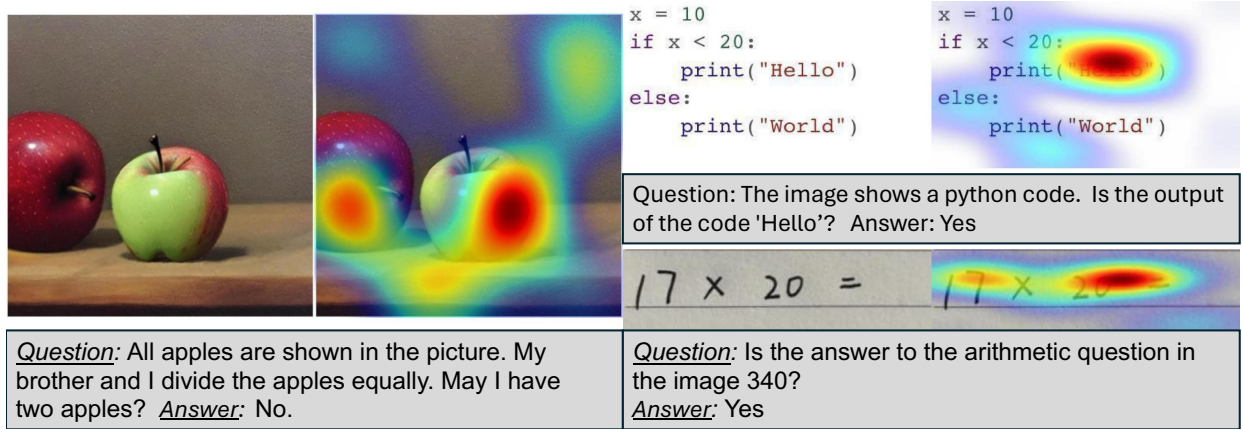


Figure 2.6: Examples from MME Cognition. Grad-CAM results are from using 16 tokens which answered all the questions correctly.

tasks, like the hallucination question "Is there a car in the image?" from POPE. However, once the “turning point” is reached, further reducing the number of visual tokens prevents the model from attending to the correct objects, leading to a sharp decline in performance.

Another notable observation comes from ScienceQA and MMMU, which contain subject-specific questions from school curricula. The model’s performance on these tasks remains robust despite a decrease in visual tokens, achieving scores of 65.0 and 32.5, respectively, with only 2 tokens. This suggests that the reasoning required for academic questions is primarily conducted by the language model (LLM); even with minimal visual hints, the LLM can interpret the image content and perform the reasoning tasks effectively.

When are fewer visual tokens better? As shown above, MQT-LLAVA with 16 tokens can achieve better performance on ScienceQA compared to MQT-LLAVA with 144 tokens. To understand why fewer tokens may benefit this task, we qualitatively analyze instances where MQT-LLAVA succeeded with 16 visual tokens, but failed with 144. We show a representative example in Figure 2.7. MQT-LLAVA with 16 visual tokens attends to all three objects, allowing it to understand their mutual relationship and answer the question correctly. On the other hand, with 144 visual tokens, MQT-LLAVA focuses on various portions of the image and attend to each object independently. This discourages the model from reasoning with the common attributes among

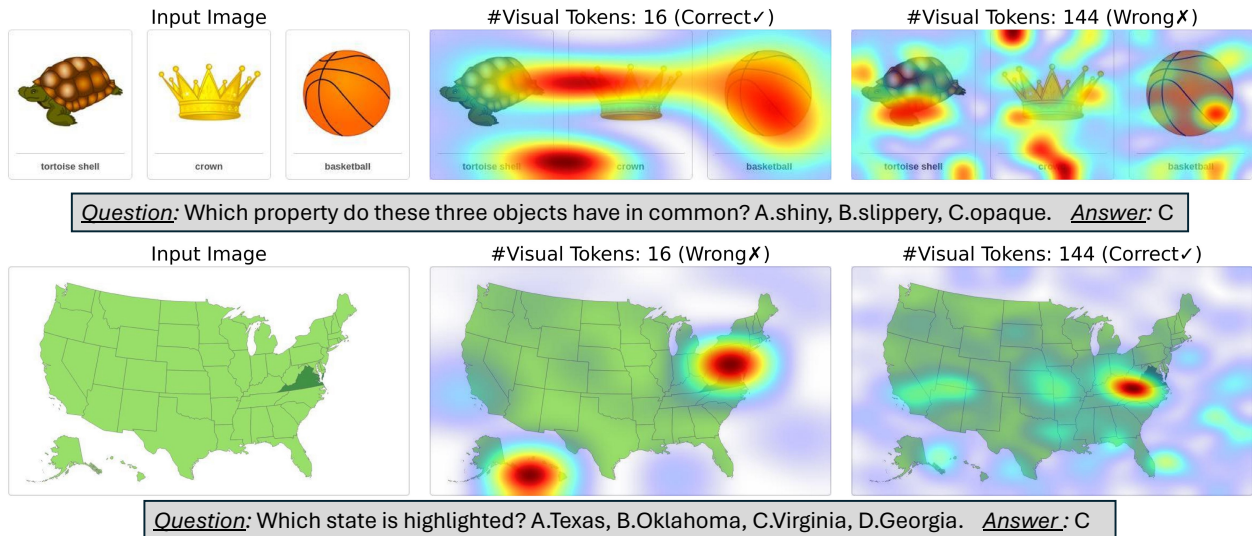


Figure 2.7: Comparison of correct and failure cases in 16 vs 144 visual tokens on Science-QA (test-set).

the three objects, thus predicting the wrong answer. In summary, fewer visual tokens seems to be preferable when fine-grained visual understanding is not required.

However, it should be noted that using fewer tokens is not always better in this case. As shown in Figure 2.7, MQT-LLAVA with 144 tokens precisely identified state of Virginia on the map and answered the question correctly. Whereas 16 tokens concentrated on another region which potentially confused its final prediction, lacking the abilities of distinguishing local details of the geographic shape on the map.

2.4.3 Ablation Studies

We ablate several design choices across 11 benchmarks in Table 2.2. Each ablation independently modifies our best variant, MQT-LLAVA, to create new variants. (i) *linear vs. log-based token number selection*. We replace our linear growth elastic tokens, i.e., $m \in \{2, 4, 6, \dots, 252, 254, 256\}$ to the log-based approach of MRL, i.e., $m \in \{2, 4, 8, 16, \dots, 128, 256\}$. This results in an average accuracy of 57.3%, 2.1% lower than MQT-LLAVA, validating our hypothesis that gradually compressing the visual tokens helps the model perform better than log-based choices. (ii) *query transformer architecture*. As mentioned in §2.2, we choose to first perform cross-attention between

Method	VisWiz	SQA ¹	VQA ^{v2}	GQA	POPE	MME ^P	MME ^C	MMMU	MMB	LLaVA ^W	MM-Vet	Avg
QT-LLaVA (Baseline)	51.1	68.1	76.8*	61.5*	84.1	1431.2	348.2	34.3	64.0	63.9	27.9	58.8
MQT-LLaVA (Ours)	53.1	67.6	76.8*	61.6*	84.4	1434.5	353.6	34.8	64.3	64.6	29.8	59.4
w/ Log-based Matryoshka Tokens	51.2	67.4	75.6*	60.3*	83.2	1418.9	314.1	32.8	62.6	59.2	27.3	57.3
w/ Project then Attention	50.5	66.8	73.4	57.1	82.3	1382.8	317.5	32.7	61.4	60.0	29.5	56.6
w/ First-stage training with Query Transformer	51.6	67.2	75.9*	60.5*	82.6	1378.6	295.4	33.2	63.1	56.5	26.8	56.7

Table 2.2: For simplicity in ablation studies, we evaluate all the models with 256 visual tokens. All models are trained with the same hyperparameters.

query tokens and visual features, then project the learned visual tokens to the LLM. We call this technique “attention then projection”. The alternative variant is “projection then attention”, which achieves lowest average performance, with a score of 56.6%. This suggests that directly applying the attention mechanism helps preserve the rich grid features, making them better projected to the LLM. (iii) *first-stage pretraining with query transformer*. As mentioned in §2.3.1, we choose to apply our elastic training paradigm only during the second stage. Experimental results demonstrate that adopting elastic training during the first stage leads average performance dropped by 2.7%. We hypothesize that the first stage aims to align the randomly initialized query tokens with vision-language awareness. Therefore, it is important to train all 256 tokens with this prior knowledge before reducing the number of tokens in the second stage.

2.5 Conclusion

In this work, we present MQT-LLaVA, a single vision-language model that enables elastic inference on various downstream tasks and computation resources. We demonstrate that our model achieves performance comparable to or better than training with a fixed number tokens. MQT-LLaVA matches the performance of LLaVA-1.5 across 11 benchmarks using less than half the number of visual tokens, and outperforms LLaVA-1.5 in 6 out of 11 benchmarks. We achieve an 8x less TFLOPs when reducing to 16 tokens while only sacrificing the performance on MMBench by 2.4 points. We hope our exploration of the trade-off between the accuracy and computational cost caused by the number of visual tokens will facilitate future research to achieve the best of both worlds.

Chapter 3

Advance Hallucination and Informativeness Evaluation of LVLM

3.1 Introduction

Large Vision-Language Models (LVLMs) Chen et al. [2023a], Liu et al. [2023b], OpenAI [2023] have shown remarkable performance across a broad range of vision-language tasks. Despite the promising progress, the issue of hallucinations has emerged as a critical concern. *Hallucination* refers to the generation of plausible-sounding but inaccurate or fabricated textual descriptions for a given image, which can compromise the reliability and trustworthiness of the models.

Recent studies have proposed various methods to *evaluate* models' *generative* hallucinations Jing et al. [2023], Wang et al. [2023a], Zhai et al. [2023] and *discriminative* hallucinations Guan et al. [2023], Li et al. [2023c], Lovenia et al. [2023]. However, they predominantly focus on hallucinations concerning object existence and their faithfulness within generated content, often neglecting other critical types of hallucinations and the assessment of coverage. This oversight can result in a lack of attention to the variety and depth of hallucinations that may occur beyond

The contents of this chapter appeared in paper Qiu et al. [2024]

Evaluation Method	Hallucination Type			Human Annotation	Faithfulness	Coverage	Open Vocab. Generation
	<i>Object</i>	<i>Attribute</i>	<i>Relation</i>				
POPE	✓	✗	✗	✗	✓	✗	✗
HaELM	✓	?	?	✗	✓	✗	✓
HallusionBench	✓	?	?	✓	✓	✗	✗
Halle-Switch	✓	✗	✗	✗	✓	✓	✓
NOPE	✓	✗	✗	✗	✓	✗	✗
Bingo	?	?	?	?	✓	✗	✗
FaithScore	✓	✓	✓	✗	✓	✗	✓
AMBER	✓	✓	✓	✓	✓	✓	✗
MERLIM	✓	✗	✗	✗	✓	✗	✗
Ours (VALOR-EVAL)	✓	✓	✓	✓	✓	✓	✓

Table 3.1: Comparison of existing hallucination evaluation benchmarks for LVLMs, including POPE Li et al. [2023c], HaELM Wang et al. [2023a], HallusionBench Guan et al. [2023], Halle-Switch Zhai et al. [2023], NOPE Lovenia et al. [2023], Bingo Cui et al. [2023], FaithScore Jing et al. [2023], AMBER Wang et al. [2023b], MERLIM Villa et al. [2023]. ? refers to features not explicitly mentioned in the paper. Open Vocab represents evaluating free-form generated captions without constraints to pre-defined vocabulary.

object identification, such as attributes and relations. Furthermore, these evaluation methods are often constrained by a predefined vocabulary, thus are inherently limited to fully appreciating the richness of the free-form generated captions. Specifically, the evaluation metrics may not capture novel expressions that extend beyond the predetermined vocabulary.

In contrast to prior studies, we introduce a human-annotated multi-dimensional evaluation benchmark VALOR-BENCH¹ by breaking down hallucinations into three categories: *object* (existence), *attributes* (color and count), and *relations* (positional and comparative). In addition, to make the test cases challenging, we utilize the *associative biases* Li et al. [2023c], Zhou et al. [2023] presented in training datasets to select images with only one component of commonly co-occurred pairs or groups, leading models to mistakenly generate associated elements that are not present. Our experimental findings validate the effectiveness of this methodology in exposing the susceptibility of current LVLMs to such biases.

In addition to constructing the benchmark dataset, we also propose a new evaluation framework, VALOR-EVAL. Existing evaluation frameworks such as the widely used CHAIR Rohrbach et al. [2018] metric, exhibit several major constraints. First, they rely on a predefined vocabulary, limiting their ability to identify hallucinations in an *open vocabulary* setting where semantic nuances –

¹VALOR is short for vision-language atttribute, relation, and object coverage and faithfulness.

such as synonyms and variations – are prevalent in model outputs and references. Additionally, focusing exclusively on hallucination overlooking the aspect of *coverage*, resulting in a preference for precise but uninformative model outputs. To address these issues, our propose VALOR-EVAL metric generalizes CHAIR by incorporating an LLM in a two-stage design, enhancing the capability to evaluate open vocabulary hallucination across object, attribute, and relation dimensions while also considering coverage. We provide a detailed comparison of existing evaluation methods in Table 3.1.

We conduct comprehensive evaluations on 10 established LVLMs across multiple dimensions with VALOR-BENCH. Our findings reveal that some LVLMs tend to prioritize precision over coverage, leading to predictions with high accuracy but limited scope. This observation underscores the need for the community to focus on achieving an *balance* between faithfulness and coverage in LVLMs.

3.2 VALOR-BENCH

In this section, we detail the methodology employed to create the benchmark, which aims to evaluate the hallucination issues of LVLMs. Constructing this benchmark involves two principle phases: the *collection* of images (Section 3.2.1) and their subsequent *annotation* (Section 3.2.2).

3.2.1 Image Collection

We aim to select images that can effectively expose the issue of model hallucinations. We hypothesize that when models are repeatedly exposed to specific combinations of features – such as object existence, object attributes, and object relations – during training, they develop a pronounced *associative bias*, which leads the models to expect these co-occurring features in similar situations. Consequently, when a model encounters an image containing only one element of a familiar combination, it may erroneously infer the presence of the associated feature. This associative bias is one primary source of model hallucinations Li et al. [2023c], Zhou et al. [2023]. To explore this

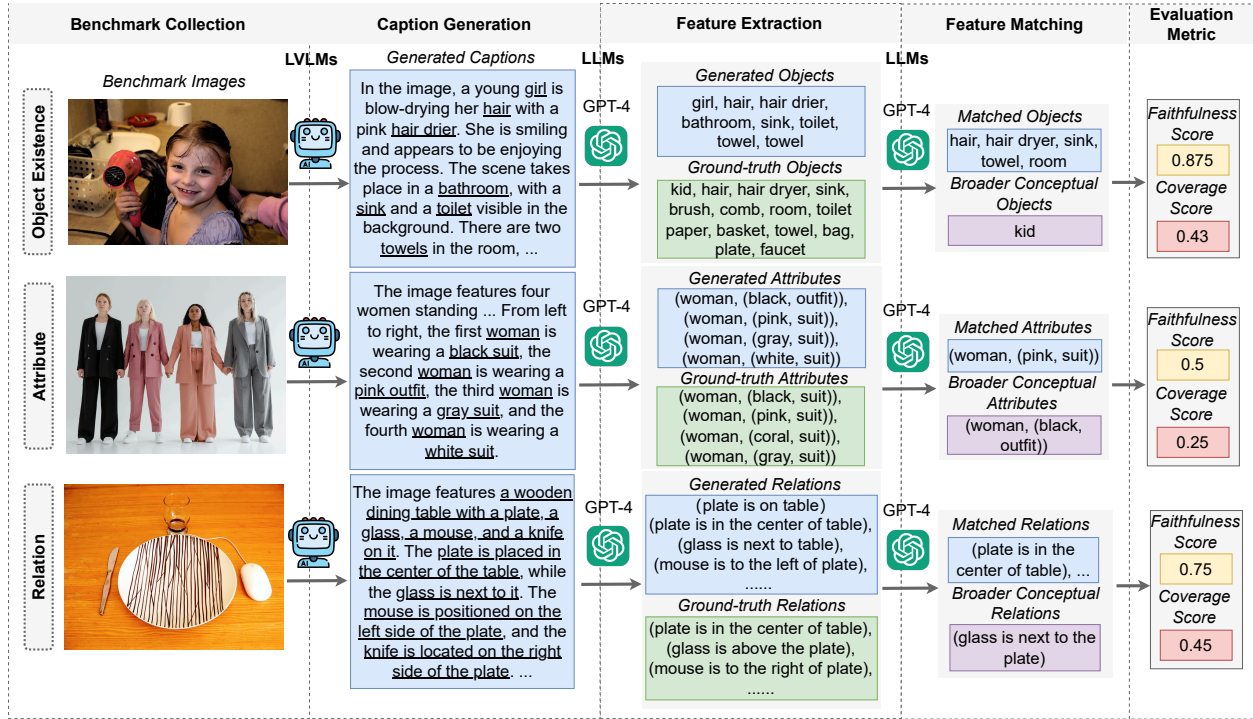


Figure 3.1: Overview of VALOR-EVAL evaluation framework: (1) Firstly, LVLMs generate captions from VALOR-BENCH benchmark images. (2) Following this, LLMs are employed to *extract* pivotal features that encapsulate from the generated descriptions. (3) Subsequently, these features are *aligned* with a pre-defined list of ground-truth features using LLMs, facilitating the creation of two essential outputs: a dictionary of matched features and a more extensive dictionary encompassing broader conceptual matches. (4) Finally, we calculate two key metrics: *faithfulness* and *coverage*. These metrics measure the LVLMs’ comprehension by evaluating how well the generated captions encapsulate the salient features of the images and the breadth of concepts they cover, respectively.

phenomenon, we initially analyze the co-occurrence statistics of *object-object*, *object-attribute*, and *object-relation-object* combinations within the extensively annotated GQA Hudson and Manning [2019a] dataset. We then curate a collection of images representing *frequently* and *infrequently* co-occur (object, object), (object, attribute), (object, relation, object) tuples. By doing so, we identify the most challenging images to construct a benchmark, to which we then add detailed human annotations for later thorough evaluation.

We first outline the definition (Section 3.2.1), then explain the process for calculating co-occurrence statistics (Section 3.2.1), and finally describe the steps for using these dependencies to select images (Section 3.2.1).

Definition We first define three principal features to assess hallucination issues in LVLMs. The first feature, **Object existence** (object-object), encompasses all visual entities within an image, covering both *foreground* and *background* elements. The second feature, **Attribute** (object-attribute), focuses on the characteristics of objects, with a particular emphasis on *color* and *counting*. Our analysis within this category is divided into two segments: *object* and *people*. For objects, we concentrate on the color and count of each item not related to people (*e.g.*, six *green apples* on the table). For people, we highlight the colors of attire and the total number of individuals depicted (*e.g.*, a woman who is wearing a *red jacket*). The third feature, **Relation** (object-relation-object), pertains to the relational information between the objects in the image. Here, we focus on *positional* and *comparative* relation. Specifically, the positional relation tests the relative position between the objects, while the comparative relation analyzes the understanding of “which object is larger than the other.”

Quantifying Co-Occurring Features To utilize co-occurring features effectively, the first step involves computing the *statistical dependencies* between different features. This analysis aids in identifying dominant co-occurrence patterns in the data, thereby spotlighting features with strong associations that the model might have internalized. We employ two statistical methods to determine these dependencies – *frequencies* and *conditional probabilities*. **Frequency** provides insights by quantifying the frequency of specific features in conjunction with particular objects, attributes, or relations, thereby illuminating the raw distribution of these features throughout the dataset. To delve deeper, we calculate the **conditional probability**, which quantifies the likelihood of encountering a specific feature given the presence of an object:

$$\mathcal{P}(\text{feature}|\text{object}) = \frac{\text{Frequency}(\text{feature}, \text{object})}{\text{Frequency}(\text{object})}, \quad (3.1)$$

where $\text{feature} \in \{\text{object}, \text{attribute}, \text{relation}\}$. Our goal is to identify objects whose conditional probability distributions exhibit significant skew. To achieve this, we explore five distinct metrics based on conditional probabilities. Detailed definitions of these five metrics are provided in

Appendix B.2.

Utilizing Co-Occurrence Statistics for Image Extraction Leveraging the identified co-occurrence statistics, we systematically extract images from existing datasets. The process includes several critical steps:

1. Identify objects (\mathbf{O}) that exhibit the *most pronounced* co-occurrence dependencies, including frequency and conditional probabilities:

$$\mathbf{O} = \{\arg \max_o P(f|o) | f \in F\}, \quad (3.2)$$

where \mathcal{F} denotes the set of all features (including object, attribute, and relation) annotated in the dataset, o represents any object annotated in the dataset, and \mathcal{P} signifies all statistical dependencies, including frequencies and five kinds of conditional probabilities.

2. Select features that are *minimally* associated with each identified object in \mathbf{O} , denoted as set \mathbf{I} , thereby spotlighting instances where common co-occurrences are *absent*:

$$\mathbf{I} = \{\arg \min_i P(i|o) | i \in F_o, o \in \mathbf{O}\}, \quad (3.3)$$

where \mathcal{F}_o denotes the set of all features (including object, attribute, and relation) annotated in the dataset related to object o and \mathcal{P} signifies all statistical dependencies.

3. Determine features that are *most frequently* co-occurring with each identified object in \mathbf{O} , denoted as set \mathbf{H} , serving as *strong* associative tendencies:

$$\mathbf{H} = \{\arg \max_h P(h|o) | h \in F_o, o \in \mathbf{O}\}, \quad (3.4)$$

where \mathcal{F}_o denotes the set of all features (including object, attribute, and relation) annotated in the dataset related to object o and \mathcal{P} signifies all statistical dependencies.

4. Collect images \mathbf{C} for each feature in \mathbf{I} corresponding to an object in \mathbf{O} , with the chosen images *including* the specified feature and object, yet *excluding* any features from \mathbf{H} , to create clear cases for testing the model’s associative bias:

$$\mathbf{C} = \{c : (o, f) | o \in \mathbf{O}, f \in \mathbf{I}, \text{ and } f \notin \mathbf{H}\} \quad (3.5)$$

where c denotes an image that contains the object o characterized by the feature f .

For each feature defined in Section 3.2.1, we adhere to the outlined steps to extract images from the GQA dataset. Subsequently, we manually review the collected images by two expert annotators to ensure that only those of high quality and with clear annotations are retained. These procedures enable us to amass a collection of images for evaluating the object existence and the relations. However, extracting images that accurately represent specific *attributes* proved to be challenging due to the limited attribute annotations in GQA. To overcome this, we source copyright-free images from the Internet², guided by the attribute-related statistics gathered in the previous step. The statistics of our proposed benchmark are detailed in Table 3.2.

3.2.2 Annotation

For each image within the distinct feature subsets, we manually annotate them based on existing annotations, adhering to the definitions discussed in Section 3.2.1. Figure 3.1 illustrates three examples in the object, attribute, and relation subsets from our collected benchmark. Below, we discuss the details of these annotations.

Object Existence. Through manual verification of existing annotations, we enhance the dataset by including additional annotations to ensure all visual entities within an image are accounted for. This contains both *foreground* and *background* entities. For example, in an image showing “a lady sitting on a bench in front of a building,” the objects to be annotated are the “lady,” “bench,” and “building.”

Attributes. In a similar vein to the approach adopted in the object subset, we further enhance images by appending detailed attribute annotations to the depicted objects. Our analysis within this category bifurcates into two subsets: *object* and *people*. Within the object sub-category, for an image described as “two green apples on a white table,” the identified attributes are “(green, apple)”

²We use Pixel, a free stock photos platform: <https://www.pexels.com/> for image retrieval.

Category	Sub-Category	# Images	Source
Object Existence	-	50	GQA
Attribute	Object	27	Pixel
	People	34	Pixel
Relation	Positional	50	GQA
	Comparative	50	GQA

Table 3.2: In the VALOR-BENCH benchmark, we categorize images into three main areas: object existence, attributes, and relations, as outlined in Section 3.2.1 and Section 3.2.1. Attributes are further split into *object* (focusing on color and count of each item not related to people) and *people* (emphasizing the attire colors and the total number of individuals. For relations, we examine both *positional* relations between objects and *comparative* sizes.

for each apple and “(white, table)” for the table. For *people* sub-category, in a scene showing “a woman wearing a red jacket with black shoes,” the identified attribute is “(woman, (red, jacket), (black, shoes))”.

Relations. In our benchmark, we capture *positional* relations between objects. For instance, the statement “the bed is to the left of the table” illustrates the positional relation between “bed” and “table”. Conversely, the inverse statement “the table is to the right of the bed” is equally valid and is annotated accordingly. Additionally, we annotate descriptions such as “a bed is on the left side of the image” to denote the positional relations of objects at the image level. For *comparative* relations, we use an annotation scheme that assigns a numerical rank based on object size, ordering objects from largest to smallest (*e.g.*, “1. bed, 2. table, 3. cup”).

Ultimately, VALOR-BENCH provides a set of tuples (I, F_G, p_G) , where I denotes the image, F_G is the feature annotations of the image, and p_G represents the prompt designed for LVLMS generation. The designed prompts p_G are shown in Section B.3 for each subset – object, attribute, and relation.

3.3 VALOR-EVAL

We propose a framework VALOR-EVAL that generalizes CHAIR, a metric that is widely adopted in existing studies Wang et al. [2023b], Zhai et al. [2023], by introducing semantic matching and

incorporating both the *faithfulness* and *coverage* aspects into the evaluation. As shown in Figure 3.1, our evaluation process has two steps: *feature extraction* and *matching* (Section 3.3.1) and *scoring* (Section 3.3.2).

3.3.1 Feature Extraction and Matching

We start the process by generating an initial response, denoted as R , using a specific LVLM with the input pair (I, p_G) , where I denotes the image and p_G represents the prompt designed for LVLMs generation from VALOR-BENCH. Then, we leverage an LLM to analyze R and extract key features. This is achieved through a series of prompts p_E , outlined in Section B.4, which are designed to *extract* features from object existence, attributes, and relations, respectively, resulting in a comprehensive list of extracted features from R , denoted as $F_R = \{f_{R_1}, f_{R_2}, \dots, f_{R_m}\}$. Next, we utilize an LLM to *align* the extracted features list F_R with a pre-annotated ground-truth features list $F_G = \{f_{G_1}, f_{G_2}, \dots, f_{G_m}\}$ from VALOR-BENCH. This alignment is facilitated through a set of carefully crafted prompts p_M , outlined in Section B.5, tailored to each feature subset, aiming to identify correlations and correspondences. Unlike previous evaluation metrics that rely on a fixed feature list and direct mapping, our approach eschews pre-processing and instead utilizes LLMs’ language comprehension capabilities to semantically match extracted features with their ground-truth counterparts. This process yields two key outputs: **matched features** dictionary (D_M) and **broader conceptual matches** dictionary (D_B).

D_M contains features $f_{R_{i_m}}$ from f_R that semantically aligned with the features $f_{G_{i_m}}$ from F_G , ensuring *precision*. For example, if we have the extracted “(plaid, shirts)” and the candidate ground-truth feature is “(checkered, shirt),” we can establish a match between these two because “plaid” and “checkered” are conceptually similar patterns often used interchangeably in the context of textiles.

D_B includes features $f_{R_{j_n}}$ from f_R that have broader conceptual meanings than the features $f_{G_{j_n}}$ from F_G , adding *conceptual depth* to the evaluation. For instance, if we have the extracted “(red, clothes)” from an image, and the ground-truth annotation is “(red, dress),” we can still consider

these features to match. This is because “clothes” is a broader category that encompasses “dress.” Therefore, despite the slight difference in specificity, the extracted features can be aligned with the ground-truth annotations based on their semantic relationship, where “dress” is a sub-type of “clothes.”

3.3.2 Evaluation Metrics

We introduce two metrics to evaluate the hallucinations in two dimensions: *faithfulness* and *coverage* based on the original CHAIR metric.

Faithfulness. In the context of image captioning, faithfulness measures how closely captions match an image’s content, emphasizing *accuracy* in depicting visual elements and their attributes and relations without introducing hallucinations. It is calculated by comparing generated features against actual image features, considering both *direct* (D_M) and *broader* conceptual similarities (D_B):

$$\text{Faithfulness}(R, F_G) = \frac{|D_M \cup \text{set}(D_B)|}{|F_R|} \in [0, 1]. \quad (3.6)$$

Coverage. It measures the *comprehensiveness* of the generated captions in capturing the key elements and attributes depicted in the image. It evaluates the proportion of ground-truth features that are successfully captured in the generated response, only through *direct* matches (D_M):

$$\text{Coverage}(R, F_G) = \frac{|\text{set}(D_M)|}{|F_G|} \in [0, 1]. \quad (3.7)$$

3.4 Experiment

In this section, we perform experiments to evaluate different existing LVLMs within our proposed framework (Section 3.4.1). We also present evidence demonstrating that our evaluation methodology aligns closely with human judgment (Section 3.4.2). Additionally, we explore the significance of

Model	Object		Attribute				Relation				Average	
	Existence		Color & Counting				Positional		Comparative		Faithful.	Cover.
			Object		People						Score	Score
	Faithful _↑	Cover _↑	Faithful _↑	Cover _↑	Faithful _↑	Cover _↑	Faithful _↑	Cover _↑	Faithful _↑	Cover _↑	(%)	(%)
InstructBLIP	74.5	24.8	72.0	23.9	47.1	9.3	50.0	13.6	66.9	35.6	62.1	21.44
LLaVA-1.5	72.1	24.7	74.6	37.8	43.3	12.1	64.8	14.9	51.9	40.1	61.34	25.92
MiniGPT-4 v2	65.0	25.4	64.5	17.9	38.9	11.6	38.8	33.1	44.7	11.2	50.38	19.84
mPLUG-Owl2	71.5	24.8	79.9	32.7	39.7	16.2	45.2	10.8	41.6	30.6	55.58	23.02
BLIVA	77.7	21.9	73.3	24.3	37.6	11.6	39.5	9.7	68.0	29.9	59.22	19.48
CogVLM	71.2	35.5	75.3	24.3	43.7	22.4	51.9	10.5	49.0	35.9	58.22	25.72
InternLM-XComposer2	82.5	23.9	75.8	26.3	50.4	13.8	62.6	11.1	64.1	38.4	67.08	22.7
Qwen-VL-Chat	70.6	28.4	75.1	38.6	38.8	16.0	56.9	8.5	51.9	24.3	58.66	23.16
Emu2	94.2	14.1	66.7	10.4	54.3	1.9	72.2	1.8	87.5	12.3	74.98	8.1
GPT-4V	61.6	38.8	78.5	36.3	34.7	23.8	46.7	12.6	51.6*	28.5*	54.62	28.0

Table 3.3: The overall evaluation results of object existence, attribute, and relation hallucination in VALOR-BENCH using GPT-4 as the LLM Agent within VALOR-EVAL. The highest is highlighted in blue, while the worst performance is highlighted in yellow. Faithfulness and coverage scores are in percentage (%). For images that contain people, GPT-4V refrains from generating comments, and we marked this score with an asterisk (*).

each design aspect of our framework through ablation studies (Section 3.4.3). Finally, we showcase qualitative examples to illustrate our findings (Section 3.4.4).

3.4.1 Model Coverage-Faithfulness Evaluation

We use the framework VALOR-EVAL to evaluate various LVLMs listed in Table B.1 in the Appendix B.1, employing GPT-4 as the evaluation LLM agent.

In the evaluation of various models, as shown in Table 3.3, Emu2 distinguishes itself by achieving the highest average faithfulness score of 74.98, signifying its consistent capability to generate responses that accurately reflect the content of the input image. However, Emu2’s performance in terms of coverage is less impressive, with the lowest average score of 8.1, suggesting that its responses, while accurate, may not comprehensively cover all elements of the image. When broken down into specific dimensions, Emu2 excels in faithfulness across categories – scoring 94.2 in object existence, 54.3 in attribute-people, 72.2 in relation-positional, and 87.5 in relation-comparative. Conversely, it lags in coverage, with scores of 14.1 in object existence, 10.4 in attribute-object, 1.9 in attribute-people, and 1.8 in relation-positional. These results point to a potential trade-off

between faithfulness and coverage in Emu2’s design, where the model prioritizes accuracy at the expense of a broader scope in its responses. This pattern supports the initial hypothesis that *some LVLMs may intentionally sacrifice coverage to improve the precision of their outputs*.

Meanwhile, GPT-4V(ision) distinguishes itself with an unparalleled average coverage score of 28.0, showcasing its adeptness in encapsulating a wide array of features from the input image. This indicates that GPT-4V excels in recognizing and addressing diverse elements within images, although it does not necessarily always maintain the highest accuracy, as seen in its lower faithfulness score of 61.6. Particularly in evaluations concerning the existence of objects, GPT-4V leads with the highest coverage score of 38.8, underlining its comprehensive approach to object detection. This approach tends to favor inclusivity, which might lead to the occasional identification of objects that are not present in the image. Furthermore, in evaluations focused on attributes related to people, GPT-4V again achieves the highest coverage score of 54.3. However, this comes with a trade-off, as it also exhibits a higher tendency towards hallucinations compared to other models, indicating a propensity to generate details or elements that may not be grounded in the actual content of the image.

Models such as LLaVA-1.5 and CogVLM showcase a more equitable performance, achieving respectable scores in both faithfulness and coverage metrics. This highlights their capability to provide responses that are not only precise but also encompassing. Notably, LLaVA-1.5 stands out for its remarkable outcomes, achieved through the efficient use of training data, underscoring the significance of leveraging high-quality instruction-tuning data to enhance model performance.

3.4.2 Effectiveness of Evaluation Framework

To demonstrate the *effectiveness* and *reliability* of our LLM-based automatic evaluation pipeline, we conduct experiments to evaluate if our evaluation framework correlates with human evaluations in both faithfulness and coverage dimensions. Specifically, we have human and our GPT-4-based

Category	Sub-Category	Faithful. (ρ)	Cover (ρ)
Object Existence	-	0.91	0.89
Attribute	Object	0.99	0.98
	People	0.98	0.96
Relation	Positional	0.78	0.86
	Comparative	0.92	0.98

Table 3.4: Pearson correlation (ρ) between our GPT-4-based evaluation framework VALOR-EVAL and human judgements.

evaluation method evaluate InstructBLIP outputs and compute the Pearson correlation (ρ) score³. As shown in Table 3.4, for object existence, the findings reveal a significantly strong Pearson correlation of 0.91 for faithfulness and 0.89 for coverage, effectively rejecting the null hypothesis that posits no correlation between the two evaluation methodologies, with a compelling p -value of 0. Additionally, our study achieved a notably high correlation of 0.98 in attribute recognition and comparative relations. When evaluating positional relations, which tend to involve longer and more complex descriptions, the correlation scores were not as high as those observed in the other categories but still indicated a very high level of correlation, with 0.78 in faithfulness and 0.86 in coverage. These results affirm the comparability of our automatic evaluation metrics to human evaluation in terms of both *efficacy* and *reliability*.

3.4.3 Ablation Study

In this section, we serve to answer **two** questions and discuss our findings.

1. How does our co-occurrence data selection method compare to other alternatives?

To illustrate the effectiveness of the co-occurrence data selection method, we set up a baseline of randomly selecting 50 images in the GQA validation split and applying human annotations, the same as for our dataset. For the ablation study, we focus on the well-studied object hallucination. We evaluate three popular models representing query tokens-based image features (InstructBLIP), linear projection-based features (LLaVA-1.5), and advanced commercial LVLMS (GPT-4V). As shown in

³We opt for Pearson correlation as our assessment metric due to its suitability for measuring *linear* relationships, as opposed to Spearman’s rank correlation, which is more attuned to *monotonic* relationships.

Model	InstructBLIP	LLaVA-1.5	GPT-4V
<i>Evaluation data: randomly selected</i>			
Faithfulness	76.5	84.5	64.1
Coverage	24.3	26.3	41.2
<i>Evaluation data: co-occurrence selected (Ours)</i>			
Faithfulness	74.5 (-2.0)	72.1 (-12.4)	61.6 (-2.5)
Coverage	24.8 (+0.5)	24.7 (-1.6)	38.8 (-2.4)

Table 3.5: Model performance comparison on our data selection method against random selection. Faithfulness and coverage scores are in percentage (%).

Table 3.5, all models tend to produce more hallucinations and exhibit significantly *lower faithfulness* compared to our benchmark. Notably, LLaVA-1.5 scores 12.4 points lower in faithfulness when evaluated against our benchmark. This suggests that our benchmark is challenging due to its reliance on co-occurrence selection. Additionally, the coverage scores for both LLaVA-1.5 and GPT-4V decreased. Upon further analysis through human review, we discover that our benchmark, on average, contains 1.69 more objects than images selected at random. This finding indicates that our data selection method can incorporate more complex objects compared to the random selection approach commonly used in other benchmark constructions.

2. How does our LLM-based evaluation framework compare with LLM-free evaluation?

We compare our proposed LLM agent augmented framework against the original CHAIR metric which is adopted by all previous studies. Because the CHAIR metric is limited to evaluating only 80 objects from the MSCOCO dataset, for a fair comparison, we randomly select 20 COCO images and re-annotate them for analysis alongside the CHAIR metric. We have made these annotations publicly available, adhering to the same list of synonyms used in the original CHAIR metric. To conduct this comparison, we utilize two accuracy scores. For Acc (F), we assess the performance by comparing the number of hallucinated objects identified by the metric against the ground-truth hallucinated objects in the caption. If an object is incorrectly identified as hallucinated when it is not, the metric imposes a penalty of -1. This score aligns with the *matching* phase of our framework, ensuring a thorough evaluation of hallucination detection accuracy. For Acc (C), we calculate the number of objects detected by metric over the unique objects mentioned in the caption, assessing

Metric	F.\uparrow	C.\uparrow	Acc (F)\uparrow	Acc (C)\uparrow
<i>Model: InstructBLIP</i>				
CHAIR	75.0	34.3	11.11	80.66
CHAIR _{LLM} (Ours)	76.9	30.4	88.89 (+77.78)	100.0 (+19.34)
<i>Model: LLaVA-1.5</i>				
CHAIR	74.3	34.1	30.00	83.52
CHAIR _{LLM} (Ours)	81.5	27.0	90.00 (+60.00)	97.08 (+13.56)
<i>Model: GPT-4V</i>				
CHAIR	79.3	54.8	5.88	82.35
CHAIR _{LLM} (Ours)	69.7	57.9	82.35 (+76.47)	98.17 (+15.82)

Table 3.6: Comparison of LLM-augmented CHAIR with original CHAIR metric. Here, F. and C. denote faithfulness and coverage scores in percentage (%). Acc (F) represents the average percentage of hallucinated objects detected by the metric. Acc (C) denotes the average percentage of objects detected by metric.

our *extraction* phase’s efficiency. As shown in Table 3.6, our framework significantly outperforms in both faithfulness and coverage accuracy by a large amount. This improvement is due to our framework’s *open vocabulary matching* ability, unlike the original CHAIR approach that struggles with new expressions without pre-defined synonyms. Notably, with complex models like GPT-4V, CHAIR’s faithfulness accuracy drops to 5.88, highlighting our method’s strength in managing diverse object descriptions.

Moreover, the limitation of CHAIR’s pre-defined object list extends to its inability to account for potential hallucinated objects, which are essential for differentiating between mere words and actual objects in captions. This leads to its failure in detecting hallucinated objects, resulting in performance degradation. In contrast, our method overcomes this by using an automatically extracted object list that dynamically matches objects, avoiding this limitation. Although approaches like Wang et al. [2023b] attempt to address this by including a selection of potential hallucinated objects, they cannot guarantee coverage of all possible hallucinated objects, particularly in complex outputs from advanced LVLMs that generate extensive captions.

3.4.4 Qualitative Results

We illustrate the qualitative results of three representative models in Figure B.1, Figure B.2 and Figure B.3 in the Appendix B.6. Each model exhibited instances of hallucination in these examples from our evaluation benchmark VALOR-BENCH. Notably, while GPT-4V generates the most comprehensive results, it is also more prone to producing hallucinations.

3.5 Conclusion

We introduce a comprehensive multi-dimensional benchmark, named VALOR-BENCH, dedicated to the evaluation of LVLMs, with a particular focus on measuring hallucinations in generative tasks. Our benchmark categorizes hallucinations into three distinct types – object, attribute, and relation – offering a detailed understanding of model inaccuracies. Furthermore, our novel evaluation framework, referred to as VALOR-EVAL, employs a two-stage approach that integrates an LLM, effectively addressing the complexities related to open vocabularies, semantic similarities, and the intricate assessment of attributes and relationships. This method significantly enhances the precision and depth of image captioning evaluations compared to previous methods. Our experimental findings highlight the persistent challenges in this field, demonstrating that even state-of-the-art models such as GPT-4V, are prone to a considerable degree of hallucination. This study emphasizes the imperative for continuous advancements in LVLM evaluation techniques and establishes a new benchmark for future endeavors aimed at reducing hallucination and bolstering the reliability of content generated by LVLMs.

Chapter 4

Advance LVLM with Visual Retrieval-Augmented Knowledge

4.1 Introduction

Retrieval-augmented generation (RAG) has emerged as a promising direction in large vision-language models (LVLMs) [Bai et al., 2023, Chen et al., 2023b, 2024, Hu et al., 2024c, Huang et al., 2023b, Liu et al., 2023a, McKinzie et al., 2024, OpenAI, 2023, Tong et al., 2024]. By incorporating external knowledge during generation, models such as Wiki-LLaVA [Caffagni et al., 2024] have demonstrated improved performance in knowledge-intensive question answering tasks.

There are several existing benchmarks evaluating retrieval-augmented LVLMs. For example, OK-VQA [Marino et al., 2019] focused on scenarios where the image content alone is insufficient to answer the questions. A-OKVQA [Schwenk et al., 2022] further extended this dataset to incorporate additional types of world knowledge. More recent works [Chang et al., 2022, Chen et al., 2023c, Mensink et al., 2023] further expanded and curated large-scale knowledge base data to evaluate pre-trained vision and language models in knowledge-intensive and information-seeking

The contents of this chapter appeared in paper Hu et al. [2024b]

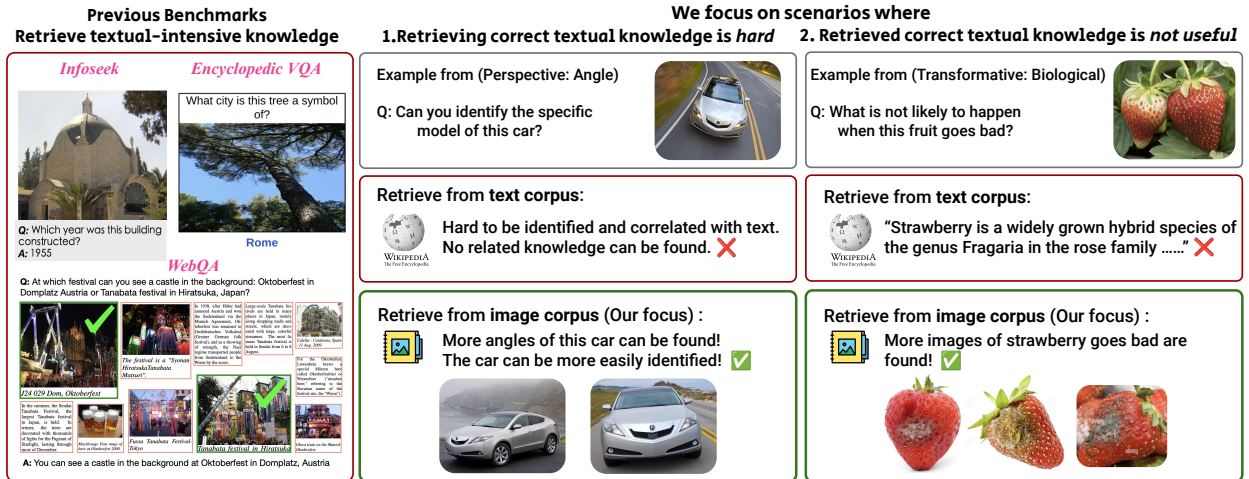


Figure 4.1: Example scenarios from MRAG-BENCH. Previous benchmarks [Chang et al., 2022, Chen et al., 2023c, Mensink et al., 2023] mainly focused on retrieving from textual knowledge. However, there are scenarios where retrieving correct textual knowledge is hard and sometimes not as useful as visual knowledge.

visual questions. However, as shown in Table 4.1, these benchmarks remain text-centric, as their questions can often be resolved with related external textual knowledge. In contrast, retrieving visual information is sometimes more beneficial than retrieving text, as humans often gain greater insights from it. Specifically, we illustrate examples in Figure 4.1 where retrieving correct textual knowledge can be *hard* and retrieved textual knowledge can be *useless*, while retrieving additional images is helpful. For instance, when presented with a top-down view of a car, humans may struggle to accurately identify it; however, with a front-facing view, they can quickly recognize the vehicle and effectively leverage the visual information.

In this paper, we introduce MRAG-BENCH, a benchmark specifically designed for vision-centric evaluation for retrieval-augmented multimodal models, with visual questions typically benefit more from retrieving visual knowledge than textual information. MRAG-BENCH consists of 16,130 images and 1,353 human-annotated multi-choice questions spanning 9 distinctive scenarios. Focusing on utilizing visually augmented knowledge in real-world scenarios, we divide our benchmark into two aspects: *perspective*, where changes in visual entity’s perspective requiring visually augmented knowledge; and *transformative*, where the visual entity undergoes transformative change physically thus requiring visually augmented knowledge. Specifically, MRAG-BENCH









Benchmarks	Knowledge Modality	Knowledge Source	Multi-Image Input	Diverse Scenarios
K-VQA [Shah et al., 2019]	Text	Wikipedia	✗	✗
OK-VQA [Marino et al., 2019]	Text	Wikipedia	✗	✗
MultiModalQA [Talmor et al., 2021]	Text	Wikipedia	✗	✗
ManyModalQA [Hannan et al., 2020]	Text	Wikipedia	✗	✓
A-OKVQA [Schwenk et al., 2022]	Text	Common/World	✗	✗
ViQuAE [Lerner et al., 2022]	Text	Wikipedia	✗	✗
WebQA [Chang et al., 2022]	Text/Caption	Wikipedia	✗	✗
Encyclopedia VQA [Mensink et al., 2023]	Text	Wikipedia	✗	✗
InfoSeek [Chen et al., 2023c]	Text	Wikipedia	✗	✗
MRAG-BENCH (Ours)	Image	   	✓	✓

Table 4.1: Compared with previous works, MRAG-BENCH focuses on evaluating LVLMS in utilizing vision-centric retrieval-augmented multimodal knowledge. “Diverse scenarios” refers to whether a benchmark categorized different scenarios during evaluation. : Web, : ImageNet [Russakovsky et al., 2015], : Flowers102 [Nilsback and Zisserman, 2008], : StanfordCars [Krause et al., 2013].

requires models to reason about visual entities that undergo perspective changes, such as *angle*, *partial*, *scope* and *occlusion*, as well as transformative changes, such as *temporal*, *incomplete*, *biological* and *deformations*. Additionally, MRAG-BENCH includes 9,673 human-selected images, which serves as the ground-truth image knowledge corpus for model evaluation.

We conduct extensive experiments on MRAG-BENCH to evaluate 10 open-source and 4 proprietary LVLMS. The results confirm that MRAG-BENCH is vision-centric, as all LVLMS show greater improvements when augmented with images compared to textual knowledge. Our results indicate that the best-performing GPT-4o model only achieve 68.68% and 74.5% of accuracy without RAG knowledge and with ground-truth (GT) RAG knowledge, respectively. This substantially outperforms the best open-source model LLaVA-OneVision by 15.39% and 15.52%, respectively. Notably, we observe while all models improve with GT knowledge, only proprietary models are able to effectively utilize noisy retrieved multimodal knowledge. This indicates the gap between open-source and close-source models still exists. Open-source models are falling short on their parametric knowledge and the ability to distinguish between high-quality and poor-quality retrieved visually augmented examples. In comparison to humans, GPT-4o achieves only a 5.82% improvement when augmented with GT knowledge and 0.28% with retrieved knowledge, whereas humans demonstrate

Statistic	Number
Total questions	1,353
- Multiple-choice questions	1,353 (100%)
- Questions newly annotated	1,353 (100%)
Total Scenarios	9
Unique number of questions	375
Unique number of answers	663
Total number of images	16,130
Unique number of images	16,130
Human selected images	9,673
Average image size (px)	1076 x 851
Maximum question length	20
Maximum answer length	9
Average question length	8.03
Average answer length	2.16

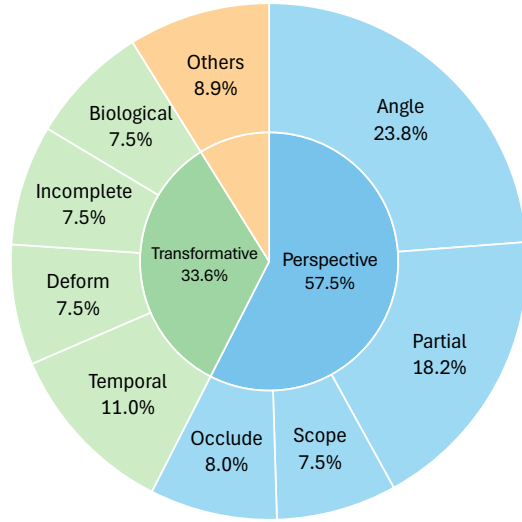


Table 4.2: Key statistics of MRAG-BENCH. Figure 4.2: Scenarios distribution of MRAG-BENCH.

a 33.16% and 22.91% improvement, respectively. These results highlight the importance of MRAG-BENCH in encouraging the community to develop LVLMS better utilizing of visually augmented knowledge.

4.2 MRAG-BENCH

4.2.1 Benchmark Overview

Our benchmark is designed for systematic evaluation of LVLMS’s vision-centric multimodal RAG abilities. To achieve this, we focus on evaluating the model’s understanding of image objects that are not commonly associated with its knowledge base, while the collected ground-truth images can help incentivize specific visual concepts within LVLMS’ memory. Therefore, we divide our benchmark into two main aspects, as illustrated in the examples in Figure 4.1:

- *perspective*, refers to the challenges in visual recognition and reasoning that arise when a visual entity is presented from varying viewpoints, scopes, or levels of visibility.
- *transformative*, refers to the challenges that arise when a visual entity undergoes fine-grained physical transformations, making it unfamiliar or not easily associated with the model’s prior

knowledge.

MRAG-BENCH consists of 16,130 images and 1,353 multiple choice questions, with key statistics shown in Table 4.2. MRAG-BENCH adheres to the following design principles: (1) it focuses on real-world scenarios where visually augmented information is useful; (2) it incorporates 9 diverse multimodal RAG scenarios covering various types of image objects; (3) it features cleaned ground-truth images for each question that align with human knowledge; and (4) it provides robust evaluation settings for deterministic evaluations. Unlike previous works focus on retrieving textual knowledge, evaluation on MRAG-BENCH focuses on retrieving vision-centric knowledge, which can be formulated as follows: Given a query tuple \mathbf{Q} composed of (query image, textual question), the multimodal retriever \mathcal{R} returns a set of relevant images $\mathbf{I} ([\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_N])$, then the LVLMM \mathcal{M} take the input (\mathbf{Q}, \mathbf{I}) and output the final answer.

4.2.2 Benchmark Composition

MRAG-BENCH provides a systematic evaluation across 9 distinctive multimodal RAG scenarios, with four scenarios focused on the *perspective* understanding of visual entities, four on *transformative* understanding, and one categorized as “others”. As illustrated in Figure 4.2, each scenario comprises 7.5% to 23.8% of the whole benchmark. The selected examples of each scenario is shown in Figure 4.3. The details of each scenario are introduced as follows.

Perspective understanding aspect. First, we have *perspective* aspect comprising [ANGLE], [PARTIAL], [SCOPE], and [OCCLUSION] dimensions.

- [ANGLE] evaluates the ability of models to utilize visual knowledge of common shooting angles to identify and reason about less common, long-tailed viewpoints of visual entities.
- [PARTIAL] evaluates the ability of models to use complete appearance knowledge to identify and reason when only a partial image of the visual entities is available.



Figure 4.3: Qualitative examples on MRAG-BENCH. For each scenario, we show the result of GPT-4o [OpenAI, 2023], Gemini Pro [Team et al., 2023], LLaVA-Next-Interleave [Li et al., 2024b] and Mantis-8B-Siglip [Jiang et al., 2024a]. The ground-truth answer is in blue.

- [SCOPE] evaluates the ability of models to leverage high-resolution, detailed images for identifying and reasoning about visual entities in longer-scoped, low-resolution images.
- [OCCLUSION] evaluates the ability of models to use ground-truth image knowledge to identify and reason when visual entities are occluded or partially hidden in natural scenes.

Transformative understanding aspect. On the other hand, the *transformative* understanding scenarios cover [TEMPORAL], [DEFORMATION], [INCOMPLETE], and [BIOLOGICAL] dimensions.

- [TEMPORAL] evaluates the ability of models to use familiar image knowledge to identify and reason about visual entities undergoing temporal changes that may not be represented in the model’s knowledge base.
- [DEFORMATION] evaluates the ability of models to use intact physical appearance knowledge to identify and reason when visual entities undergo deformation not captured in the model’s knowledge base.
- [INCOMPLETE] evaluates the ability of models to compare and contrast the complete layout and structure of image knowledge to identify and reason about missing parts and the correct layout of visual entities.
- [BIOLOGICAL] evaluates the ability of models to utilize image knowledge after biological transformations of the visual entities.

[OTHERS] aims to evaluate the ability of models to leverage geographic image knowledge to accurately identify and reason about the correct regions of origin for the visual entities of interest. All these scenarios work in tandem to comprehensively evaluate LVLMS’ abilities of leveraging visually augmented knowledge.

4.2.3 Data Collection

As the guidelines discussed in § 4.2.1, our benchmark collection involves a clean ground-truth image corpus that can resonate with model’s internal knowledge and a query question and image that challenge model’s memory according to our definition of 9 diverse scenarios. To collect a dataset for systematic evaluation of vision-centric multimodal RAG scenarios, we manually annotate all multiple-choice question answering (MCQA) data while sourcing images from either publicly available datasets or manually scraping them from the web.

Collection of *perspective* aspect. To collect diverse image objects and knowledge that are not extensively represented in LVLMS’ memories [Zhang et al., 2024c], we considered three sources of

data, ImageNet [Russakovsky et al., 2015], Oxford Flowers102 [Nilsback and Zisserman, 2008], and StanfordCars [Krause et al., 2013]. To construct a high quality image corpus, for each of the image class that we included in our benchmark, we examined the validation set and excluded the unqualified images which can't provide sufficient visual information for the recognition of this class. Among the selected corpus, we further humanly picked five representative examples covering the diverse aspects of each class object, as the five ground-truth examples in our experimental results (See §4.3). For constructing the query images, we adhered to our scenario definitions and manually selected qualified images for the [ANGLE], [SCOPE], and [OCCLUSION] scenarios. For the [PARTIAL] scenario, we randomly cropped images by 50% in both height and width. Then we performed another human inspection to ensure the quality of the cropped images, filtering out examples where the visual object did not occupy the dominant area of the image. We repeated the random cropping process until satisfactory images were obtained, filtering to 20.4 GT images per question on average.

Collection of *transformative* aspect. We chose to manually scrape images from the web based on the definitions of the *transformative* aspect. To construct the image corpus, we employed Bing Image Search for each of the image object keyword predefined by us, please refer to Appendix C.1.1 for more details. We filtered out image objects that did not form a clear transformative pair between the query image and the ground-truth image, retaining approximately 74% of the keyword names in the process. For ground-truth image examples, we employed automatic scripts to download the top 15 images related to its keyword names and human filtered out the unqualified image. On average, this results to 5.9 images per question and the five ground-truth images used during our evaluation are manually selected same as in *perspective* aspect.

According to our guidelines, additional related image object knowledge from the same geographic region can assist in identifying that region more effectively. For the [OTHERS] scenario, we source the data from the GeoDE dataset [Ramaswamy et al., 2023]. For each distinct image object category, we randomly sampled 3 out of 6 regions to serve as the answers for each question

and selected the corresponding image as the query image.

Quality control. After constructing the entire benchmark, we implemented two quality control procedures: an automatic check with predefined rules and a manual examination of each instance. The automatic check verifies the correct MCQA format, assesses image validity and filters out redundant images in the corpus, more details are presented in Appendix C.1.1. The manual examination is conducted by two experts in this field, who checked the correspondence between query images and ground-truth image examples, and filtered or revised ambiguous questions and uncorrelated query image and ground-truth images.

4.3 Experiments

In this section, we first introduce the experimental setup and evaluation metric (§ 4.3.1). Then, we present a comprehensive evaluation of 14 recent LVLMs (§ 4.3.2). We demonstrate the importance of visual knowledge and discuss the critical findings revealed by the results from MRAG-BENCH.

4.3.1 Experimental Setup

We evaluate 14 popular LVLMs on MRAG-BENCH, including 4 proprietary models and 10 open-sourced models that can accept multi-image inputs:

- **Proprietary models:** GPT-4o (0513) [OpenAI, 2023], GPT-4-Turbo [OpenAI, 2023], Gemini Pro [Team et al., 2023], and Claude 3.5 Sonnet [Anthropic, 2024].
- **Open-source models:** OpenFlamingo (v2-9B) [Awadalla et al., 2023], Idefics (v2-8B) [Laurençon et al., 2024], VILA (v1.5-13B) [Lin et al., 2023], LLaVA-NeXT-Interleave-7B [Li et al., 2024b], LLaVA-OneVision [Li et al., 2024a], Mantis (clip-llama3, and siglip-llama3 versions; 8B) [Jiang et al., 2024a], mPLUG-Owl3-7B [Ye et al., 2024], Deepseek-VL-7B-chat [Lu et al., 2024a], and Pixtral-12B [Team, 2024].

Model	Overall	Perspective				Transformative				Others
		Angle	Partial	Scope	Occlusion	Temporal	Deformation	Incomplete	Biological	
Random chance	24.83	27.64	23.98	24.51	19.44	22.15	25.49	29.41	25.49	22.5
Human performance	38.47	25.16	34.96	31.37	41.67	21.48	24.51	58.82	54.9	53.33
+ Retrieved RAG	61.38 ^{+22.91}	62.42 ^{+37.26}	60.16 ^{+25.2}	58.82 ^{+27.45}	62.96 ^{+21.29}	54.36 ^{+32.88}	49.02 ^{+24.51}	78.43 ^{+19.61}	63.73 ^{+8.83}	62.5 ^{+9.17}
+ GT RAG	71.63 ^{+33.16}	83.85 ^{+58.69}	70.33 ^{+35.37}	66.67 ^{+35.3}	69.44 ^{+27.77}	59.73 ^{+38.25}	68.63 ^{+44.12}	83.33 ^{+24.51}	73.53 ^{+18.63}	69.17 ^{+15.84}
<i>Open-Source LLMs</i>										
OpenFlamingo-v2-9B	26.83	27.95	26.02	31.37	30.56	29.53	34.31	20.59	17.65	21.67
+ Retrieved RAG	28.31 ^{+1.48}	29.5 ^{+1.55}	28.86 ^{+2.84}	28.43 ^{-2.94}	30.56 ^{+0.0}	34.23 ^{+4.7}	31.37 ^{-2.94}	22.55 ^{+1.96}	21.57 ^{+3.92}	22.5 ^{+0.83}
+ GT RAG	28.90 ^{+2.07}	26.71 ^{-1.24}	33.74 ^{+7.72}	28.43 ^{-2.94}	33.33 ^{+2.77}	35.57 ^{+6.04}	27.45 ^{-6.86}	27.45 ^{+6.86}	25.49 ^{+7.84}	18.33 ^{-3.34}
Idefics2-8B	31.04	31.06	33.33	31.37	38.89	30.2	35.29	25.49	24.51	26.67
+ Retrieved RAG	30.16 ^{-0.88}	29.81 ^{-1.25}	27.64 ^{-5.69}	29.41 ^{-1.96}	36.11 ^{-2.78}	36.24 ^{+6.04}	28.43 ^{-6.86}	27.45 ^{+1.96}	32.35 ^{+7.84}	25.83 ^{-0.84}
+ GT RAG	37.03 ^{+5.99}	36.34 ^{+5.28}	35.37 ^{+2.04}	38.24 ^{+6.87}	54.63 ^{+15.74}	47.65 ^{+17.45}	36.27 ^{+0.98}	24.51 ^{-0.98}	34.31 ^{+9.8}	25.83 ^{-0.84}
VILA1.5-13B	43.68	45.34	41.87	52.94	48.15	50.34	38.24	21.57	30.39	57.5
+ Retrieved RAG	35.48 ^{-8.2}	33.54 ^{-11.8}	28.86 ^{-13.01}	29.41 ^{-23.53}	40.74 ^{-7.41}	47.65 ^{-2.69}	33.33 ^{-4.91}	22.55 ^{+0.98}	33.33 ^{+2.94}	54.17 ^{-3.33}
+ GT RAG	47.01 ^{+3.33}	45.65 ^{+0.31}	46.75 ^{+4.88}	39.22 ^{-13.72}	51.85 ^{+3.7}	53.69 ^{+3.35}	43.14 ^{+4.9}	25.49 ^{+3.92}	44.12 ^{+13.73}	69.17 ^{+11.67}
Mantis-8B-clip-llama3	40.8	45.03	39.43	42.16	49.07	49.66	36.27	28.43	19.61	45.0
+ Retrieved RAG	36.88 ^{-3.92}	36.65 ^{-8.38}	34.96 ^{-4.47}	42.16 ^{0.0}	47.22 ^{-1.85}	50.34 ^{+0.68}	33.33 ^{-2.94}	18.63 ^{-9.8}	21.57 ^{+1.96}	42.5 ^{-2.5}
+ GT RAG	44.72 ^{+3.92}	48.14 ^{+3.11}	46.75 ^{+7.32}	43.14 ^{+0.98}	54.63 ^{+5.56}	57.05 ^{+7.39}	45.1 ^{+8.83}	19.61 ^{-8.82}	18.63 ^{-0.98}	51.67 ^{+6.67}
Mantis-8B-siglip-llama3	45.01	46.89	45.12	57.84	58.33	45.64	45.1	26.47	29.41	45.0
+ Retrieved RAG	39.62 ^{-5.39}	42.55 ^{-4.34}	35.37 ^{-9.75}	47.06 ^{-10.78}	47.22 ^{-11.11}	42.95 ^{-2.69}	45.1 ^{0.0}	23.53 ^{-2.94}	29.41 ^{0.0}	40.83 ^{-4.17}
+ GT RAG	48.85 ^{+3.84}	54.66 ^{+7.77}	52.85 ^{+7.73}	51.96 ^{-5.88}	58.33 ^{0.0}	48.99 ^{+3.35}	50.0 ^{+4.9}	21.57 ^{-4.9}	33.33 ^{+3.92}	49.17 ^{+4.17}
Deepseek-VL-7B-chat	43.39	45.34	47.56	47.06	45.37	46.31	48.04	28.43	20.59	49.17
+ Retrieved RAG	34.66 ^{-8.73}	33.54 ^{-11.8}	32.11 ^{-15.45}	33.33 ^{-13.73}	37.04 ^{-8.33}	43.62 ^{-2.69}	40.2 ^{-7.84}	20.59 ^{-7.84}	26.47 ^{+5.88}	45.0 ^{-4.17}
+ GT RAG	50.33 ^{+6.94}	54.04 ^{+8.7}	56.5 ^{+8.94}	50.98 ^{+3.92}	56.48 ^{+11.11}	57.05 ^{+10.74}	50.0 ^{+1.96}	21.57 ^{-6.86}	23.53 ^{+2.94}	60.83 ^{+11.66}
LLaVA-NeXT-Interleave-7B	43.46	44.41	43.5	40.2	64.81	44.97	44.12	32.35	26.47	45.83
+ Retrieved RAG	40.35 ^{-3.11}	40.06 ^{-4.35}	33.33 ^{-10.17}	39.22 ^{-0.98}	56.48 ^{-8.33}	43.62 ^{-1.35}	44.12 ^{+0.0}	27.45 ^{-4.9}	36.27 ^{+9.8}	49.17 ^{+3.34}
+ GT RAG	52.99 ^{+9.53}	54.97 ^{+10.56}	54.88 ^{+11.38}	49.02 ^{+8.82}	62.04 ^{-2.77}	52.35 ^{+7.38}	47.06 ^{+2.94}	38.24 ^{+5.89}	48.04 ^{+21.57}	61.67 ^{+15.84}
mPLUG-Owl3-7B	49.74	48.45	50.81	54.9	58.33	54.36	51.96	30.39	45.1	51.67
+ Retrieved RAG	41.83 ^{-7.91}	40.06 ^{-8.39}	36.59 ^{-14.22}	40.2 ^{-14.7}	50.0 ^{-8.33}	50.34 ^{-4.02}	46.08 ^{-5.88}	20.59 ^{-9.8}	51.96 ^{+6.86}	46.67 ^{-5.0}
+ GT RAG	56.32 ^{+6.58}	58.39 ^{+9.94}	58.94 ^{+8.13}	58.82 ^{+3.92}	62.96 ^{+4.63}	61.74 ^{+7.38}	59.8 ^{+7.84}	26.47 ^{-3.92}	50.0 ^{+4.9}	58.33 ^{+6.66}
LLaVA-OneVision	53.29	58.39	56.1	49.02	60.19	47.65	53.92	37.25	52.94	51.67
+ Retrieved RAG	50.11 ^{-3.18}	50.93 ^{-7.46}	48.78 ^{-7.32}	50.0 ^{+0.98}	60.19 ^{+0.0}	50.34 ^{+2.69}	48.04 ^{-5.88}	33.33 ^{-3.92}	53.92 ^{+0.98}	54.17 ^{+2.5}
+ GT RAG	58.98 ^{+5.69}	62.42 ^{+4.03}	63.82 ^{+7.72}	59.8 ^{+10.78}	66.67 ^{+6.48}	59.73 ^{+12.08}	53.92 ^{+0.0}	30.39 ^{-6.86}	57.84 ^{+4.9}	60.83 ^{+9.16}
Pixtral-12B	47.97	52.48	45.53	58.82	50.0	51.68	49.02	38.24	42.16	37.5
+ Retrieved RAG	45.97 ^{-2.0}	51.86 ^{-0.62}	40.24 ^{-5.29}	53.92 ^{-4.9}	50.93 ^{+0.93}	49.66 ^{-2.02}	47.06 ^{-1.96}	19.61 ^{-18.63}	47.06 ^{+4.9}	46.67 ^{+9.17}
+ GT RAG	59.28 ^{+11.31}	63.04 ^{+10.56}	63.41 ^{+17.88}	65.69 ^{+6.87}	66.67 ^{+16.67}	61.74 ^{+10.06}	59.8 ^{+10.78}	20.59 ^{-17.65}	50.98 ^{+8.82}	65.0 ^{+27.5}
<i>Proprietary LLMs</i>										
GPT-4-Turbo	57.21	64.29	59.35	54.9	56.48	62.42	47.06	41.18	59.8	50.0
+ Retrieved RAG	58.95 ^{+1.74}	66.53 ^{+2.24}	59.94 ^{+0.59}	53.94 ^{-0.96}	66.74 ^{+10.26}	59.73 ^{-2.69}	49.06 ^{+2.0}	38.27 ^{-2.91}	62.78 ^{+2.98}	58.83 ^{+8.83}
+ GT RAG	62.85 ^{+5.64}	68.94 ^{+4.65}	69.51 ^{+10.16}	60.78 ^{+5.88}	67.59 ^{+11.11}	63.33 ^{+0.91}	51.96 ^{+4.9}	38.24 ^{-2.94}	59.8 ^{+0.0}	62.5 ^{+12.5}
Gemini Pro	61.71	68.01	69.92	73.53	71.3	70.47	42.16	39.22	53.92	40.83
+ Retrieved RAG	65.93 ^{+4.22}	73.29 ^{+5.28}	69.92 ^{+0.0}	69.61 ^{-3.92}	73.15 ^{+1.85}	75.84 ^{+5.37}	49.02 ^{+6.86}	34.31 ^{-4.91}	56.86 ^{+2.94}	65.0 ^{+24.17}
+ GT RAG	71.40 ^{+9.69}	77.33 ^{+9.32}	79.27 ^{+9.35}	78.43 ^{+4.9}	75.93 ^{+4.63}	78.52 ^{+8.05}	54.9 ^{+12.74}	36.27 ^{-2.95}	61.76 ^{+7.84}	72.5 ^{+31.67}
Claude 3.5 Sonnet	59.87	70.19	57.72	56.86	57.41	68.46	48.04	49.02	62.75	47.5
+ Retrieved RAG	63.56 ^{+3.69}	73.91 ^{+3.72}	70.73 ^{+13.01}	56.86 ^{+0.0}	62.96 ^{+5.55}	70.47 ^{+2.01}	55.88 ^{+7.84}	31.37 ^{-17.65}	62.75 ^{+0.0}	53.33 ^{+5.83}
+ GT RAG	71.10 ^{+11.23}	78.88 ^{+8.69}	80.49 ^{+22.77}	76.47 ^{+19.61}	70.37 ^{+12.96}	75.17 ^{+6.71}	67.65 ^{+19.61}	36.27 ^{-12.75}	65.69 ^{+2.94}	59.17 ^{+11.67}
GPT-4o	68.68	76.09	70.42	69.61	74.07	73.82	61.21	47.62	58.82	65.83
+ Retrieved RAG	68.96 ^{+0.28}	77.95 ^{+1.86}	78.86 ^{+8.44}	69.61 ^{+0.0}	75.0 ^{+0.93}	73.83 ^{+0.01}	54.9 ^{+7.28}	26.47 ^{-34.74}	59.8 ^{+0.98}	68.33 ^{+2.5}
+ GT RAG	74.50 ^{+5.82}	84.47 ^{+8.38}	77.46 ^{+7.04}	82.35 ^{+12.74}	79.63 ^{+5.56}	77.18 ^{+3.36}	68.62 ^{+7.41}	30.95 ^{-16.67}	62.75 ^{+3.93}	80.0 ^{+14.17}

Table 4.3: Accuracy scores on MRAG-BENCH. The highest scores for **open-source** models in each section and **proprietary** models are highlighted in blue and red, respectively. Both Retrieved RAG and GT RAG employ top-5 image examples (except for the incomplete scenario, where a single example is intuitively sufficient). The relative difference in performance compared to the score without RAG is shown in subscript, with **blue** indicating performance drops and **red** indicating improvements.

Evaluation setup. We follow standard MCQA evaluation setup and employ accuracy score as our metric. We adopt default generation hyper-parameters selected by each model. Following Lu et al. [2024b], we employ GPT-3.5-turbo to extract the multiple choice answer in rare cases where our pre-defined automatic extraction rules failed. We refer the readers to Appendix C.1.1 and C.2 for more details on evaluation prompts for both without multimodal RAG and with multimodal RAG scenarios, answer extraction prompt and human performance evaluation protocol.

4.3.2 Main Results

As shown in Table 4.3, the average performance of the most advanced LVLMs is not better than 68.68% without multimodal RAG knowledge, and 74.5% with ground-truth knowledge, which demonstrates MRAG-BENCH to be a challenging benchmark. The mean accuracies of open-source LVLMs are between 26.83% and 53.29% without RAG knowledge and between 28.90% and 59.28% with ground-truth knowledge, which fall behind from advanced proprietary LVLMs. Notably, MRAG-BENCH proves to be knowledge-intensive as average humans achieved 38.47% without RAG knowledge, while proprietary LVLMs generally perform well, suggesting that their extensive training data equips them with a broader knowledge base. However, when provided with either retrieved or ground-truth knowledge, humans achieve the most significant improvements of 22.91% and 33.16%, respectively. This underscore the need of LVLMs to better utilize visually augmented information like humans.

Can LVLMs utilize retrieved and ground-truth image knowledge well? As illustrated in Table 4.3, all models demonstrate improvement when ground-truth image RAG knowledge is provided. Among the open-source models, they achieve improvements ranging from 2.07% to 11.31% when using ground-truth RAG knowledge, whereas 5.64% to 9.69% improvements are observed from proprietary LVLMs. Interestingly, when images from the multimodal retriever is provided, almost all open-source LVLMs on average demonstrate a declined performance while proprietary models can still gain improvement. This indicate proprietary models possess emerging

abilities to distinguish between good and bad image knowledge sources, which is a critical skill in the multimodal RAG domain. We further conducted a qualitative analysis to investigate the reasons behind this, as detailed in the following paragraphs.

Fine-grained results. We also report fine-grained scores across 9 scenarios on MRAG-BENCH in Table 4.3. Remarkably, GPT-4o surpasses most other baselines in various categories, with exceptions in problems related to partial, incomplete and biological scenarios. Notably, GPT-4o outperforms human performance on all perspective aspect as well as on temporal and deformation scenarios within the transformative aspect. We conjecture that incomplete and biological scenarios are less likely to be included in the training knowledge. Interestingly, all models exhibit a decline in performance on incomplete scenarios, with only a few exceptions, while humans find this task relatively easy, achieving 58.82% and 83.33% scores with ground-truth knowledge. This further highlights the importance of leveraging retrieved visually augmented knowledge to address questions that do not directly incentivize knowledge stored in the models’ memories.

Why can proprietary models better utilize retrieved images? We conduct an error analysis on an open-source model (LLaVA-Next-Interleave) and a proprietary model (Gemini Pro). For a fair comparison, we filtered results where LLaVA-Next-Interleave answered correctly without or with GT knowledge but was misled to wrong answer with retrieved examples. One example is illustrated in Figure 4.4, the retrieved images contain two correct examples and three false examples. While Gemini Pro is able to utilize all retrieved images, LLaVA-Next-Interleave leverages bad examples and makes wrong prediction. This example helps explain why do almost all open-source models have lower performance with retrieved knowledge.

4.4 Analysis

In this section, we conduct quantitative analysis addressing three important questions: 1) To what extent can LVLMs benefit more from visual knowledge than from textual knowledge on MRAG-

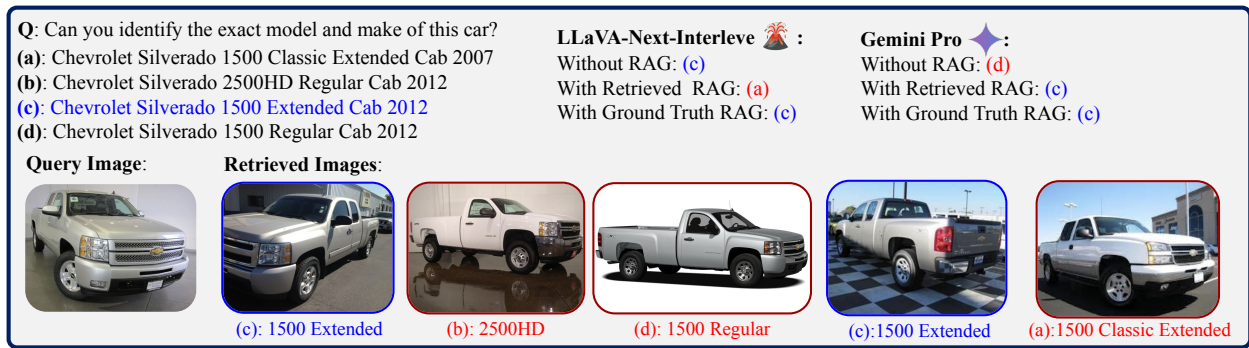


Figure 4.4: Qualitative Example of Proprietary model (Gemini Pro) identifies and utilizes correct examples, while open-source model (LLaVA-Next-Interleave) is misled by noisy retrieved information, resulting in incorrect answers.

BENCH? (§ 4.4.1) 2) How does the performance of LVLMs vary with examples retrieved from different retrievers? (§ 4.4.2) 3) How many ground-truth visual knowledge examples are required for LVLMs to continue benefiting? (§ 4.4.3)

4.4.1 How much can visual knowledge benefit more than textual knowledge?

We used the Wikipedia corpus as of 2023/07/01 as our text knowledge corpus¹. To ensure a fair comparison, we employed the same multimodal retriever (CLIP) for retrieving either text or image knowledge. The top-5 ranked documents or images are used for augmenting the input. We selected one open-source (LLaVA-Next-Interleave) and one proprietary (GPT-4-Turbo) LVLM to examine their preference for textual knowledge versus image knowledge on MRAG-BENCH. As shown in Table 4.4, when both models utilized retrieved knowledge, LLaVA-Next-Interleave demonstrated a 2.36% improvement with image knowledge over text knowledge, while GPT-4-Turbo showed a 2.34% improvement. When using GT knowledge, LLaVA-Next-Interleave exhibited an 11.09% improvement with image knowledge over text knowledge, compared to a 3.87% improvement for GPT-4-Turbo. Interestingly, when both GT image and text knowledge are provided, LLaVA-Next-Interleave indicated less improvement than with GT image alone whereas GPT-4-Turbo further pushed its performance. All these results demonstrate that retrieving visual knowledge is more

¹<https://www.kaggle.com/datasets/jjinho/wikipedia-20230701>

Model	Overall	Perspective				Transformative				Others
		Angle	Partial	Scope	Occlusion	Temporal	Deformation	Incomplete	Biological	
LLaVA-NeXT-Interleave-7B	43.46	44.41	43.5	40.2	64.81	44.97	44.12	32.35	26.47	45.83
+ Retrieved Text RAG	37.99 _{-5.47}	37.58 _{-6.83}	34.96 _{-8.54}	33.33 _{-6.87}	50.0 _{-14.81}	41.61 _{-3.36}	35.29 _{-8.83}	30.39 _{-1.96}	27.45 _{+0.98}	51.67 _{+5.84}
+ Retrieved Image RAG	40.35 _{-3.11}	40.06 _{-4.35}	33.33 _{-10.17}	39.22 _{-0.98}	56.48 _{-8.33}	43.62 _{-1.35}	44.12 _{+0.0}	27.45 _{-4.9}	36.27 _{+9.8}	49.17 _{+3.34}
+ GT Text RAG	41.09 _{-2.37}	41.93 _{-2.48}	39.02 _{-4.48}	38.24 _{-1.96}	56.48 _{-8.33}	44.97 _{+0.0}	43.14 _{-0.98}	30.39 _{-1.96}	21.57 _{-4.9}	50.83 _{+5.0}
+ GT Image RAG	52.99 _{+9.53}	54.97 _{+10.56}	54.88 _{+11.38}	49.02 _{+8.82}	62.04 _{-2.77}	52.35 _{+7.38}	47.06 _{+2.94}	38.24 _{+5.89}	48.04 _{+21.57}	61.67 _{+15.84}
+ GT Image & Text RAG	47.82 _{+4.36}	47.83 _{+3.42}	48.78 _{+5.28}	44.12 _{+3.92}	58.33 _{-6.48}	49.66 _{+4.69}	48.04 _{+3.92}	30.39 _{-1.96}	35.29 _{+8.82}	62.5 _{+16.67}
GPT-4-Turbo	57.21	64.29	59.35	54.9	56.48	62.42	47.06	41.18	59.8	50.0
+ Retrieved Text RAG	56.61 _{-0.6}	61.8 _{-2.49}	59.35 _{+0.0}	59.8 _{+4.9}	58.33 _{+1.85}	59.06 _{-3.36}	49.02 _{+1.96}	33.33 _{-7.85}	60.78 _{+0.98}	52.5 _{+2.5}
+ Retrieved Image RAG	58.95 _{+1.74}	66.53 _{+2.24}	59.94 _{+0.59}	53.94 _{-0.96}	66.74 _{+10.26}	59.73 _{-2.69}	49.06 _{+2.0}	38.27 _{-2.91}	62.78 _{+2.98}	58.83 _{+8.83}
+ GT Text RAG	58.98 _{+1.77}	68.01 _{+3.72}	63.41 _{+4.06}	65.69 _{+10.79}	63.89 _{+7.41}	59.73 _{-2.69}	38.24 _{-8.82}	37.25 _{-3.93}	58.82 _{-0.98}	50.83 _{+0.83}
+ GT Image RAG	62.85 _{+5.64}	68.94 _{+4.65}	69.51 _{+10.16}	60.78 _{+5.88}	67.59 _{+11.11}	63.33 _{+0.91}	51.96 _{+4.9}	38.24 _{-2.94}	59.8 _{+0.0}	62.5 _{+12.5}
+ GT Image & Text RAG	65.11 _{+7.9}	72.05 _{+7.76}	72.76 _{+13.41}	67.65 _{+12.75}	70.37 _{+13.89}	71.81 _{+9.39}	46.08 _{-0.98}	39.22 _{-1.96}	60.78 _{+0.98}	57.5 _{+7.5}

Table 4.4: LVLMS performance on MRAG-BENCH with textual knowledge v.s visual knowledge. Both the open-source and proprietary model benefit more from image knowledge.

helpful than retrieving text on MRAG-BENCH.

4.4.2 How does retriever performance affect LVLMS?

We picked four recent best-performing multimodal retrievers, including CLIP [Radford et al., 2021], MagicLens [Zhang et al., 2024a], E5-V [Jiang et al., 2024b], VISTA [Zhou et al., 2024] and evaluated their performance (Recall@5). The detailed retriever performance can be found at Table C.2 in Appendix C.3. We selected LLaVA-Next-Interleave as the end model to assess its performance. As shown in Figure 4.5, when retrievers achieve higher Recall@5 scores (i.e., better retrieved examples), the LVLMS’s accuracy tends to improve, demonstrating a strong 95% positive correlation. Interestingly, despite similar Recall@5 scores from CLIP and VISTA retrievers, LLaVA-Next-Interleave demonstrated a 2.07% gap in overall accuracy. We conjecture that the order of the correctly retrieved examples may also impact the model’s final performance. The sensitivity to the order of retrieved examples is a common issue that persists across various models. Although this phenomenon, known as position bias, has been examined in text-based RAG [Lu et al., 2022b, Wang et al., 2023c], its impact on visual RAG remains unexplored, presenting a promising direction for future research.

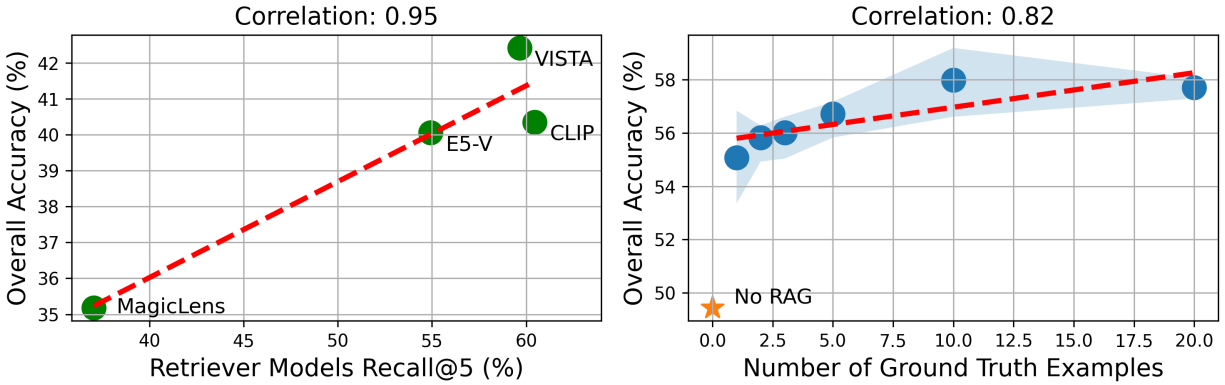


Figure 4.5: Left: LLaVA-Next-Interleave results with 4 different multimodal retrievers. Its performance using retrieved images correlates 95% with retriever’s Recall@5 scores. Right: Average results of three random seed runs. Improve the number of ground-truth RAG examples shows steady increase of model’s performance, reaches the maximum with 10 examples.

4.4.3 How many ground-truth image examples are needed?

For simplicity, all our experiments used five retrieved or ground-truth image examples. However, it is worth exploring how many examples LVLMs can effectively leverage. As noted in § 4.2.3, the perspective aspect of our benchmark includes an average of 20.4 ground-truth examples. To investigate further, we perform an analysis focusing on the perspective and others aspects, covering a total of 892 questions. As shown in Figure 4.5, we evaluated LLaVA-Next-Interleave using 1, 2, 3, 5, 10, 20 GT examples, averaging the results across three random seeds for sampling the GT examples. LLaVA-Next-Interleave saw the greatest improvement of 5.64% with just one GT example. Performance continued to increase steadily, reaching a peak at 10 GT examples, which was 0.29% higher than with 20 GT examples. One possible explanation could be LLaVA-Next-Interleave may not able to better leverage visually augmented knowledge in long context scenarios. Moreover, the complexity of questions affects the number of images needed too, one ground-truth example sometimes help the model the most on MRAG-BENCH. We encourage the research on adaptatively deciding the number of necessary images based on the complexity of questions.

4.5 Conclusion

In this work, we introduce MRAG-BENCH, a benchmark specifically designed for vision-centric evaluation for retrieval-augmented multimodal models. Our evaluation of 14 LVLMs highlights that visually augmented knowledge brings more improvements on MRAG-BENCH compared to textual knowledge. Moreover, the top-performing model, GPT-4o, struggles to effectively utilize the retrieved knowledge, achieving only a 5.82% improvement when augmented with relevant information, compared to a 33.16% improvement demonstrated by human participants. We further conduct extensive analysis and propose several promising directions for future research. Our findings underscore the significance of MRAG-BENCH in motivating the community to develop LVLMs that better utilize retrieved visual knowledge.

Chapter 5

Future Work and Conclusion

5.1 Current Challenges and Future Prospects

The field of Large Vision-Language Models (LVLMs) has witnessed significant progress in recent years, with advancements expanding their capabilities across diverse tasks and domains. Yet, despite these achievements, several challenges remain that hinder the realization of truly general and human-aligned multimodal AI. This section highlights current limitations in LVLMs, ongoing advancements in the field, and promising directions for future research.

5.1.1 Scaling Contextual Understanding Across Modalities

While LVLMs excel at single-image tasks, extending their capabilities to handle video, multi-image inputs, and long-context scenarios has become a trending focus in recent research and has made promising progress [Chen et al., 2024, Xue et al., 2024, Zhang et al., 2024b]. Still, current models struggle with maintaining coherence and reasoning over extended temporal or spatial contexts, limiting their applicability in dynamic or multi-view tasks. By incorporating techniques such as memory-augmented networks and hierarchical encoding schemes, future models can improve temporal reasoning, multi-perspective synthesis, and task continuity across extended contexts.

Expanding LVLMs into 3D vision-language tasks, robotics, and embodied AI opens exciting

possibilities for real-world applications. By integrating 3D spatial reasoning, physical interaction, and embodied memory systems, models can enhance their ability to perform complex tasks such as indoor navigation, object assembly, and collaborative robotics.

Unified Multimodal Architectures Current LVLM architectures often separate vision and language processing, leading to inefficiencies and limitations in cross-modal integration. Unified architectures that more effectively combine vision and language representations could achieve greater alignment, coherence, and generalization across tasks. While current research has greatly explored this direction, the performance is still lagging behind cross-modal integration [Fang et al., 2024, Li et al., 2024c, Wu et al., 2024].

5.1.2 Efficient and Lightweight LVLMs

The increasing size and complexity of LVLMs demand prohibitive computational resources, which pose significant challenges for real-world deployment. Although token pruning, model distillation, and parameter-efficient fine-tuning approaches have demonstrated promise, achieving efficient, lightweight LVLMs without sacrificing performance remains an open research problem [Hu et al., 2024a, Wen et al., 2024, Xu et al., 2024]. Innovations in these methods and sparsity-based architectures will be critical for creating scalable, resource-efficient LVLMs. These approaches can democratize access to high-performance multimodal AI by reducing computational requirements, enabling deployment in mobile devices, embedded systems, and edge computing scenarios.

5.1.3 Tool-Integrated and Real-World Interaction

LVLMs have yet to fully leverage external tools such as search engines, real-world APIs, physical sensors and other tools from various modalities. The lack of seamless integration with external systems restricts their ability to perform complex, real-world tasks requiring additional contextual knowledge, dynamic planning, or interaction with external environments. Future LVLMs should seamlessly integrate with external tools to enhance their reasoning and tool-using capabilities.

By leveraging external tools dynamically, these models can extend their utility across real-world applications, including dynamic information retrieval, real-time analytics, and interaction with IoT devices.

5.2 Conclusion

In this thesis, we have explored critical advancements in the development of Large Vision-Language Models (LVLMs) to address challenges in efficiency, evaluation, and knowledge integration. Specifically, we introduced MQT-LLaVA, an adaptive module for encoding visual input with dynamic tokenization, achieving significant computational efficiency. We also developed VALOR-EVAL, a comprehensive evaluation framework that addresses nuanced issues like attribute and relational hallucinations, enhancing the trustworthiness of LVLM outputs. Finally, we proposed MRAG-Bench, a benchmark for assessing the integration of visually augmented knowledge, shedding light on how retrieval-augmented approaches can improve multimodal reasoning.

Our work highlights the remarkable progress LVLMs have made in tackling complex, real-world tasks while emphasizing the challenges that remain. We identified critical barriers, such as scaling LVLMs for long-context scenarios, enhancing their computational efficiency, expanding their applicability to new domains like 3D and robotics, and enabling seamless integration with external tools and real-world environments.

Future developments in LVLMs will require unified architectures that better align vision and language processing, robust evaluation frameworks to guide their refinement, and innovative approaches to enable human-like reasoning and interaction.

By addressing these interconnected challenges, LVLMs have the potential to become not only more capable and efficient but also more aligned with human intelligence. This thesis provides a foundation for advancing the next generation of multimodal AI, paving the way for systems that perceive, reason, and interact with the world as humans do, opening new avenues for both research and impactful real-world applications.

Appendices

Appendix A

Additional Results from Chapter 2

A.1 Additional Results

We present the results of choosing a random number of visual tokens, 77 as shown in Table A.1, to demonstrate our flexibility in selecting any number of tokens during inference.

To demonstrate that the visual tokens used for visualization in Figure 2.4 are not cherry-picked, we present all the first eight tokens in Figure A.1.

Method	LLM	Res.	#Tokens	VizWiz	SQA ^I	VQA ^{v2}	GQA	POPE	MME ^P	MME ^C	MMMU	MMB	LLaVA ^W	MM-Vet	Avg
QT-LLaVA	Vicuna-1.5-7B	336	256	51.1	68.1	76.8*	61.5*	84.1	1431.2	348.2	34.3	64.0	63.9	27.9	58.8
MQT-LLaVA	Vicuna-1.5-7B	336	256	53.1	67.6	76.8*	61.6*	84.4	1434.5	353.6	34.8	64.3	64.6	29.8	59.4
MQT-LLaVA	Vicuna-1.5-7B	336	144	52.0	67.5	76.4*	61.4*	83.9	1446.4	351.8	34.4	64.4	61.4	29.9	58.9
MQT-LLaVA	Vicuna-1.5-7B	336	77	51.6	67.1	75.8*	60.4*	83.6	1457.0	336.1	34.0	64.0	59.9	29.3	58.3
MQT-LLaVA	Vicuna-1.5-7B	336	64	51.5	67.0	75.3*	60.0*	83.6	1464.3	352.9	34.4	63.5	59.4	28.9	58.3

Table A.1: Results of MQT-LLaVA with different numbers of visual tokens. To demonstrate our flexibility in selecting any number of tokens up to 256, we chose a random number of visual tokens during inference, 77, which was not seen during training.

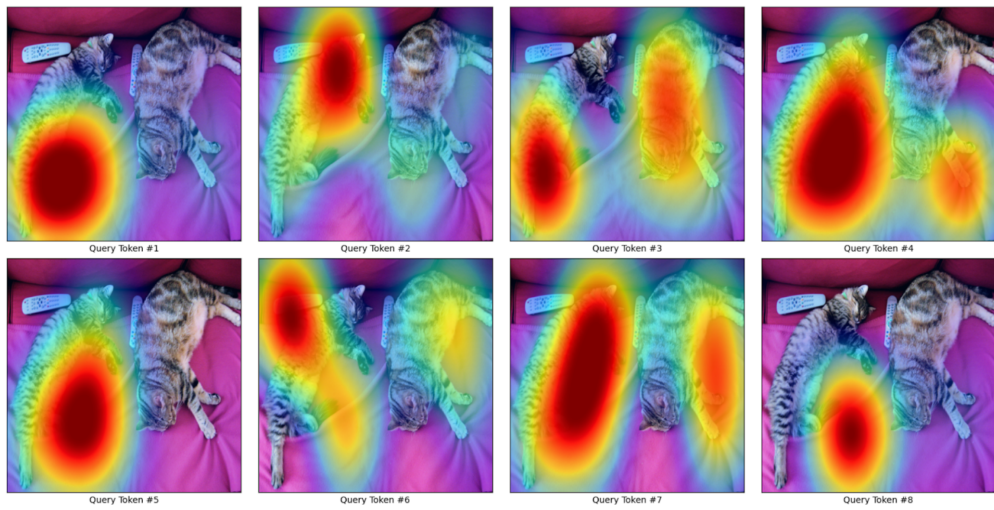


Figure A.1: Grad-CAM visualization from all the tokens in our model when inference with 8 tokens. Input: “How many cats are there in the image? Answer: 2”.

Appendix B

Additional Results from Chapter 3

Model	Visual Encoder	Alignment Network	Language Model
InstructBLIP	EVA CLIP ViT-G/14 _{1.1B}	Q-Former	Vicuna _{7B}
LLaVA-1.5	CLIP ViT-L/14-336px _{0.4B}	MLP	Vicuna-v1.5 _{13B}
MiniGPT-v2	EVA CLIP ViT-G/14 _{1.1B}	Linear Projection	LLaMA-2 _{7B}
mPLUG-Owl2	CLIP ViT-L/14 _{0.4B}	Cross Attention	LLaMA-2 _{7B}
BLIVA	EVA CLIP ViT-G/14 _{1.1B}	Q-Former & Linear Projection	Vicuna _{7B}
CogVLM	EVA2-CLIP-E/14 _{4.7B}	MLP	Vicuna-v1.5 _{7B}
InternLM-Xcomposer2	CLIP ViT-L/14-336px _{0.4B}	Partial Low-Rank Adaptation	InternLM2 _{7B}
Qwen-VL	CLIP ViT-G/14 _{1.9B}	Cross Attention	QwenLM _{13B}
Emu2	EVA2-CLIP-E-plus/14 _{5.0B}	Linear Projection	LLaMA _{33B}
GPT-4(V)	Unknown	Unknown	GPT-4

Table B.1: Architectures of mainstream LVLMs evaluated in our benchmark. InstructBLIP Dai et al. [2023], LLaVA-1.5 Liu et al. [2023a], MiniGPT-v2 Chen et al. [2023a], mPLUG-Owl2 Ye et al. [2023b], BLIVA Hu et al. [2024c], CogVLM Wang et al. [2024], InternLM-XComposer2 Dong et al. [2024], Qwen-VL Bai et al. [2023], Emu2 Sun et al. [2024] and GPT-4V OpenAI [2023].

B.1 Large Vision-Language Models

The recent advancements in large language models (LLMs) Bai et al. [2023], Chiang et al. [2023], OpenAI [2023], Touvron et al. [2023a,c] have sparked a wave of research focused on enhancing vision-language pre-trained models (VLPMS) Alayrac et al. [2022], Kim et al. [2021], Li et al. [2023a]. By incorporating the versatile capabilities of LLMs, these studies aim to improve the

language understanding and generation abilities of VLPMs significantly. In this paper, we refer to the enhanced VLPMs with the integration of LLMs as *Large Vision-Language Models* (LVLMs) Li et al. [2023c]. LVLMs excel in comprehending both the visual semantics of objects in images and the linguistic semantics associated with these objects by leveraging the extensive parametric knowledge embedded in the LLMs. This dual understanding enables LVLMs to conduct intricate reasoning about the concepts related to these objects. Consequently, LVLMs demonstrate strong performance in various traditional multi-modal tasks, such as visual question answering, image captioning, and object detection, highlighting their versatility and robustness in these domains Dai et al. [2023], Hu et al. [2024c], Huang et al. [2023a, 2024], Liu et al. [2023a,b], OpenAI [2023], Ye et al. [2023a], Zhu et al. [2023]. Table B.1 shows comparison of these LVLMs.

B.2 Conditional Probabilities

1. $\mathcal{P}(\text{feature}|\text{object})_{\max}$: maximum conditional probability, highlighting the strongest feature-object associations.
2. $\mathcal{P}(\text{feature}|\text{object})_{\text{avg}}$: average conditional probability, offering a broad view of how features tend to cluster around objects.
3. $\mathcal{P}(\text{feature}|\text{object})_{\max} - \mathcal{P}(\text{feature}|\text{object})_{\text{avg}}$: the difference between the maximum and average conditional probabilities, revealing objects with outlier features.
4. $\mathcal{P}(\text{feature}|\text{object})_{\text{avg}} - \mathcal{P}(\text{feature}|\text{object})_{\min}$: the spread between average and minimum conditional probabilities, indicating the range of commonality among features.
5. $\mathcal{P}(\text{feature}|\text{object})_{\max} - \mathcal{P}(\text{feature}|\text{object})_{\min}$: the range between maximum and minimum conditional probabilities, capturing the full spectrum of feature variability.

B.3 Captions Generation Prompts

- **Object Existence:** Write a detailed description of the image. Provide information about all objects in front and background.
- **Attribute (Object):** Write a detailed description of the image. Provide information about the total number and colors of all objects from left to right and up to bottom.
- **Attribute (People):** Write a detailed description of the image. Provide information about the total number of people and colors of clothes for each person from left to right.
- **Relation (Positional):** Describe the positional relationship between all the objects in the image in detail, using left, right, top, and bottom etc, from the view of the observer.
- **Relation (Comparative):** Rank the size of all the objects in the image in detail, from large to small.

B.4 Features Extraction Prompts

The feature extraction prompts for objects, color and counting attributes, positional relation and comparative relation are illustrated in Table B.2, Table B.3, Table B.4, Table B.5, and Table B.6, respectively.

B.5 Features Matching Prompts

The features matching prompts for objects, color and counting attributes, positional relation and comparative relation are illustrated in Table B.7, Table B.8, Table B.9, Table B.10, and Table B.11, respectively.

System message

You are a language assistant who helps extract information from given sentences.

Prompt

Given an image with a caption that is generated by a vision-language model.

Please act as a linguistic master and extract all the objects from the captions.

Format your response in JSON format, with the key being “objects” and the value being a list of objects.

Please only extract objects without including attributes. For example, extract “field” instead of “grassy field”. Also be mindful of plural forms. For example, extract "cow" instead of “cows”.

Please only extract the object that is a concrete entity in the real world instead of abstract concepts, actions, and moves.

It cannot be an abstract notion such as day, time, scene, moment, image, game, sport, setting, plot, atmosphere, surroundings, group etc.

It cannot be any words describing the emotions such as excitement, enthusiasm, etc.

It cannot be any words describing the positions in the image, such as foreground, background, left, right, etc.

For clarity, consider these examples: {In-context examples}

With these examples in mind, please help me extract the objects based on the factual information in the caption.

Here is the caption: {Input Caption}

Table B.2: Prompt template for extracting **objects**. {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.

B.6 Qualitative Results

We illustrate the qualitative results of three representative models in Figure B.1, Figure B.2 and Figure B.3. Each model exhibited instances of hallucination in these examples from our benchmark VALOR-BENCH. Notably, while GPT-4V generates the most comprehensive results, it is also more prone to producing hallucinations.

System message

You are a language assistant who helps extract information from given sentences.

Prompt

Given an image with a caption that is generated by a vision-language model.

Please act as a linguistic master and extract the total number and colors of all objects as mentioned in the captions.

Your answer should be a dictionary of this format: {"total num of objects": "(NUM, OBJECT)", "objects": {"ORDER": "(ATTRIBUTE, OBJECT)"}}. Remember OBJECT should be in singular format.

For clarity, consider these examples: {In-context examples}

With these examples in mind, please help me extract the objects and attributes based on the factual information in the caption.

Here is the caption: {Input Caption}

Table B.3: Prompt template for extracting **attributes (object)**. {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.

System message

You are a language assistant who helps extract information from given sentences.

Prompt

Given an image with a caption that is generated by a vision-language model.

Please act as a linguistic master and extract the total number of people and colors of clothes for each person as mentioned in the captions.

Your answer should be a dictionary of this format: {"total num of people": "(NUM, PERSON)", "clothes": {"ORDER": "person": "PERSON", "object": "(ATTRIBUTE, OBJECT)", "action": "ACTION"}}. OBJECT can be clothes or accessories (e.g., bags, socks).

For clarity, consider these examples: {In-context examples}

With these examples in mind, please help me extract the objects and attributes based on the factual information in the caption.

Here is the caption: {Input Caption}

Table B.4: Prompt template for extracting **attributes (people)**. {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.

System message

You are a language assistant that helps to extract information from given sentences.

Prompt

Given an image with a caption that is generated by a vision language model.

Please act as a linguistic master and extract a set of words describing the spatial or positional relations between all the visual objects from the captions. Your answer should be a list of values that are in format of object1 relation with object2 with the relation being left, right, top, bottom, middle etc. Do not extract the attribute along with the object and don't extract any relation that is an verb, replace it with simply which object is (on or to the left or etc) the other object or the image. Formulate your response into a JSON object with the key being "relations" and the value being a list of relations. If there are no relations found, please return an empty list.

For clarity, consider these examples: **{In-context examples}**

With these examples in mind, please help me extract the relations based on the information in the caption.

Here is the caption: **{Input Caption}**

Table B.5: Prompt template for extracting **positional relations**. **{In-context examples}** are in-context examples. **{Input caption}** are captions generated by evaluated models.

System message

You are a language assistant that helps to extract ranking from given sentences.

Prompt

Given an image with a caption that is generated by a vision language model.

Given an image with a caption that is generated by a vision language model. Please act as a linguistic master and extract the rank of all the objects from large to small as mentioned in the captions. Your answer should be a dict of values which the keys represent the ranks starting from 1 and values are the No.1 largest object to smallest. If the caption does not mention the order of the object, you can by default view the order of objects appearance as from largest to smallest. If there are no objects mentioned in the caption, you can return an empty dict.

For clarity, consider these examples: {In-context examples}

With these examples in mind, please help me extract the relations based on the information in the caption.

Here is the caption: {Input Caption}

Table B.6: Prompt template for extracting **comparative relations**. {In-context examples} are in-context examples. {Input caption} are captions generated by evaluated models.

System message

You are given a task to match objects from two lists that have the same meaning.

Prompt

Input Lists:

1. “gt-objects”: Ground truth objects in the image.
2. “generated-objects”: Objects identified by a vision-language model.

Matching Criteria:

- For each object in “generated-objects”, find the object in the “gt-objects” that have the same meaning and add it to the “matched-objects” dictionary.
- By the same meaning, we mean the words can be synonyms, can be plural/singular forms of each other and can also have different length of words to express the same meaning of objects, etc.
- Note since we find the matched object for each object in “generated-objects”, it’s ok that multiple objects in “generated-objects” match one object in “gt-object”, list all matches.
- There is special scenario that when you can’t find the matched object in “gt-objects” but you can find one or more object is a subset or a sub category of the generated object, which means that the generated object is a broader concept of the object in “gt-objects”, add it to the “broader-concept” dictionary instead of the “matched-objects”. If there are many objects are a subset or a sub category of the generated object, you can pick anyone of them. Note we are matching for each object in “generated-objects”. If you can find the matched object in “gt-objects”, you should not add it to the “broader-concept” dictionary.

Output:

1. A “broader-concept” dictionary: only if an object from “generated-objects” denotes a broader category of a concept in “gt-objects”. Key = word from “generated-objects”, Value = word from “gt-objects”.
2. A “matched-objects” dictionary: Key = word from “generated-objects”, Value = word from “gt-objects”. It should not contain any words from the “broader-concept” dictionary.

For clarity, consider these examples: **{In-context examples}**

With these examples in mind, please help me extract the broader-concept, and matched-objects from the following two objects lists.

1. gt-objects: **{Input Ground Truth Objects}**
2. generated-objects: **{Input Generated Objects}**

Table B.7: Prompt template for matching **objects** in image caption and reference caption. **{In-context examples}** are in-context examples. **{Input Ground Truth Objects}** are the ground truth objects list **{Input Generated Objects}** are the extracted objects list from the extraction step which are originally captions generated by evaluated models.

System message

You are given a task to match (attributes, objects) from two lists that have the same meaning.

Prompt

Inputs:

1. “gt-att-obj”: A dictionary with order being the key and the ground-truth (attribute, object) pair being the value. Sometimes one object can be, for example “(black, bag), (white, bag), (striped, bag)”, it means either “black” or “white” or “striped” is correct for an attribute related with the “bag” and should be matched.

2. “generated-att-obj”: A dictionary with order being the key and the generated (attribute, object) pair being the value. The order is the order of the object in the generated caption.

Matching Criteria:

- For each (attribute, object) in “generated-att-obj”, find the (attribute, object) in the “gt-att-obj” that have the same meaning and add it to the “matched-att-obj” dictionary.

- By the same meaning, we mean the words can be synonyms, can be plural/singular forms of each other and can also have different length of words to express the same meaning of attributes or objects, etc.

- If you find that the “generated-att-obj” can be matched with the “gt-att-obj” but the attribute or object in “generated-att-obj” is a broader concept of the attribute or object in “gt-att-obj”, for example, one object in “generated-att-obj” is “person”, but the “gt-att-obj” don’t have “person” but specifically have “man”, which is a subcategory of “person”, add it to the “broader-concept” dictionary instead of the “matched-att-obj”.

Output:

1. A “broader-concept” dictionary: {“ORDER2”: {“(ATTRIBUTE1, OBJECT1)”:“(ATTRIBUTE2, OBJECT2)”}} only if an (ATTRIBUTE1, OBJECT1) with ORDER1 from “generated-att-obj” denotes a broader category of an (ATTRIBUTE2, OBJECT2) with ORDER2 in “gt-att-obj”. Notify that Key must be the (ATTRIBUTE1, OBJECT1) from “generated-att-obj”, Value must be (ATTRIBUTE2, OBJECT2) from “gt-att-obj”. If none, it should be an empty dictionary. ORDER1 should be the same as ORDER2.

2. A “matched-att-obj” dictionary: {“ORDER2”: {“(ATTRIBUTE1, OBJECT1)”:“(ATTRIBUTE2, OBJECT2)”}} only if an (ATTRIBUTE1, OBJECT1) with ORDER1 from “generated-att-obj” can be mapped to an (ATTRIBUTE2, OBJECT2) with ORDER2 in “gt-att-obj” with the matching criteria. Key must be (ATTRIBUTE1, OBJECT1) from “generated-att-obj”, Value must be (ATTRIBUTE2, OBJECT2) from “gt-att-obj”. It should not contain any (ATTRIBUTE1, OBJECT1) or (ATTRIBUTE2, OBJECT2) from the “broader-concept” dictionary. ORDER1 should be the same as ORDER2.

- The keys in “broader-concept” and “matched-att-obj” must be the same as “gt-att-obj”.

For clarity, consider these examples: [{In-context examples}](#)

With these examples in mind, please help me extract the broader-concept, and matched-objects from the following two objects lists.

1. gt-objects: [{Input Ground Truth Attributes}](#)

2. generated-objects: [{Input Generated Attributes}](#)

Table B.8: Prompt template for matching **attributes (object)** in image caption and reference caption.

System message

You are given a task to match (attributes, objects) from two lists that have the same meaning.

Prompt

Inputs:

1. “gt-att-obj”: A dictionary with order being the key and the ground-truth (attribute, object) pair being the value. Sometimes one object can be, for example “(black, bag), (white, bag), (striped, bag)”, it means either “black” or “white” or “striped” is correct for an attribute related with the “bag” and should be matched.

2. “generated-att-obj”: A dictionary with order being the key and the generated (attribute, object) pair being the value. The order is the order of the object in the generated caption.

Matching Criteria:

- For each (attribute, object) in “generated-att-obj”, find the (attribute, object) in the “gt-att-obj” that have the same meaning and add it to the “matched-att-obj” dictionary.

- By the same meaning, we mean the words can be synonyms, can be plural/singular forms of each other and can also have different length of words to express the same meaning of attributes or objects, etc.

- If you find that the “generated-att-obj” can be matched with the “gt-att-obj” but the attribute or object in “generated-att-obj” is a broader concept of the attribute or object in “gt-att-obj”, for example, one object in “generated-att-obj” is “person”, but the “gt-att-obj” don’t have “person” but specifically have “man”, which is a subcategory of “person”, add it to the “broader-concept” dictionary instead of the “matched-att-obj”.

Output:

1. A “broader-concept” dictionary: {“ORDER2”: {“(ATTRIBUTE1, OBJECT1)”:“(ATTRIBUTE2, OBJECT2)”}} only if an (ATTRIBUTE1, OBJECT1) with ORDER1 from “generated-att-obj” denotes a broader category of an (ATTRIBUTE2, OBJECT2) with ORDER2 in “gt-att-obj”. Notify that Key must be the (ATTRIBUTE1, OBJECT1) from “generated-att-obj”, Value must be (ATTRIBUTE2, OBJECT2) from “gt-att-obj”. If none, it should be an empty dictionary. ORDER1 should be the same as ORDER2.

2. A “matched-att-obj” dictionary: {“ORDER2”: {“(ATTRIBUTE1, OBJECT1)”:“(ATTRIBUTE2, OBJECT2)”}} only if an (ATTRIBUTE1, OBJECT1) with ORDER1 from “generated-att-obj” can be mapped to an (ATTRIBUTE2, OBJECT2) with ORDER2 in “gt-att-obj” with the matching criteria. Key must be (ATTRIBUTE1, OBJECT1) from “generated-att-obj”, Value must be (ATTRIBUTE2, OBJECT2) from “gt-att-obj”. It should not contain any (ATTRIBUTE1, OBJECT1) or (ATTRIBUTE2, OBJECT2) from the “broader-concept” dictionary. ORDER1 should be the same as ORDER2.

- The keys in “broader-concept” and “matched-att-obj” must be the same as “gt-att-obj”.

For clarity, consider these examples: [In-context examples](#)

With these examples in mind, please help me extract the broader-concept, and matched-objects from the following two objects lists.

1. gt-objects: [Input Ground Truth Attributes](#)

2. generated-objects: [Input Generated Attributes](#)

Table B.9: Prompt template for matching **attributes (people)** in image caption and reference caption.

System message

You are given a task to match (object-1 positional relation with object-2) from a ground truth dictionary and a list based on their meaning.

Prompt

Inputs:

1. “gt-relations”: A dictionary of ground truth relations. Each key is a number with no meaning of order. Each key represents different relations. The values is a list of one or two relations, if there are two relations, they are synonyms. Sometimes in one relation it contains for example “image / table”, it means either image or table in this phrase is correct.
2. “generated-relations”: A list of generated relations from a model.

Matching Criteria:

- For each relation in “generated-relations”, find the corresponding relation in “gt-relations” based on their meaning, if there is none, skip it.
- If you find a match, add it to the “matched-relations” dictionary. Note that if there are two relations in a item of “gt-relations”, it means the same meaning of the relation, you can pick either one of them as the match to the relation in “generated-relations”.
- If you find that the generated relation is a broader concept of a relation in “gt-relations” such as the generated relation is near each other, next to, in touch etc. but the gt-relation specifically have their relation is specifically left, right, behind or front, etc, which is more than near, add it to the “broader-concept” dictionary.

Output:

1. A “broader-concept” dictionary: only if an relation from “generated-relations” denotes a broader category of a concept in “gt-relations” Notify that Key must be the item from “generated-relations”, Value must be item from “gt-relation”. If none, it should be an empty dictionary.
2. A “matched-relations” dictionary: only if an relation from “generated-relations” can be mapped to an relation in “gt-relations” with the matching criteria. Key must be word from “generated-relations”, Value must be word from “gt-relations”. It should not contain any words from the “broader-concept” dictionary.

For clarity, consider these examples: {In-context examples}

With these examples in mind, please help me extract the broader-concept, and matched-relations from the following two inputs.

1. gt-relations: {Input Ground Truth Relations}
2. generated-relations: {Input Generated Relations}

Table B.10: Prompt template for matching **positional relations** in image caption and reference caption. {In-context examples} are in-context examples. {Input Ground Truth Relations} are the ground truth relation list {Input Generated Relations} are the extracted relation list from the extraction step which are originally captions generated by evaluated models.

System message

You are given a task to match the correct objects with the same meaning from a ground truth dictionary and a generated dictionary.

Prompt

Inputs:

1. “gt-objects”: A dictionary of ground truth objects. Each key is a number starting rank No.1 and increment each time by 1. Each value is the corresponding object with the rank. Sometimes one object can be, for example “ground / court”, it means either ground or court is correct and should be matched.
2. “generated-objects”: A dictionary with rank being the key and the object being the value. The rank is the rank of the object in the generated caption.

Matching Criteria:

- For each object in “generated-objects”, find the object in the “gt-objects” that have the same meaning and add it to the “matched-objects” dictionary.
- By the same meaning, we mean the words can be synonyms, can be plural/singular forms of each other and can also have different length of words to express the same meaning of objects, etc.
- Notice that the final matched-objects must follow the order of values in “generated-objects”.
- If you find that the “generated-objects” can be matched with the “gt-objects” but the object in “generated-objects” is a broader concept of the objects in “gt-objects”, for example, one object in “generated-objects” is “person”, but the “gt-objects” don’t have “person” but specifically have “man”, which is a subcategory of “person”, add it to the ‘broader-concept’ dictionary instead of the “matched-objects”.

Output:

1. A “broader-concept” dictionary: only if an object from “generated-objects” denotes a broader category of a concept in “gt-objects” Notify that Key must be the item from “generated-objects”, Value must be item from “gt-objects”. If none, it should be an empty dictionary.
2. A “matched-objects” dictionary: only if an object from “generated-objects” can be mapped to an object in “gt-objects” with the matching criteria. Key must be word from “generated-objects”, Value must be word from “gt-objects”. It should not contain any words from the “broader-concept” dictionary.

For clarity, consider these examples: {In-context examples}

With these examples in mind, please help me extract the broader-concept, and matched-relations from the following two inputs.

- gt-relations: {Input Ground Truth Relations}
- generated-relations: {Input Generated Relations}


Table B.11: Prompt template for matching **comparative relations** in image caption and reference caption.


[Input Image]



[Input Instruction] Write a detailed description of the image. Provide information about all objects in front and background.

[Model Response]

InstructBLIP  In the image, a young girl is blow-drying her hair with a pink hair dryer. She is smiling and appears to be enjoying the process. The scene takes place in a bathroom, with a sink and a **toilet** visible in the background. There are several **bottles** scattered around the room, including one near the sink and another close to the toilet. Additionally, a **towel** can be seen hanging on the left side of the room.

GPT4-Vision  In the image, there is a young girl with wet hair who appears to be sitting down, smiling towards the **camera**. The girl has a pink hair dryer in her hand, which she is holding up as if displaying it. ... Above the sink, to the right, there is what looks like a roll of **paper towels**. Next to the paper towels, there's a red object that might be a **soap dish** containing a bar of **soap**. The wall in the background has a neutral color, possibly a shade of beige or light brown. ... Additionally, there's a portion of someone's arm, clothed in a pink garment, reaching towards the girl's hair with a black **comb**, likely in the act of helping to style or dry her hair. ...


LLaVA-v1.5  The image features a young girl standing in a bathroom, holding a pink hair dryer in her hand. She is smiling and appears to be enjoying the experience. Another person is present in the background, possibly assisting or observing the girl. In the bathroom, there is a sink located near the left side of the room. A **bottle** can be seen on the left side of the sink, and a **bowl** is placed on the countertop. The overall atmosphere of the scene is cheerful and lighthearted, with the girl having fun while using the hair dryer.

Figure B.1: Object existence evaluation example from three representative models in our benchmark VALOR-BENCH. Text in red indicating models' hallucinations.

[Input Image]




[Input Instruction] Describe the relationship between all the objects in the image in detail, using left, right, top, and bottom etc, from the view of the observer.


[Model Response]

InstructBLIP 

From the view of the observer, the relationship between the objects in the image can be described as follows: **1. On the left side of the image, there is a white plate with black stripes. 2. On the right side of the image, there is a wine glass. 3. At the top of the image, there is a knife. 4. At the bottom of the image, there is a mouse.**

GPT4-Vision 

In the image, we see a wooden table serving as the surface for the objects. ... On the right side of the plate, parallel to the knife and directly opposite it, there is a computer mouse with its **cord extending towards the top right corner** of the image and out of view. Above the plate (towards the top of the image), a clear glass, which appears empty, is placed **upside down with its open end facing downward and the base pointing towards the sky**. These objects seem to mimic the arrangement of a dining setting, with the mouse substituting for what might typically be a drinking cup positioned to the right side of a plate and the knife on the left side ...

LLaVA-v1.5 

In the image, there is a wooden dining table with a white plate on top of it. The plate is placed in the center of the table, and it has a mouse and a glass of water on it. The mouse is positioned to the right of the plate, while the **glass of water is located to the left of the plate**. Additionally, there is a fork on the left side of the plate and **a knife on the right side**. The fork is placed horizontally, with its handle facing upwards, while the knife is positioned vertically, **with its blade facing downwards**. The arrangement of these objects suggests that the table is set for a meal, with the mouse and glass of water possibly serving as decorative or functional elements.

Figure B.2: Positional relation evaluation example from three representative models in our benchmark VALOR-BENCH. Text in red indicating models' hallucinations.

[Input Image]



[Input Instruction] Rank the size of all the objects in the image in detail, from large to small.


[Model Response]

InstructBLIP 

From largest to smallest, the objects in the image include a cutting board, a knife, a banana, and cheese slices. The cutting board is the largest object, followed by the **knife, the banana**, and the cheese slices.

GPT4-Vision 

Starting from the largest object to the smallest, the order would roughly be as follows: 1. The cutting board – It occupies the most space in the image. 2. The knife with the light-colored handle – This knife lies diagonally across the cutting board and appears to be quite long. 3. **The piece of cheese** – Although it's not the length of the knife, it is thicker and occupies a more three-dimensional space. 4. **The banana peel** – It's spread out next to the cutting board, but it's not as large as the other objects mentioned before. 5. The knife with the red handle ... 6. The slices/sticks of cheese

LLaVA-v1.5 

1. **Banana**: The banana is the largest object in the image, taking up a significant portion of the cutting board. 2. **Cutting board**: The cutting board is the second largest object, providing a surface for the banana and other items. 3. Knife: The knife is the third largest object, with a blade that is longer than the banana. 4. Cheese: The cheese is the fourth largest object, occupying a smaller area than the knife. 5. Potato: The potato is the smallest object in the image, located near the banana and knife.

Figure B.3: Comparative relation evaluation example from three representative models in our benchmark VALOR-BENCH. Text in red indicating models' hallucinations.

Appendix C

Additional Results from Chapter 4

C.1 MRAG-BENCH Details

C.1.1 Dataset Curation Details

Dataset collection of transformative aspect We chose to manually scrape images from the web based on the definitions of the transformative aspect. To construct the image corpus, we employed Bing Image Search for each of the image object keyword predefined by us. We filtered some of the search results where the image objects do not have a clear pair of query image and ground-truth image example, around 74% keyword names were kept during this process. Here we listed all the keywords that are already filtered and used for search of query image except in biological scenario, it's for search of ground-truth image example. Each search keyword is composed of an "image object" and a "condition". For example, "A young kitten image of Himalayan Cat", here Himalayan Cat is the image object and a young kitten is the condition. For each of keyword listed below, we searched again for its ground-truth examples (except for biological scenario, it's for query images), in which only "image object" is kept and "condition" is removed. All searched results are further picked and downloaded by humans to ensure quality. Here is a list of the filtered keywords for transformative aspect:

Transformative: Temporal

- A young kitten image of Himalayan Cat
- A young kitten image of Chartreux
- A young kitten image of Burmese
- A young kitten image of Turkish Van
- A young kitten image of American Shorthair
- A young kitten image of British Shorthair
- A young kitten image of Maine Coon
- A young kitten image of Burma (Myanmar)
- A young kitten image of Selkirk Rex
- A young kitten image of Siberian
- A young kitten image of Persian
- A young kitten image of Manx
- A young kitten image of Ocicat
- A young kitten image of Russian Blue
- A young kitten image of Bengal Cat
- A young kitten image of Devon Rex
- A young kitten image of American Bobtail
- A young kitten image of Balinese
- A young kitten image of LaPerm
- A young kitten image of Egyptian Mau
- A young kitten image of Japanese Bobtail
- A young kitten image of Ragdoll
- A young kitten image of Abyssinian
- A young kitten image of American Wirehair
- A young kitten image of Oriental Shorthair
- A young kitten image of Cornish Rex

- A young kitten image of Kurilian Bobtail
- A young kitten image of Singapura Cat
- A young kitten image of Birman
- A young kitten image of Burmilla
- A young kitten image of Korat
- A young kitten image of Tonkinese
- A young kitten image of Somali Cat
- A young kitten image of Norwegian Forest Cat
- A young kitten image of Turkish Angora
- A young kitten image of Siamese
- A picture of Sainte-Chapelle under construction
- A picture of Washington Monument under construction
- A picture of Hearst Castle under construction
- A picture of Time Square under construction
- A picture of Wrigley Building under construction
- A picture of Eiffel Tower under construction
- A picture of The Arc de Triomphe under construction
- A picture of Golden Gate Bridge under construction
- A picture of White House under construction
- A picture of Palace of Versailles under construction
- A picture of Opéra Garnier under construction
- A picture of San Simeon under construction
- A picture of The Louvre under construction
- A picture of Cathédrale Notre-Dame de Paris under construction
- A picture of Sacré-Cœur Basilica under construction
- A picture of Brooklyn Bridge under construction
- A picture of Panthéon under construction

- A picture of Capitol Building under construction
- A picture of Independence Hall under construction
- A picture of Mont Saint-Michel under construction
- A picture of St Patrick's Cathedral under construction
- A picture of Space Needle under construction
- A picture of Château de Chambord under construction
- A picture of Versailles under construction

Transformative: Deformation

- An image of Toyota Camry damaged
- An image of Ford F-150 damaged
- An image of Ferrari 458 damaged
- An image of Audi Q5 damaged
- An image of Lamborghini LP640 damaged
- An image of McLaren 675LT damaged
- An image of Mercedes SLC damaged
- An image of Lamborghini Aventador damaged
- An image of Lamborghini LP570 damaged
- An image of Porsche 911 GT3 RS damaged
- An image of Audi A6 damaged
- An image of Audi A4 damaged
- An image of Lamborghini Aventador SV damaged
- An image of GMC Sierra 2500 HD damaged
- An image of Infiniti G37 damaged
- An image of GMC Yukon damaged
- An image of Honda Accord damaged
- An image of Infiniti FX35 damaged

- An image of Tesla Model 3 damaged
- An image of Acura RDX 2020 damaged
- An image of BMW 7 Series damaged
- An image of Audi A5 Sportback damaged
- An image of Hyundai IX35 damaged
- An image of Cadillac XTS damaged
- An image of BMW M3 damaged
- An image of Acura MDX damaged
- An image of Audi A3 damaged
- An image of BMW X3 damaged
- An image of Porsche Boxster damaged
- An image of Mercedes CLA45 AMG damaged
- An image of Jaguar XJ damaged

Transformative: Incomplete

- MacBook Keyboard missing keys
- Windows Keyboard missing keys
- Laptop Keyboards (Generic) missing keys
- Mechanical Keyboard missing keys
- Ergonomic Keyboard missing keys
- Compact Keyboard missing keys
- Gaming Keyboard missing keys
- Chiclet Keyboard missing keys
- Tenkeyless (TKL) Keyboard missing keys
- Virtual Keyboard (On-screen) missing keys
- Numeric Keypad missing keys
- ISO Keyboard Layout missing keys

- ANSI Keyboard Layout missing keys
- Ortholinear Keyboard missing keys
- Bluetooth/Wireless Keyboard missing keys

Transformative: Biological

- An image of Lime after oxidation
- An image of breadfruit after oxidation
- An image of dragonfruit after oxidation
- An image of starfruit after oxidation
- An image of Raspberry after oxidation
- An image of Zucchini after oxidation
- An image of Pear after oxidation
- An image of passionfruit after oxidation
- An image of Blackberry after oxidation
- An image of durian after oxidation
- An image of persimmon after oxidation
- An image of Apple after oxidation
- An image of bell pepper after oxidation
- An image of olive after oxidation
- An image of Mango after oxidation
- An image of nectarine after oxidation
- An image of tomato after oxidation
- An image of quince after oxidation
- An image of coconut after oxidation
- An image of soursop after oxidation
- An image of Kiwi after oxidation
- An image of cucumber after oxidation

- An image of apricot after oxidation
- An image of Honeydew after oxidation
- An image of Peach after oxidation
- An image of pomegranate after oxidation
- An image of carrot after oxidation
- An image of fig after oxidation
- An image of Papaya after oxidation
- An image of Blueberry after oxidation
- An image of Banana after oxidation
- An image of jackfruit after oxidation
- An image of Lemon after oxidation
- An image of tamarind after oxidation
- An image of lychee after oxidation
- An image of Pineapple after oxidation
- An image of Cantaloupe after oxidation
- An image of Orange after oxidation
- An image of Rambutan after oxidation
- An image of guava after oxidation
- An image of sweet potato after oxidation
- An image of Plum after oxidation
- An image of Avocado after oxidation
- An image of Watermelon after oxidation
- An image of potato after oxidation
- An image of Grapefruit after oxidation
- An image of Grapes after oxidation
- An image of pumpkin after oxidation
- An image of Cherry after oxidation

- An image of Strawberry after oxidation
- An image of custard apple after oxidation

Quality control We employ two types of quality control throughout the annotation process: an automatic check with predefined rules and a manual examination of each instance. The automatic check verifies correct MCQA format in which each question should only have one correct answer, metadata values, assesses image validity (checking the accessibility of each image) and filters out redundant images in the corpus (images that are repetitively downloaded). The manual examination is conducted by two experts in this field, who checked the correspondence between query images and ground-truth image examples, and filtered or revised ambiguous questions and uncorrelated query image and ground-truth images.

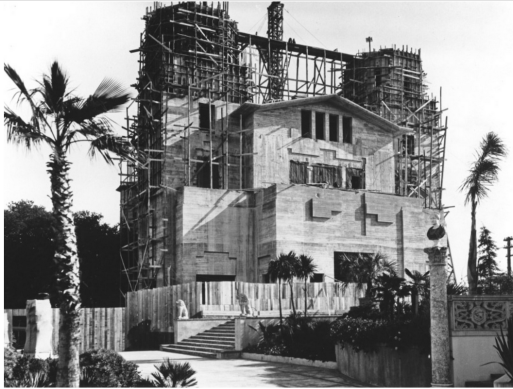
C.1.2 Human Evaluation Protocol

Three human annotators in domain conducted the human evaluation. The interface for human evaluation without RAG knowledge and with RAG knowledge are shown in Figure C.1 and Figure C.2.

C.2 Experiment Setting Details

C.2.1 Model Prompts

Following Lu et al. [2024b] and Liu et al. [2023a] our prompt consists of four parts, the instruction, question, options, and a prefix of the answer. For images, we insert them into the text to form a coherent prompt as the image placeholder (`{Image}`) indicated below. The complete prompt is as follows:



2. What is this building after its complete of construction?

- A: Biltmore Estate
- B: Vizcaya Museum and Gardens
- C: The Breakers
- D: Hearst Castle



3. What are the missing keys?

- A: *
- B: +
- C: x
- D: 8

Figure C.1: Human evaluation interface without RAG examples

Model Prompts for No RAG Evaluation

Instruction: Answer with the option's letter from the given choices directly.

{Image}

Question: {QUESTION}

Choices:

(A) {OPTION_A}

(B) {OPTION_B}

(C) {OPTION_C}

(D) {OPTION_D}

Answer:

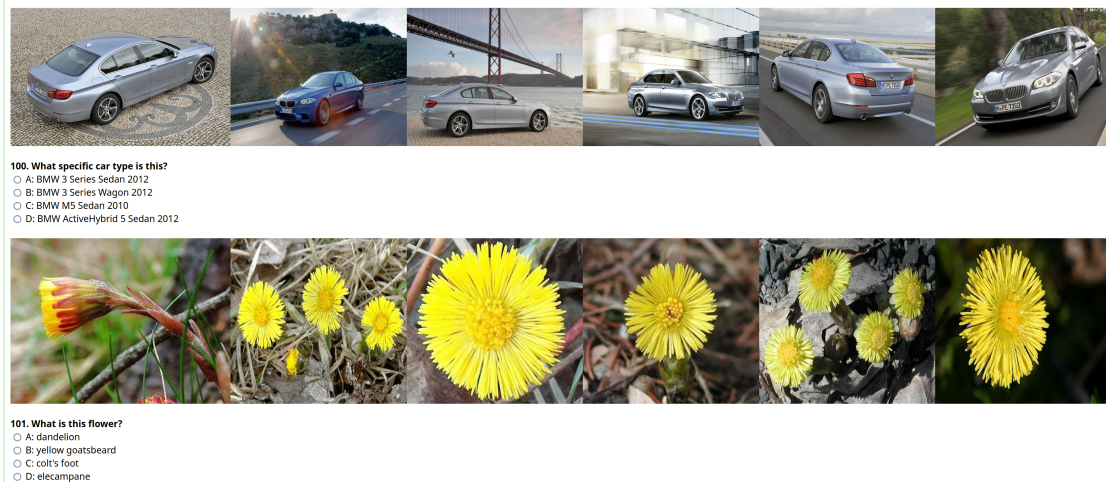


Figure C.2: Human evaluation interface with ground-truth RAG examples

Model Prompts for RAG Evaluation

Instruction: You will be given one question concerning several images. The first image is the input image, others are retrieved examples to help you. Answer with the option's letter from the given choices directly.

{Image} {Image} {Image} {Image} {Image} {Image}

Question: {QUESTION}

Choices:

(A) {OPTION_A}

(B) {OPTION_B}

(C) {OPTION_C}

(D) {OPTION_D}

Answer:

C.2.2 Evaluation Tool

Following Lu et al. [2024b], we first use a rule-based automatic tool to extract the exact answer. First, the tool detects if a valid option index appears in the model output. If no direct answer is found, the tool matches the output to the content of each option. If there is still no match, we employ GPT-3.5-turbo to automatically extract the answer following our prompts in Table C.1. If

GPT-3.5-turbo finds there is still no match, we will randomly select an option as the answer.

<p>Prompt</p> <p>Please read the following example. Then extract the multiple choice letter with the answer corresponding to the choice list from the model response and type it at the end of the prompt. You should only output either A, B, C, or D.</p> <p>{In-context examples}</p> <p>Question: {QUESTION} Choice List: (A) {OPTION_A} (B) {OPTION_B} (C) {OPTION_C} (D) {OPTION_D} Model Response: {Response} Extracted answer:</p>

Table C.1: Prompt template to extract multiple choice answer from model’s response. **{In-context examples}** are in-context examples.

C.3 More Results

We present the Recall@5 scores per each scenarios on 4 multimodal retrievers as shown in Table C.2 and LLaVA-Next-Interleave’s accuracy score affected by these retrievers in Table C.3.

Model	Overall	Perspective				Transformative				Others
		Angle	Partial	Scope	Occlusion	Temporal	Deformation	Incomplete	Biological	
MagicLens	37.03	41.61	33.33	36.27	36.11	12.75	10.78	79.41	29.41	56.67
E5-V	54.92	49.69	48.78	61.76	66.67	38.93	22.55	73.53	71.57	82.50
VISTA	59.65	66.15	67.48	64.71	63.89	38.26	8.82	33.33	94.12	80.83
CLIP	60.46	70.19	54.47	71.57	73.15	44.30	31.37	67.65	40.2	81.67

Table C.2: Recall@5 scores with 4 retriever models on MRAG-BENCH.

Model	Overall	Perspective				Transformative				Others
		Angle	Partial	Scope	Occlusion	Temporal	Deformation	Incomplete	Biological	
MagicLens	35.18	34.78	29.67	30.39	34.26	40.94	36.27	27.45	49.02	39.17
E5-V	40.06	38.82	39.84	41.18	46.3	38.93	41.18	27.45	48.04	41.67
VISTA	42.42	40.37	35.77	40.2	52.78	45.64	42.16	36.27	50.98	48.33
CLIP	40.35	40.06	33.33	39.22	56.48	43.62	44.12	27.45	36.27	49.17

Table C.3: LLaVA-Next-Interleave accuracy scores on MRAG-BENCH with 4 different retrievers.

Bibliography

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Anthropic. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>, 2024.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL <https://doi.org/10.5281/zenodo.7733589>.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding,

- localization, text reading, and beyond. *ArXiv preprint*, abs/2308.12966, 2023. URL <https://arxiv.org/abs/2308.12966>.
- Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms, 2024. URL <https://arxiv.org/abs/2404.15406>.
- Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16474–16483. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01600. URL <https://doi.org/10.1109/CVPR52688.2022.01600>.
- 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*, 2023a. URL <https://arxiv.org/abs/2310.09478>.
- Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.
- 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 Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore, 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.925. URL <https://aclanthology.org/2023.emnlp-main.925>.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *ArXiv preprint*, abs/2404.16821, 2024. URL <https://arxiv.org/abs/2404.16821>.

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 opensource chatbot impressing gpt-4 with 90% chatgpt quality. *ArXiv preprint*, 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.

Chenheng Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *ArXiv preprint*, 2024. URL <https://arxiv.org/abs/2401.16420>.

Rongyao Fang, Chengqi Duan, Kun Wang, Hao Li, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Hongsheng Li, and Xihui Liu. Puma: Empowering unified mllm with multi-granular visual generation, 2024. URL <https://arxiv.org/abs/2410.13861>.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *ArXiv preprint*, abs/2310.14566, 2023. URL <https://arxiv.org/abs/2310.14566>.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7879–7886, Apr. 2020. doi: 10.1609/aaai.v34i05.6294. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6294>.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models. In *The 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024a.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*, 2024b.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence*,

- EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 2256–2264. AAAI Press, 2024c. doi: 10.1609/AAAI.V38I3.27999. URL <https://doi.org/10.1609/aaai.v38i3.27999>.
- Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. *ArXiv preprint*, 2023a. URL <https://arxiv.org/abs/2312.10160>.
- Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *ArXiv preprint*, 2024. URL <https://arxiv.org/abs/2403.12027>.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models, 2023b.
- Drew A. Hudson and Christopher D. Manning. Gqa: a new dataset for compositional question answering over real-world images. *ArXiv preprint*, 2019a. URL <https://arxiv.org/abs/1902.09506>.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019b.
- IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. <https://huggingface.co/blog/idefics>, 2023.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *ArXiv preprint*, abs/2405.01483, 2024a. URL <https://arxiv.org/abs/2405.01483>.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models, 2024b. URL <https://arxiv.org/abs/2407.12580>.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. FAITHSCORE: Evaluating hallucinations in large vision-language models. *ArXiv preprint*, 2023. URL <https://arxiv.org/abs/2311.01477>.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proc. of ICML*, 2021. URL <http://proceedings.mlr.press/v139/kim21k.html>.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013.

Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hannaneh Hajishirzi, Sham Kakade, Ali Farhadi, Prateek Jain, et al. Matformer: Nested transformer for elastic inference. *arXiv preprint arXiv:2310.07707*, 2023.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper-Conference.pdf.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G. Moreno, and Jesus Lovon. ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 2022. doi: 10.1145/3477495.3531753. URL <https://universite-paris-saclay.hal.science/hal-03650618>.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL <https://arxiv.org/abs/2408.03326>.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL <https://arxiv.org/abs/2407.07895>.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML, 2022*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of ICML, 2023a*. URL <https://arxiv.org/abs/2301.12597>.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proc. of EMNLP, 2023c*. URL <https://aclanthology.org/2023.emnlp-main.20>.

Zhaowei Li, Wei Wang, YiQing Cai, Xu Qi, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model, 2024c. URL <https://arxiv.org/abs/2408.02503>.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv preprint*, 2023a. URL <https://arxiv.org/abs/2310.03744>.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *ArXiv preprint*, abs/2307.06281, 2023c. URL <https://arxiv.org/abs/2307.06281>.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ArXiv preprint*, 2023. URL <https://arxiv.org/abs/2310.05338>.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022a.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning

of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL <https://aclanthology.org/2022.acl-long.556>.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00331. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Anan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024.

Thomas Mensink, Jasper R. R. Uijlings, Lluís Castrejón, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araújo, and Vittorio Ferrari. Encyclopedic VQA: visual questions about detailed properties of fine-grained categories. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3090–3101. IEEE, 2023. doi: 10.

1109/ICCV51070.2023.00289. URL <https://doi.org/10.1109/ICCV51070.2023.00289>.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

OpenAI. Gpt-4 technical report. 2023. URL <https://api.semanticscholar.org/CorpusID:257532815>.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. VALOR-EVAL: Holistic coverage and faithfulness evaluation of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1783–1805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.105. URL <https://aclanthology.org/2024.findings-acl.105>.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL <http://proceedings.mlr.press/v139/radford21a.html>.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16,*

2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d08b6801f24dda81199079a3371d77f9-Abstract-Datasets_and_Benchmarks.html.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proc. of EMNLP*, 2018. URL <https://aclanthology.org/D18-1437>.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015. doi: 10.1007/s11263-015-0816-y.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8876–8884, Jul. 2019. doi: 10.1609/aaai.v33i01.33018876. URL <https://ojs.aaai.org/index.php/AAAI/article/view/4915>.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodal{qa}: complex question answering over

text, tables and images. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=ee6W5UgQLa>.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Mistral AI Team. Announcing pixtral 12b. <https://mistral.ai/news/pixtral-12b/>, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *ArXiv preprint*, 2023a. URL <https://arxiv.org/abs/2302.13971>.

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*, 2023b.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov,

- Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, 2023c. URL <https://arxiv.org/abs/2307.09288>.
- Andrés Villa, Juan Carlos Le'on Alc'azar, Alvaro Soto, and Bernard Ghanem. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. *ArXiv preprint*, 2023. URL <https://arxiv.org/abs/2312.02219>.
- Junyan Wang, Yi Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Mingshi Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. Evaluation and analysis of hallucination in large vision-language models. *ArXiv preprint*, 2023a. URL <https://arxiv.org/abs/2308.15126>.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *ArXiv preprint*, 2023b. URL <https://arxiv.org/abs/2311.07397>.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv preprint*, 2024. URL <https://arxiv.org/abs/2311.03079>.
- Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information*

Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023c. URL http://papers.nips.cc/paper_files/paper/2023/hash/3255a7554605a88800f4e120b3a929e1-Abstract-Conference.html.

Yuxin Wen, Qingqing Cao, Qichen Fu, Sachin Mehta, and Mahyar Najibi. Efficient vision-language models by summarizing visual tokens into compact registers, 2024. URL <https://arxiv.org/abs/2410.14072>.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2024. URL <https://arxiv.org/abs/2409.04429>.

Bingxin Xu, Yuzhang Shang, Yunhao Ge, Qian Lou, and Yan Yan. freepruner: A training-free approach for large multimodal model acceleration, 2024. URL <https://arxiv.org/abs/2411.15446>.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos, 2024. URL <https://arxiv.org/abs/2408.10188>.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL <https://arxiv.org/abs/2408.04840>.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv preprint*, 2023a. URL <https://arxiv.org/abs/2304.14178>.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv preprint*, 2023b. URL <https://arxiv.org/abs/2311.04257>.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-Switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *ArXiv preprint*, 2023. URL <https://arxiv.org/abs/2310.01779>.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Mingwei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *The Forty-first International Conference on Machine Learning (ICML)*, page to appear, 2024a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b. URL <https://arxiv.org/abs/2406.16852>.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruva Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *ArXiv preprint*, abs/2405.18415, 2024c. URL <https://arxiv.org/abs/2405.18415>.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval, 2024. URL <https://arxiv.org/abs/2406.04292>.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit

Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. 2023. URL <https://arxiv.org/pdf/2310.00754>.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.