SURf: Teaching Large Vision-Language Models to Selectively Utilize Retrieved Information

Jiashuo Sun[1*], **Jihai Zhang**[2], **Yucheng Zhou**[3], **Zhaochen Su**[4], **Xiaoye Qu**[5†], **Yu Cheng**[2†]
[1] Xiamen University, [2] The Chinese University of Hong Kong
[3] SKL-IOTSC, CIS, University of Macau, [4] Soochow University
[5] Shanghai AI Laboratory

Abstract

Large Vision-Language Models (LVLMs) have become pivotal at the intersection of computer vision and natural language processing. However, the full potential of LVLMs' Retrieval-Augmented Generation (RAG) capabilities remains underutilized. Existing works either focus solely on the text modality or are limited to specific tasks. Moreover, most LVLMs struggle to selectively utilize retrieved information and are sensitive to irrelevant or misleading references. To address these challenges, we propose a self-refinement framework designed to teach LVLMs to **S**electively **U**tilize **R**etrieved In**f**ormation (SURf). Specifically, when given questions that are incorrectly answered by the LVLM backbone, we obtain references that help correct the answers (positive references) and those that do not (negative references). We then fine-tune the LVLM backbone using a combination of these positive and negative references. Our experiments across three tasks and seven datasets demonstrate that our framework significantly enhances LVLMs' ability to effectively utilize retrieved multimodal references and improves their robustness against irrelevant or misleading information. The source code is available at `https://github.com/GasolSun36/SURf`.

1 Introduction

Large Vision-Language Models (LVLMs) have become crucial at the intersection of computer vision and natural language processing (NLP), empowering various applications by generating contextually relevant textual descriptions from visual inputs (Liu et al., 2023b; gpt, 2023; Dai et al., 2023; Bai et al., 2023; Ye et al., 2023; Zhu et al., 2023; Fan et al., 2024; Sun et al., 2024). These models capture and translate complex visual patterns into coherent linguistic representations. The development of LVLMs is driven by continuous improvements in model architecture, training methodologies, and data diversity (Wang et al., 2024b,a; Yu et al., 2023; Qu et al., 2024b), resulting in better performance and broader applicability.

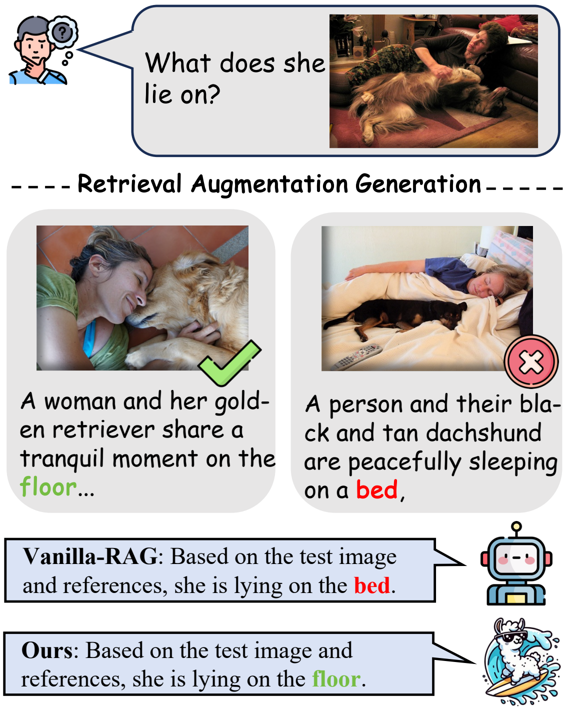

What does she lie on?
Retrieval Augmentation Generation
A woman and her golden retriever share a tranquil moment on the floor...
A person and their black and tan dachshund are peacefully sleeping on a bed,
Vanilla-RAG: Based on the test image and references, she is lying on the bed.
Ours: Based on the test image and references, she is lying on the floor.

Figure 1: Illustration of multimodal RAG. RAG can introduce misleading content, causing LVLMs to generate incorrect responses. SURf can selectively utilize information from images and descriptions, e.g., the first image-caption pair.).

Although LVLMs excel in visual language representation, they struggle with image generalization and understanding (Qu et al., 2024c). Similarly, LLMs face these challenges in the NLP domain but can mitigate them by incorporating additional knowledge or references through Retrieval-Augmented Generation (RAG), ensuring high trustworthiness (Karpukhin et al., 2020; Asai et al., 2023; Xu et al., 2023). However, in LVLMs,

[*] Work done during internship at Shanghai AI Laboratory.
[†] Both are corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7611–7629
November 12-16, 2024 ©2024 Association for Computational Linguistics

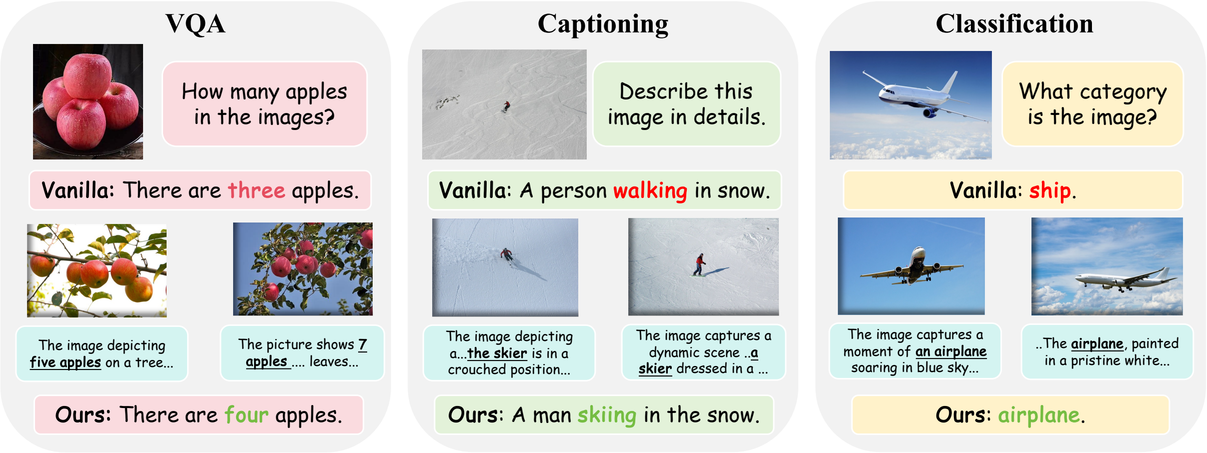

VQA
How many apples in the images?
Vanilla: There are three apples.
The image depicting five apples on a tree...
The picture shows 7 apples .... leaves...
Ours: There are four apples.
Captioning
Describe this image in details.
Vanilla: A person walking in snow.
The image depicting a...the skier is in a crouched position...
The image captures a dynamic scene ..a skier dressed in a ...
Ours: A man skiing in the snow.
Classification
What category is the image?
Vanilla: ship.
The image captures a moment of an airplane soaring in blue sky...
..The airplane, painted in a pristine white...
Ours: airplane.

Figure 2: The illustration of Multimodal RAG for VQA, Captioning and Classification Tasks. Providing images similar to the test images along with their descriptions as references can help LVLMs answer questions more accurately.

the full potential of RAG remains under-explored. Firstly, many previous multimodal RAG-related works have only focused on the text modality (Ramos et al., 2023c,a,b), without fully utilizing the LVLMs' understanding of visual content. Secondly, the few works that integrate multimodal references are often limited to specific tasks like image captioning, ignoring the broader potential of applying RAG technology (Yang et al., 2023b; Yasunaga et al., 2023). Finally, a significant issue overlooked by existing research is the potential irrelevance or even disruptive nature of retrieved content in practical applications. Under this circumstance, vanilla LVLMs fail to dynamically select retrieval content, but treat them indiscriminately, leading to a performance decline (Lin et al., 2023b; Qu et al., 2024a).

In this paper, we propose a self-refinement framework that enables LVLMs to selectively utilize the retrieved information from both image and text sources while effectively enhancing the model's robustness against irrelevant or misleading content. Specifically, we identify the visual questions that are wrongly answered by LVLM and use image-caption pairs to prompt the LVLMs to generate responses. Secondly, we assess the contribution of the introduced image-caption pairs by invoking external evaluation tools, thereby constructing a training dataset with positive and negative samples. Subsequently, we build a RAG instruction dataset to further train the LVLMs, allowing them to better benefit from RAG tasks and improve their robustness against irrelevant retrieval content. It is worth noting that we only reconstruct data from the SFT phase of the LVLMs without using any additional new datasets. We extensively evaluate our method across seven datasets and benchmarks in three different tasks: VQA, image captioning, and image classification. The experimental results demonstrate that our approach can further enhance the RAG capabilities of existing LVLMs and significantly improve their robustness in generating responses when faced with irrelevant images or content.

Our contributions are summarized as follows: (1) We empirically demonstrate that integrating Multimodal RAG with LVLMs can improve model performance, while also revealing that current LVLMs are highly sensitive to irrelevant and misleading retrieval information, which presents a significant challenge. (2) We design a lightweight and cost-effective self-refinement framework specifically aimed at teaching LVLMs to selectively utilize relevant information. (3) Through extensive experiments and evaluations, we show that our approach enhances the models' ability to effectively utilize retrieval information, making them more robust against irrelevant and misleading references.

2 Robust Multimodal RAG

2.1 Preliminaries

The RAG consists of two main components: a retriever and a generator. The retriever fetches relevant information from a large document collection, and the generator uses the retrieved document to produce the final output. We can represent the functioning of the RAG in LVLMs with the following formulas:

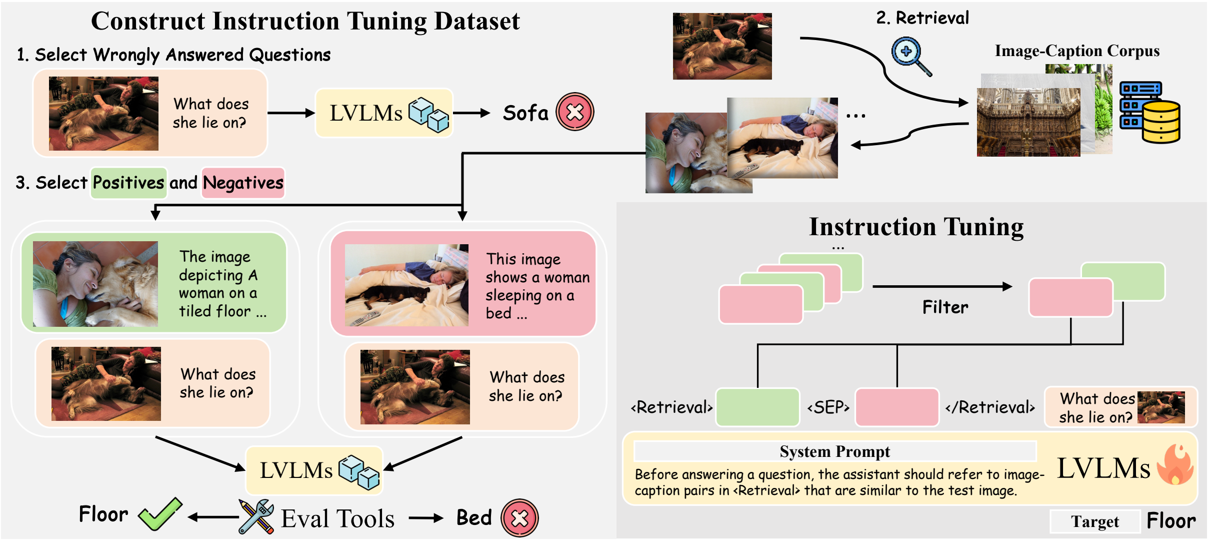

Construct Instruction Tuning Dataset
1. Select Wrongly Answered Questions
What does she lie on?
LVLMs
Sofa
2. Retrieval
Image-Caption Corpus
...
3. Select Positives and Negatives
The image depicting A woman on a tiled floor ...
What does she lie on?
This image shows a woman sleeping on a bed ...
What does she lie on?
LVLMs
Floor
Eval Tools
Bed
Instruction Tuning
...
Filter
<Retrieval>
<SEP>
</Retrieval>
What does she lie on?
System Prompt
Before answering a question, the assistant should refer to image-caption pairs in <Retrieval> that are similar to the test image.
LVLMs
Target
Floor

Figure 3: Illustration of our training framework. First, we collect questions that LVLMs initially answered incorrectly. Next, we retrieve the Top-N image-caption pairs from the corpus, allowing the LVLM to reattempt the questions. We then evaluate the answers to see if they have improved (positive) or worsened (negative). After that, we filter for the highest-quality training data and use it for instruction tuning to train the LVLMs.

Given an input x (e.g., a question q or instruction with a feature vector of an image i_{test}), the retriever fetches k relevant images $\{i_1, i_2, \ldots, i_k\}$ from an image set of image-caption collection $\mathbb{D}$. The probability distribution of the retriever can be represented as $\bar{P}(i \mid x)$. The generator uses the retrieved images $\{i_1, i_2, \ldots, i_k\}$, the corresponding captions $\{c_1, c_2, \ldots, c_m\}$ and the input x to generate the output y (e.g., an answer, image caption, or classification label). The conditional probability distribution of the generator can be represented as:

$$P(y \mid x, \{[i_1, c_1], [i_2, c_2] \ldots, [i_k, c_k]\} \qquad (1)$$

The final output of the LVLM with RAG is based on the joint probability of the input x and the set of retrieved image-caption pairs $\{[i_1, c_1], [i_2, c_2], \ldots, [i_k, c_k]\}$:

$$P(y \mid x) = \sum_{j=1}^{k} P(y \mid x, r_j) \bar{P}(i_j \mid x) \qquad (2)$$

where $r_i = [i_j, c_j]$ represents the retrieved image-caption pair, with c_j being the caption corresponding to the retrieved image i_j.

2.2 Multimodal RAG Benefit LVLMs

RAG has been proven to improve model performance on downstream tasks while maintaining a high level of trustworthiness in the field of NLP (Karpukhin et al., 2020; Asai et al., 2023; Xu et al., 2023; Jin et al., 2024). However, in LVLMs, the full potential of RAG remains under-explored.

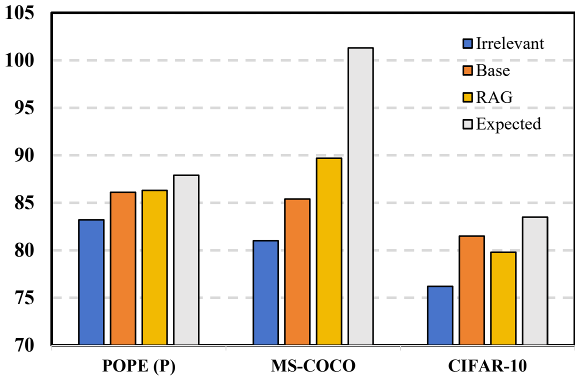


Figure 4: Performance of the base model (LLaVA-1.5-7B) without using RAG (Base), RAG with irrelevant content (Irrelevant), and RAG on POPE-popular, MS-COCO, and CIFAR-10.

As shown in Figure 2, when addressing tasks such as VQA, captioning, and classification, we can enhance the model performance by retrieving relevant images and their corresponding descriptions to provide a pattern mapping for the input x. The collection of pattern state is denoted as $\mathbb{M} = \{M_0, M_1, \cdots, M_n\}$, and $M_i \sim ([I, T] \in \mathbb{M})$, where I and T denote the image and description, respectively. Next, our goal is to learn this mapping:

$$f : x \to f(x | M_{i_1}, M_{i_2}, \ldots, M_{i_k}) \qquad (3)$$

To better understand the impact of RAG on model performance, we conducted experiments comparing direct inference with RAG-enhanced inference across three datasets. Figure 4 illustrates the performance differences. It can be observed that retrieving and incorporating additional multimodal information (both image and text) significantly improves the model's performance in tasks across VQA, Image Captioning, and Classification.

2.3 Irrelevant Harms Model Performance

Typically, the retrieval process $\bar{P}(i_j \mid x)$ or $\bar{P}(c_j \mid x)$ is typically implemented by computing image-to-image or image-to-text similarity in CLIP embedding space (Ramos et al., 2023c,a). However, this retrieval process is not always reliable, leading to the inclusion of irrelevant or misleading references. For example in Figure 1, the similarity scores returned the Top-2 images most similar to the test image. Nevertheless, these two images contribute differently to the original question. The latter image misleads the model and causes incorrect responses.

Figure 4 demonstrates the impact of irrelevant information on RAG. It can be seen that the performance of RAG is even worse than without introducing any additional information, which indicates the negative impact of irrelevant or disturbing information on current LVLMs. We believe that the RAG of current LVLMs still has significant potential. If we can teach the model to selectively utilize the retrieved information and ignore the irrelevant or misleading ones, the performance of RAG in LVLMs will be further improved, potentially approaching the results shown by the gray bars.

2.4 Robust RAG Training Framework

Since RAG has great potential to help improve the accuracy of model generation, and regardless of how the retriever is optimized, achieving perfect retrieval recall is unattainable (Radford et al., 2021; Cherti et al., 2023). Therefore, we choose to optimize $P(y \mid x, r_i)$, through teaching the model to learn to selectively utilize the retrieved information. We propose a self-refinement framework that enables LVLMs to selectively refer to relevant information from both image and text sources while effectively enhancing the model's robustness against irrelevant or misleading content.

2.4.1 Construction of Positive and Negative Examples

Introducing both relevant and irrelevant content during training can enhance the model's ability to distinguish and select relevant information (Lin et al., 2023b). Therefore, at this stage, we construct positive and negative examples (denoted as $\mathbb{C}_{\text{pos}}$ and $\mathbb{C}_{\text{neg}}$) for subsequent robust training.

We hypothesize that if the model initially answers a question incorrectly but can answer correctly after including an example (both image and description), that example contains useful information (positive). Otherwise, the example is considered misleading or irrelevant (negative). Specifically, we first collect the data used by the LVLM during the SFT stage and use a fixed-parameter LVLM to answer questions based on images, recording incorrect examples. Then, we perform retrieval from the image-caption corpus to obtain the Top-N images and their corresponding descriptions. These are then appended to the test image and question, allowing the LVLM to answer the question again. We use specific evaluation tools to determine whether the answer has improved, remained unchanged, or worsened. Image-caption pairs that successfully improve the answer are considered positive examples of the current question, while those that do not cause any change or worsen the answer are considered negative examples of the current question.

Notably, the data we construct is sourced from the examples in the existing LVLM training data used during the instruction fine-tuning stage, requiring no new external data.

2.4.2 Data Filtering

Due to the token length limitation in LVLMs, we need to further filter the positive and negative examples obtained in the previous step. We exclude examples from the Top-N image-caption pairs that contain only positive or negative examples. For positive examples, we select the image with the highest similarity to the test image to ensure the inclusion of highly relevant information and to avoid model training collapse:

$$p_{\text{pos}} \sim \max_{i_j \in \mathbb{C}_{\text{pos}}} p_V^{\theta}(x, i_j) \tag{4}$$

For negative examples, we choose the image with the highest similarity to the test image as hard negatives. These hard negatives are more similar to the positive examples, thus requiring the model

	VQA					Captioning		Classification		Avg.
	POPE (R)	POPE (P)	POPE (A)	MMstar	Vizwiz^V	MS-COCO	Vizwiz^C	CIFAR-10	EmoSet	
7B Parameter Model										
Zero-shot	87.3	86.1	84.2	30.3	50.0	198.6	134.5	81.5	52.8	89.48
Vanilla-RAG	87.9	86.3	83.3	32.1	48.3	178.1	169.6	79.7	50.4	90.63
Rerank-RAG	88.3	86.3	83.4	31.4	49.3	210.0	164.2	80.9	50.5	93.81
Filter-RAG	88.5	86.6	**83.9**	31.8	51.5	231.1	172.0	82.2	51.8	97.71
SURf	**89.8**	**87.9**	83.6	**33.5**	**54.3**	**238.4**	**177.4**	**83.5**	**53.1**	**100.17**
13B Parameter Model										
Zero-shot	87.1	86.2	84.5	32.8	53.6	210.0	150.2	82.6	56.4	93.71
Vanilla-RAG	88.3	86.4	83.4	33.1	50.2	218.5	160.9	80.7	55.6	95.23
Rerank-RAG	88.4	86.4	83.6	33.5	50.9	223.1	162.1	82.0	56.0	96.22
Filter-RAG	88.6	86.5	83.8	33.7	51.7	226.7	164.1	83.2	56.5	97.20
SURf	**89.5**	**87.7**	**84.6**	**34.5**	**54.6**	**250.9**	**177.5**	**85.1**	**58.1**	**102.50**

Table 1: Performance comparison of our model on 7B and 13B parameters using four methods across seven tasks. In POPE, (R), (P), and (A) stand for Random, Popular, and Adversarial subsets, respectively (applies to all tables below.). Vizwiz^V and Vizwiz^C represents VQA and captioning based on Vizwiz. The best performance in the table is highlighted in **bold**.

	Para.	Shots	**POPE (R)**		**POPE (P)**		**POPE (A)**		**MS-COCO**
			Acc.	F1	Acc.	F1	Acc.	F1	CIDEr ↑
Flamingo (Alayrac et al., 2022)	9B	4-shots	-	-	-	-	-	-	93.1
OpenFlamingo (Awadalla et al., 2023)	9B	4-shots	48.5	48.1	49.5	49.0	48.9	48.5	89.0
Otter (Li et al., 2023a)	9B	4-shots	82.5	81.8	74.7	73.9	69.9	69.4	92.2
MMICL (Zhao et al., 2023)	12.1B	4-shots	87.3	86.6	82.7	82.1	81.0	80.7	95.7
SURf	7B	2-shots	**89.8**	**89.3**	**87.9**	**87.6**	**83.6**	**83.9**	**101.3**

Table 2: Performance of our 7B model compared to four ICL models on the three POPE subsets (VQA) and MS-COCO (captioning). The results of the ICL models are directly from the original paper.

to develop higher discriminative capabilities to accurately identify them:

$$p_{\text{neg}} \sim \max_{i_j \in \mathbb{C}_{\text{neg}}} p_V^{\theta}(x, i_j) \tag{5}$$

2.4.3 RAG Instruction-Tuning

Using the high-quality positive and negative example pairs generated through the above process, we fine-tune the existing model with RAG instructions. The retrieved images and their corresponding descriptions are concatenated sequentially before the test image, enclosed by special characters `<Retrieval>` and `</Retrieval>`. This ensures that the model can effectively distinguish between retrieved-context and the actual test input, enhancing its ability to leverage relevant information while minimizing the impact of irrelevant or misleading data.

The algorithm of our method is shown in the Appendix Algorithm 1.

3 Experiment

3.1 Datasets

We evaluated our model using seven datasets across three distinct tasks: VQA: POPE (Li et al., 2023c), MMStar (Chen et al., 2024), Vizwiz-VQA (Chen et al., 2022), Image Captioning: MS-COCO (Lin et al., 2014), Vizwiz-Caption (Gurari et al., 2020), Image Classification: CIFAR-10 (Krizhevsky, 2009), EmoSet (Yang et al., 2023a). For more detailed information and metrics can be found in the Appendix A.2.

3.2 Baselines

We compared four methods among LlaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2023b):

Zero-shot Directly prompting LVLMs to generate responses.

Vanilla-RAG Concatenating the Top-N image-caption pairs from the database, which have the highest CLIP score similarity to the test image before the questions and images for the LVLMs to respond.

Rerank-RAG Building on Vanilla-RAG, we prompt the LLM to generate a caption for the test

	POPE (R)		POPE (P)		POPE (A)				MS-COCO			CIFAR-10
	Acc.	F1	Acc.	F1	Acc.	F1	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Acc.
Zero-shot	87.3	86.0	86.1	84.9	84.2	83.4	22.3	28.0	50.9	75.3	22.1	81.5
Vanilla-RAG	**87.9**	**86.5**	**86.3**	**85.0**	**83.3**	**82.5**	**24.5**	**28.3**	**51.4**	**79.8**	**22.4**	**79.7**
w/ 1k	87.5	86.3	85.4	84.2	82.2	81.3	22.5	27.8	50.5	75.0	21.8	79.5
w/ 5k	87.4	86.2	85.3	84.1	82.2	81.3	22.2	27.7	50.4	75.0	21.6	77.1
w/ 10k	87.2	86.0	85.0	83.8	82.1	81.2	22.0	27.4	50.1	75.4	21.2	76.7
w/ 100k	87.0	85.9	84.9	83.7	82.0	81.1	22.1	27.3	50.3	74.8	21.5	76.4
w/ 1,000k	86.7	85.0	84.5	83.1	81.8	80.8	22.0	27.5	49.9	73.6	21.2	75.6
Ours	**89.8**	**89.3**	**87.9**	**87.6**	**83.6**	**83.9**	**27.9**	**29.9**	**55.1**	**101.3**	**24.2**	**83.5**
w/ 1k	88.9	88.3	87.8	87.3	83.1	83.0	26.8	29.4	54.2	97.3	23.7	83.1
w/ 5k	89.3	88.7	87.8	87.3	83.1	83.4	26.5	29.1	53.6	97.4	23.5	82.4
w/ 10k	89.4	88.8	87.6	87.3	83.2	83.5	26.9	29.4	54.2	97.7	23.8	83.4
w/ 100k	89.1	88.5	87.7	87.2	83.3	83.4	26.6	29.2	53.9	96.5	23.6	80.5
w/ 1,000k	89.2	88.7	87.9	87.4	83.6	83.6	27.1	29.4	54.3	98.4	23.7	80.9

Table 3: Performance comparison of our model and vanilla-RAG on three tasks when introducing irrelevant image-caption pairs. "1k to 1,000k" indicates the range of similarity between the introduced images and the test images, with larger values indicating less relevance.

	POPE (R)		POPE (P)		POPE (A)				MS-COCO			CIFAR-10
	Acc.	F1	Acc.	F1	Acc.	F1	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Acc.
Vanilla-RAG	87.9	86.5	86.3	85.0	83.3	82.5	24.9	28.5	52.8	89.7	22.6	79.5
w/ Switch	87.2	86.0	85.7	84.6	82.4	82.0	22.2	28.0	50.8	75.1	22.0	78.4
Ours	89.8	89.3	87.9	87.6	83.6	83.9	27.9	29.9	55.1	101.3	24.2	83.5
w/ Switch	89.6	89.1	87.9	87.6	83.6	83.8	26.8	29.3	54.1	97.1	23.7	83.4

Table 4: Performance comparison of our model and vanilla-RAG on three tasks in the random switching of retrieved content setting.

image. We then calculate the BERT-Score between this caption and the descriptions of the retrieved images, ranking the image-caption pairs with higher relevance scores at the top.

Filter-RAG Enhancing Rerank-RAG by removing any image-caption pairs with a similarity score less than S.

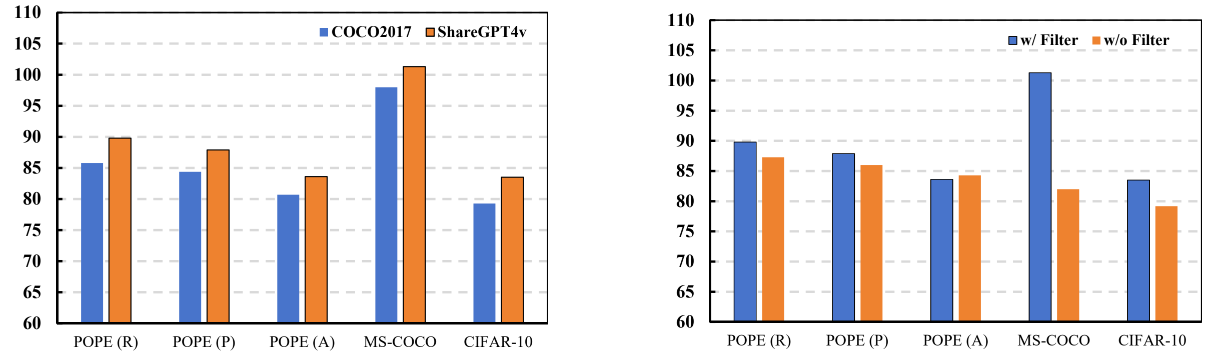
Additionally, we compared four In-Context Learning (ICL) models: Flamingo (Alayrac et al., 2022), OpenFlamingo (Awadalla et al., 2023), Otter (Li et al., 2023a), and MMICL (Zhao et al., 2023). For all approaches, we used greedy decoding as the decoding strategy.

3.3 Implementation details

We collected 60,000 initial incorrect responses from LVLMs and generated 10,000 samples with positive and negative sample pairs. After filtering, we refined this to 2,000 samples for the final training data. We use LLaVA-1.5 as the LVLM backbone of our model SURf-7B and SURf-13B and use CLIP (ViT-L with a resolution of 336*336) (Radford et al., 2021) as the vision encoder. Our 7B and 13B models are further trained from the instruction-finetuned LLaVA-1.5-7B and LLaVA-1.5-13B models following previous works (Lin et al., 2024, 2023a; Li et al., 2023b; Liu et al., 2023c) since LLaVA is the most popular used LVLMs. We use 8 A100-80G to training 1 hour for 2 epochs. For the VQA and image captioning task, we use exact match and Bert-Score (Zhang et al., 2020) as the evaluation tool respectively, mentioned in Section 2.4.2. We use ShareGPT4v-PT (Chen et al., 2023) as our database for RAG, which includes approximately 1,246k image-caption pairs with an average caption length of 826. For the retrieval system, we use FAISS (Johnson et al., 2021) with flat indexes to pre-index the computed embeddings of all images in the database.

3.4 Experimental Results

Compare to Baselines Table 1 presents the comparison of our model, trained with our method, against four other methods. On the VQA task, our model significantly outperforms previous methods, with a VQA accuracy improvement of approximately 3.7% for the 7B model compared to zero-shot and 2.3% compared to Filter-RAG, achieving



(a) Performance of our model using different retrieval sources.

(b) Performance of our model on downstream tasks with and without data filtering.

Figure 5: Ablation Study of Database Size and Data Filter.

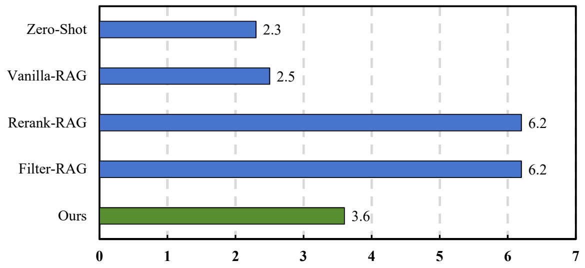


Figure 6: Efficiency analysis of our model compared to four methods. We report the running times per sample.

state-of-the-art results. Furthermore, on the captioning task, the improvement of our model is even more pronounced (detailed results can be found in the Appendix).

In contrast, for the classification task, vanilla-RAG may perform worse than direct inference. However, our training method enables the model to selectively refer to the retrieved content, resulting in a final performance that significantly surpasses zero-shot. For the 13B model, the improvement in captioning is even more significant, with an approximate 34.1% increase compared to zero-shot. Additionally, the table illustrates that simple methods, such as reranking and filtering, cannot effectively address the problem of irrelevant content introduced by retrieval.

Compare to ICL models The experiments in Table 2 compare our model with various ICL models, as ICL models are very similar to ours at the input level. Despite having fewer parameters and exemplars (For the ICL models, more shots correspond to better performance, we used their 4-shot results for comparison since they only reported results for 4-shot or 32-shot scenarios.) in the prompts compared to the other models, our model achieves the best performance on both the POPE and MS-COCO datasets. Specifically, it improves the average accuracy by 3.4% and the F1 score by 3.8% on POPE compared to the second-best model. This demonstrates that our model can effectively utilize the retrieved content to enhance the performance of downstream tasks.

Robustness Tables 3 and 4 present the results of our robustness tests. In Table 3, we maintain the image-caption pair with the highest CLIP similarity score among the retrieved content to ensure effective information. We also introduce image-caption pairs from the Top-K (from 1k to 1,000k) positions as forced irrelevant information. The results show that the performance of vanilla-RAG significantly declines on the three datasets as more irrelevant image-caption pairs are introduced. In contrast, our model's performance remains very stable. Notably, the model's performance when introducing 100k and 1000k irrelevant image-caption pairs is better than when introducing 1k pairs. This improvement is because, after training with hard negative samples, our model can easily distinguish content unrelated to the test image and question, thereby focusing more on other relevant information in the retrieval.

Table 4 shows that our model remains robust even after randomly shuffling the examples, whereas vanilla-RAG exhibits a significant decline in performance. This demonstrates that training the model with our proposed framework enables it to selectively extract relevant information from the retrieved content, making it less sensitive to the order of the examples.

3.5 Ablation Study

Size of the Database We conducted experiments using different databases as retrieval sources, with results shown in Figure 5(a). It can be seen that using COCO-2017 (approximately 118k image-caption pairs) and ShareGPT-4V (approximately 1,246k image-caption pairs) results in notable differences in model performance for VQA and classification tasks, while the metrics for the captioning task show minimal differences. The reason is that VQA and classification tasks are more challenging for the model compared to captioning, requiring a larger retrieval source to provide more diverse reference image-caption pairs.

Data Filter Figure 5(b) presents the results with and without using the data filtering step. It can be seen that the performance of the model trained without data filtering is significantly worse compared to when data filtering is used. This highlights the importance of filtering positive and negative samples and training with hard negative sampling in our training framework.

Efficiency Analysis We compared the efficiency of our method with four other methods, as shown in Figure 6. We calculate the average running time for 1,000 samples in image captioning tasks with the max token length set to 256. It can be observed that our method increases the time by approximately 1.3 seconds per sample compared to the zero-shot approach. This increase is primarily due to the time required to convert the image to an embedding, retrieval time, and the additional overhead introduced by the increased length of the prompt. However, this slight increase in time is acceptable considering the performance improvement.

In contrast, Rerank-RAG and Filter-RAG are slower because they require additional prompts for the LVLMs to generate captions for the current image, which are then used for text similarity comparisons.

4 Related Work

4.1 Large Vision-Language Models

Large vision-language models (LVLMs) have greatly benefited from advancements in large language models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Su et al., 2024a,c), which integrate a vision encoder with a language model backbone. Leveraging the success of LLMs through pre-training and instruction tuning (Liu et al., 2023b; Ye et al., 2023; Zhu et al., 2023; Bai et al., 2023; Dai et al., 2023), LVLMs like LLaVA (Liu et al., 2023b) employ GPT-4V[1] to generate diverse instruction datasets, thereby enhancing their capacity to understand images and follow human instructions. Despite these successes, current MLLMs still face significant challenges with hallucinations (Su et al., 2022; Li et al., 2023d; Liu et al., 2023a; Wang et al., 2024a; Zhao et al., 2024; Leng et al., 2023; Zhang et al., 2024; Zhou et al., 2024). These issues often result from misalignment between the vision and language components, leading to neglecting image details and generating incorrect content. Our work aims to improve LVLMs' ability to selectively reference retrieved information and enhance robustness against misleading content.

4.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) has become a powerful approach in natural language processing, combining the strengths of retrieval-based methods and generative models (Merth et al., 2024; Asai et al., 2023; Xu et al., 2023; Lin et al., 2023b; Su et al., 2024b). In the NLP domain, RAG aims to select the most relevant documents from a large corpus using techniques such as BM25 (Robertson and Zaragoza, 2009) and neural retrievers like DPR (Karpukhin et al., 2020). However, the challenge in the multimodal domain is considerably higher, as the retrieval dimension encompasses images along with text. Previous works (Yang et al., 2023b; Ramos et al., 2023c; Yasunaga et al., 2023; Ramos et al., 2023a,b; Xia et al., 2024c,a,b) have shown that retrieving similar images based on a test image and using their corresponding captions can enhance model performance on captioning tasks. Nevertheless, these methods often fail to address how to manage irrelevant image-caption pairs, which can decrease model accuracy. Our work focuses on improving LVLMs' ability to selectively reference pertinent retrieved information and increase robustness against misleading content, thereby enhancing performance across various downstream tasks.

5 Conclusion

This paper introduces a robust self-refinement multimodal RAG training framework designed for LVLMs. Our approach incorporates retrieval information for initially incorrect answers, filtering in beneficial positive examples and excluding

[1] https://openai.com/research/gpt-4v-system-card

detrimental negative ones. We implement a hard-negative sampling strategy to preserve the training data of the highest quality and employ RAG-based instruction fine-tuning. Experimental results across seven datasets spanning three different tasks show that our method significantly enhances the capability of LVLMs to effectively utilize multimodal retrieval information, while also improving their resilience against misleading content.

6 Limitation

Our method mainly has three limitations:

- Our retrieval approach heavily depends on large-scale, high-quality data sources. While using only the training data as the data source is a feasible solution, the performance is slightly inferior compared to large-scale data sources in complex tasks. Future work should explore how to leverage small sample data sources for inference through retrieval.

- Despite our method having been extensively evaluated on tasks such as Visual Question Answering (VQA), image captioning, and image classification, its generalization to other visual tasks, such as image generation and image segmentation, remains unexplored. Future work should investigate the adaptability of our framework to a broader range of tasks.

- Given that the retrieval process currently supports a maximum of three image-caption pairs due to lengthy descriptions, future optimizations could include using shorter captions, employing methods to compress descriptions, or increasing the maximum input tokens for LVLMs. These improvements would enable more image-caption pairs to be included, enhancing the accuracy of downstream tasks of LVLMs.

References

2023. Gpt-4v(ision) system card.

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. 2022. Flamingo: a visual language model for few-shot learning. In *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*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19076–19085. IEEE.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2818–2829. IEEE.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi.

2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Lizhou Fan, Wenyue Hua, Xiang Li, Kaijie Zhu, Mingyu Jin, Lingyao Li, Haoyang Ling, Jinkui Chi, Jindong Wang, Xin Ma, et al. 2024. Nphardeval4v: A dynamic reasoning benchmark of multimodal large language models. *arXiv preprint arXiv:2403.01777*.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 417–434. Springer.

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2023b. RA-DIT: retrieval-augmented dual instruction tuning. *CoRR*, abs/2310.01352.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2023c. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv:2311.05437*.

Thomas Merth, Qichen Fu, Mohammad Rastegari, and Mahyar Najibi. 2024. Superposition prompting: Improving and accelerating retrieval-augmented generation. *Preprint*, arXiv:2404.06910.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. 2024a. Alleviating hallucination in large vision-language models with active retrieval augmentation. *arXiv preprint arXiv:2408.00555*.

Xiaoye Qu, Mingyang Song, Wei Wei, Jianfeng Dong, and Yu Cheng. 2024b. Mitigating multilingual hallucination in large vision-language models. *arXiv preprint arXiv:2408.00550*.

Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. 2024c. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. *arXiv preprint arXiv:2408.17150*.

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. 2021. Learning transferable visual models from natural language supervision. In *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.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023a. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3648–3663. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023b. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1635–1651. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023c. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2840–2849. IEEE.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. 2024a. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Juntao Li, Lijun Wu, and Min Zhang. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change. *arXiv preprint arXiv:2210.17127*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024b. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024c. Timo: Towards better temporal reasoning for language models. *arXiv preprint arXiv:2406.14192*.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. 2024. Visual agents as fast and slow thinkers. *arXiv preprint arXiv:2408.08862*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024b. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. 2024a. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Preprint*, arXiv:2406.06007.

Peng Xia, Ming Hu, Feilong Tang, Wenxue Li, Wenhao Zheng, Lie Ju, Peibo Duan, Huaxiu Yao, and Zongyuan Ge. 2024b. Generalizing to unseen domains in diabetic retinopathy with disentangled representations. *Preprint*, arXiv:2406.06384.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024c. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. RECOMP: improving retrieval-augmented lms with compression and selective augmentation. *CoRR*, abs/2310.04408.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023a. Emoset: A large-scale visual emotion dataset with rich attributes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20326–20337. IEEE.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Mohammad Shoeybi, Ming-Yu Liu, Yuke Zhu, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. 2023b. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11844–11857. Association for Computational Linguistics.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. 2024. Avoiding feature suppression in contrastive learning: Learning what has not been learned before. *arXiv preprint arXiv:2402.11816*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: empowering vision-language model with multi-modal in-context learning. *CoRR*, abs/2309.07915.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Appendix

A.1 Algorithm

Algorithm 1 Robust RAG Training Framework

```
Require: Input question $q$ and image $i_{\text{test}}$, Image-Caption collection $\mathbb{D}$, Evaluate Tools $T$, Vision Encoder $p_V^{\theta}$, LVLMs $M_\theta$, SFT data collection $C$, Training data set $S$, Positive set $\mathbb{C}_{\text{pos}}$, Negative set $\mathbb{C}_{\text{neg}}$
$S \leftarrow []$
for each instruction $x$ in $C$ do
    response $\leftarrow M_\theta(x)$
    state $\leftarrow T$(response)
    if not state then
        $S \leftarrow S \cup \{x\}$
    end if
end for
for each instruction $x$ in $S$ do
    $[i, c] \leftarrow$ Retrieval from $\mathbb{D}$ with query $i_{\text{test}}$
    response $\leftarrow M_\theta(x, [i, c])$
    state $\leftarrow T$(response)
    if state then
        $C_{\text{pos}} \leftarrow \mathbb{C}_{\text{pos}} \cup \{(x,$
            $[\max_{i_j \in \mathbb{C}_{\text{pos}}} p_V^{\theta}(x, i_j), c_j])\}$
    else
        $C_{\text{neg}} \leftarrow \mathbb{C}_{\text{neg}} \cup \{(x,$
            $[\max_{i_j \in \mathbb{C}_{\text{neg}}} p_V^{\theta}(x, i_j), c_j])\}$
    end if
end for
$S \leftarrow [\mathbb{C}_{\text{pos}}, \mathbb{C}_{\text{neg}}]$
while $M_\theta$ has not converged do
    Update parameters of $M_\theta$ on $S$
end while
```

A.2 Data Analysis

In this section, we introduce the datasets used in our experiments. The statistics of these datasets are shown in Table 5.

POPE POPE (Li et al., 2023c) offers a method to assess object hallucination in LVLMs by querying if specific objects exist in images. The queries are balanced between existent and non-existent objects (50% each). There are three sampling settings: random, popular, and adversarial. The evaluation pivots on two key metrics: Accuracy and the F1 score.

MMStar MMStar (Chen et al., 2024) is an advanced benchmark designed to evaluate the capabilities of LVLMs across multiple dimensions. The benchmark includes 1,500 meticulously selected challenge samples. These samples are initially chosen from existing benchmarks using an automated pipeline, followed by a rigorous human review to ensure high quality.

Dataset/Benchmark	Answer Type	Test
POPE	Yes/No	9,000
MMStar	Multiple Choice	1,500
Vizwiz-VQA	Single word or Phrase	8,000
MS-COCO	Text	5,000
Vizwiz-Caption	Text	7,750
CIFAR-10	Class Name	10,000
EmoSet	Class Name	800*

Table 5: The statistics of the datasets used in this paper. * denotes we randomly selected 800 samples from EmoSet to constitute the test set.

Vizwiz-VQA Vizwiz-VQA (Chen et al., 2022) is the task of returning the answer to a question about an image. It has 8,000 test samples with the unique label "Unanswerable."

MS-COCO The MS-COCO (Lin et al., 2014) dataset is a large-scale dataset for object detection, segmentation, key-point detection, and captioning. We use this dataset only for the image captioning task.

Vizwiz-Caption VizWiz-Caption (Gurari et al., 2020) is a specialized dataset for evaluating and improving image captioning systems, particularly for visually impaired users. It consists of images taken by visually impaired individuals using their smartphones, accompanied by human-generated captions.

CIFAR-10 CIFAR-10 (Krizhevsky, 2009) is a well-known benchmark dataset primarily used for evaluating image classification algorithms. The dataset is split into 50,000 training images and 10,000 test images, divided into ten different classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

EmoSet EmoSet (Yang et al., 2023a) comprises 3.3 million images in total, with 118,102 of these images carefully labeled by human annotators, making it five times larger than the largest existing dataset. We randomly sampled 100 instances from each class to serve as our test set.

	POPE (R)		**POPE (P)**		**POPE (A)**		**MS-COCO**					**CIFAR-10**
	Acc.	F1	Acc.	F1	Acc.	F1	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Acc.
Zero-shot	87.3	86.0	86.1	84.9	84.2	83.4	22.8	28.2	51.4	85.4	22.2	81.5
								1-shot				
Vanilla-RAG	87.7	86.3	85.2	84.1	82.8	81.9	22.1	27.8	50.7	76.2	21.9	79.8
Ours	89.6	89.1	87.8	87.4	83.3	83.7	26.6	29.4	54.0	96.2	23.7	82.4
								2-shot				
Vanilla-RAG	87.9	86.5	86.3	85.0	83.3	82.5	24.9	28.5	52.8	89.7	22.6	79.5
Ours	**89.8**	**89.3**	**87.9**	**87.6**	**83.6**	**83.9**	**27.9**	**29.9**	**55.1**	**101.3**	**24.2**	**83.5**
								3-shot				
Vanilla-RAG	87.5	86.0	85.5	84.2	82.6	81.6	23.0	28.1	51.4	79.2	22.2	79.1
Ours	89.3	88.7	87.8	87.4	83.2	83.3	27.1	29.3	54.4	98.7	23.6	82.0

Table 6: Number of exemplars.

	POPE (R)		**POPE (P)**		**POPE (A)**		**MS-COCO**					**CIFAR-10**
	Acc.	F1	Acc.	F1	Acc.	F1	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Acc.
Vanilla-RAG	87.9	86.5	86.3	85.0	83.3	82.5	24.5	28.3	51.4	79.8	22.4	79.7
1k	86.6	84.8	85.8	84.0	83.2	81.6	25.3	26.7	51.0	98.1	22.7	79.2
2k	**89.8**	**89.3**	**87.9**	**87.6**	**83.6**	**83.9**	**27.9**	**29.9**	**55.1**	**101.3**	**24.2**	**83.5**
3k	88.8	87.9	87.3	86.5	83.2	82.9	27.5	29.4	53.5	99.3	23.9	81.5
4k	88.2	87.5	87.0	86.3	82.9	82.5	26.8	28.8	51.4	96.8	23.0	79.6

Table 7: Effect of training data size.

A.2.1 Metrics

Unless otherwise specified, we use exact match as the evaluation metric for VQA and classification tasks. For captioning tasks, we use BLEU-4, ROUGE-L, CIDEr, METEOR, and SPICE as evaluation metrics[2].

A.3 Additional Ablation Study and Experiment Analysis

A.3.1 Sensitivity to the Number of Examplars

Table 6 shows the performance of our model with different numbers of examples. Due to the long captions of ShareGPT-4V, only three examples can fit within a 4096 context window. Our method demonstrates robustness with 1, 2, and 3 examples, indicating adaptability to various numbers of examples. However, the performance peaks with 2 examples and declines with 1 and 3 examples. The decline with 1 example may be due to insufficient information, while 3 examples may introduce excessive irrelevant information.

A.3.2 Effect of Training Data Size

Table 7 shows the experiments on the amount of training data. Using only 2k data points, our model is already able to utilize RAG and achieve the best performance fully. Although the performance with 3k and 4k data points is slightly worse than with 2k, the results still surpass those of vanilla-RAG. This indicates that our framework can sufficiently leverage its capabilities using just 2k samples self-generated by the model.

A.3.3 Effect of Different Retrieved Content

In this section, we explore the performance differences when using image-caption pairs versus using only captions for retrieval across three tasks, as shown in Table 8. For VQA and classification tasks, using both image and caption yields the best results, as the additional information from the image is beneficial for tasks that require a strong understanding of the image. However, for the captioning task, using only captions performs better since this task requires the model to generate a relevant response based solely on the retrieved descriptions.

A.3.4 Detail Results of Captioning Tasks

In the main table, the metrics for the captioning task are the sum of BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE. We present the detailed results of our 7B and 13B models on MS-COCO and Vizwiz-Caption in Table 9.

A.3.5 Effect of Irrelevant Content

We also tested Qwen-VL (Bai et al., 2023) and mPLUG-Owl2 (Ye et al., 2023) under three settings (Base, Irrelevant, and RAG) across three tasks. The

[2] We use the official COCO evaluation toolkit.

	POPE (R)		**POPE (P)**		**POPE (A)**		**MS-COCO**					**CIFAR-10**
	Acc.	F1	Acc.	F1	Acc.	F1	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Acc.
Vanilla-RAG												
w/ image-caption	87.9	86.5	86.3	85.0	83.3	82.5	24.5	28.3	51.4	79.8	22.4	79.7
w/ caption	87.4	86.3	85.9	84.7	83.0	82.3	24.7	28.5	51.8	80.6	22.9	79.2

Table 8: Performance of Vanilla-RAG on downstream tasks with different retrieval content.

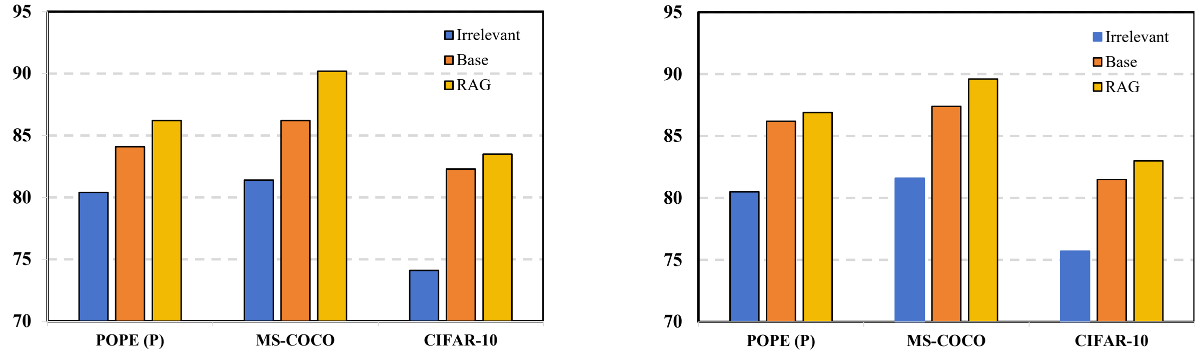
	MS-COCO					**Vizwiz-Caption**					**Avg.**
	B@4	METEOR	ROUGE-L	CIDEr	SPICE	B@4	METEOR	ROUGE-L	CIDEr	SPICE	
						7B Parameter Model					
Zero-shot	22.3	28.0	50.9	75.3	22.1	15.2	19.3	40.9	47.3	11.8	33.51
Vanilla-RAG	24.5	28.3	51.4	79.8	22.4	21.0	21.6	45.2	67.1	14.7	37.60
Rerank-RAG	24.7	28.6	52.0	82.1	22.6	20.5	20.9	44.3	64.6	13.9	37.42
Filter-RAG	26.8	29.4	54.2	97.0	23.7	21.4	21.8	45.9	68.0	14.9	40.31
Ours	**27.9**	**29.9**	**55.1**	**101.3**	**24.2**	**22.4**	**22.3**	**46.3**	**71.1**	**15.3**	**43.57**
						13B Parameter Model					
Zero-shot	22.8	28.2	51.4	85.4	22.2	18.0	19.8	42.5	57.2	12.7	36.52
Vanilla-RAG	24.9	28.5	52.8	89.7	22.6	21.6	21.3	45.4	59.4	13.2	37.94
Rerank-RAG	25.1	28.6	53.5	93.1	22.8	21.8	21.2	45.5	60.2	13.4	38.52
Filter-RAG	25.5	28.7	53.9	95.6	23.0	22.0	**21.4**	45.6	61.5	13.6	39.08
Ours	**30.8**	**29.1**	**56.0**	**111.5**	**23.5**	**24.5**	21.1	**46.2**	**71.2**	**14.5**	**44.83**

Table 9: Full results of 7B and 13B Robust-LlaVA on MS-COCO and Vizwiz-Caption. The best performance in the table is highlighted in **bold**.

results are shown in Figures 7(a) and 7(b). It can be observed that irrelevant content has a significant impact on the current LVLMs.

A.4 Case Study

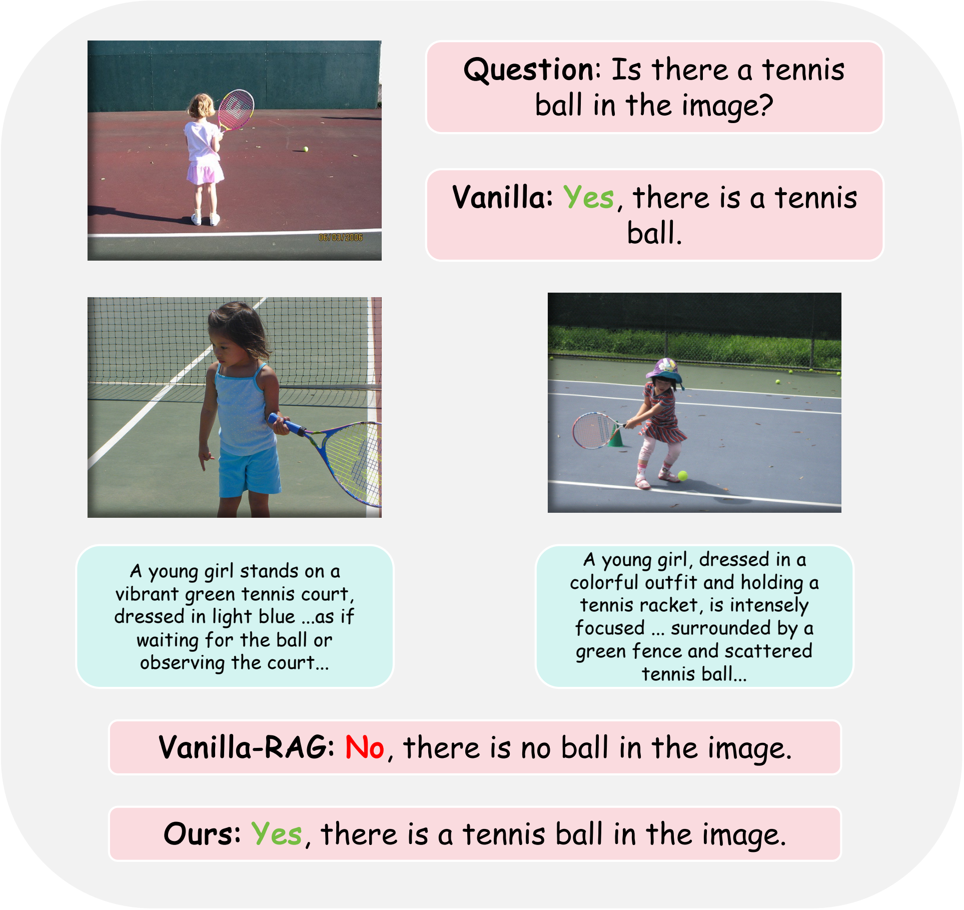
We present four examples comparing our method with zero-shot and vanilla-RAG, as shown in Figures 8, 9, 10, and 11.



(a) Performance of the base model (Qwen-VL) without using RAG (Base), RAG with irrelevant content (Irrelevant), and RAG on POPE-popular, MS-COCO, and CIFAR-10.

(b) Performance of the base model (mPLUG-Owl2) without using RAG (Base), RAG with irrelevant content (Irrelevant), and RAG on POPE-popular, MS-COCO, and CIFAR-10.

Figure 7: Ablation Study of Database Size and Data Filter.



Question: Is there a tennis ball in the image?

Vanilla: Yes, there is a tennis ball.

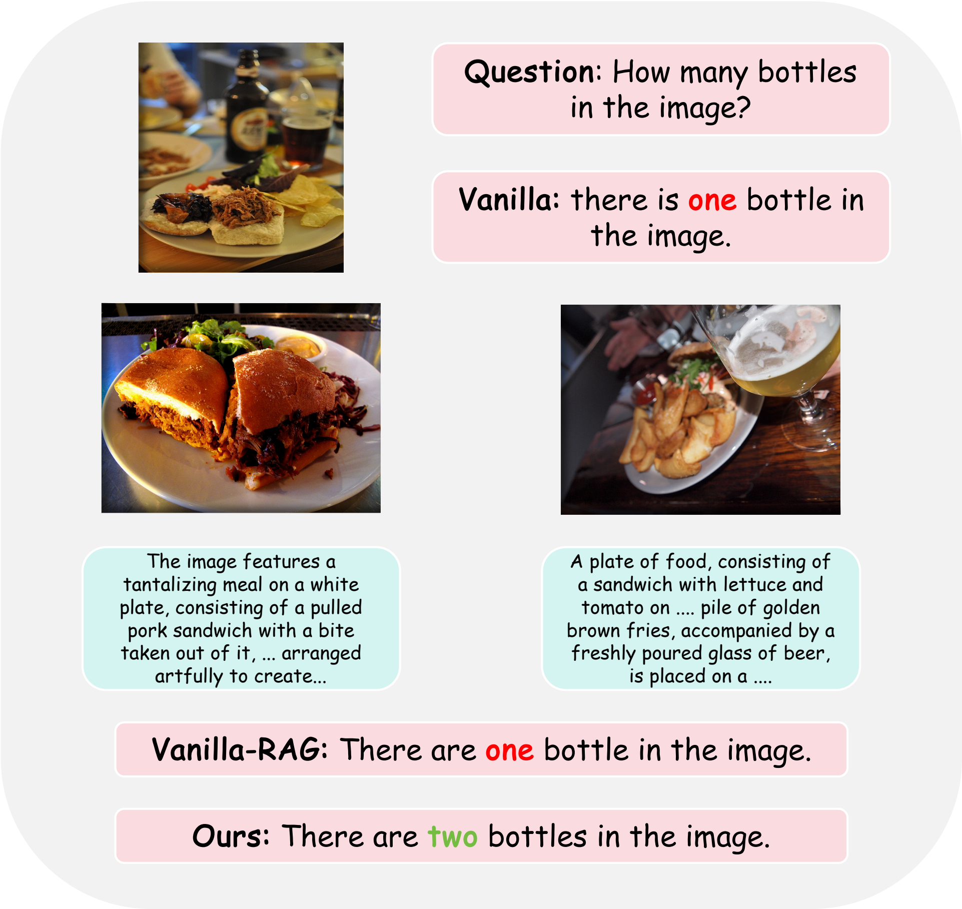
A young girl stands on a vibrant green tennis court, dressed in light blue ...as if waiting for the ball or observing the court...

A young girl, dressed in a colorful outfit and holding a tennis racket, is intensely focused ... surrounded by a green fence and scattered tennis ball...

Vanilla-RAG: No, there is no ball in the image.

Ours: Yes, there is a tennis ball in the image.

Figure 8: Case for comparing our method with zero-shot and vanilla-RAG.



Question: How many bottles in the image?

Vanilla: there is one bottle in the image.

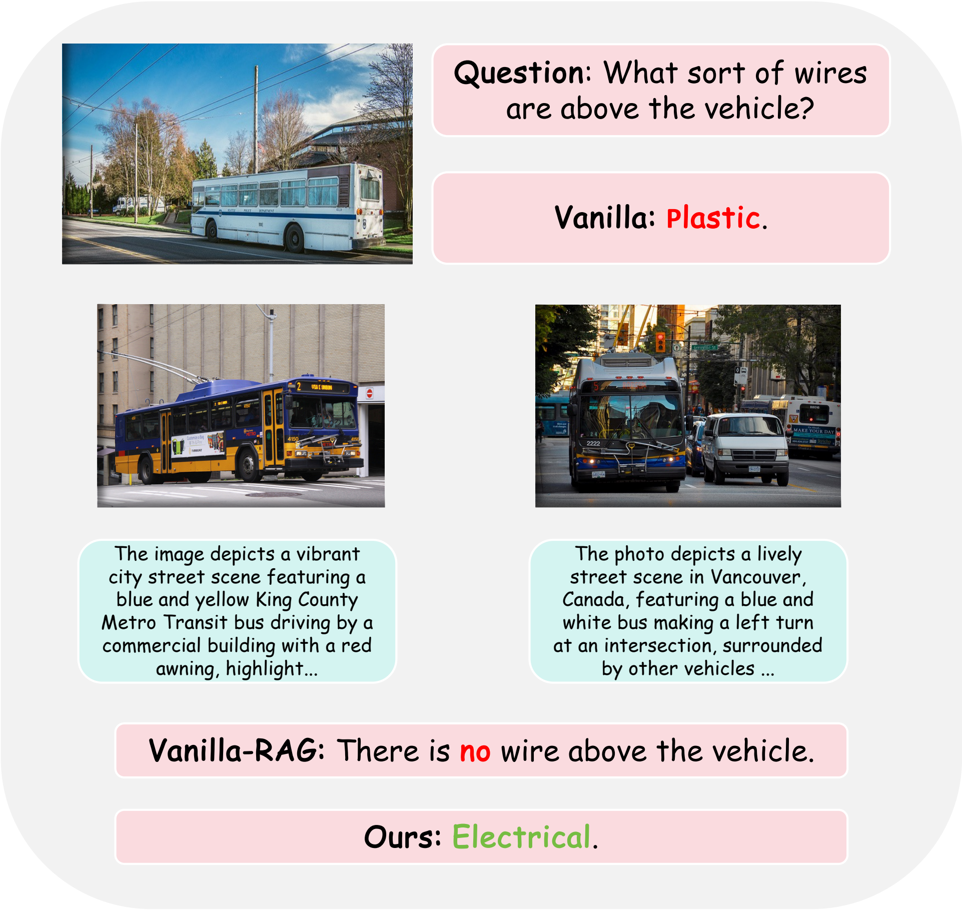
The image features a tantalizing meal on a white plate, consisting of a pulled pork sandwich with a bite taken out of it, ... arranged artfully to create...

A plate of food, consisting of a sandwich with lettuce and tomato on .... pile of golden brown fries, accompanied by a freshly poured glass of beer, is placed on a ....

Vanilla-RAG: There are one bottle in the image.

Ours: There are two bottles in the image.

Figure 9: Case for comparing our method with zero-shot and vanilla-RAG.



Question: What sort of wires are above the vehicle?

Vanilla: Plastic.

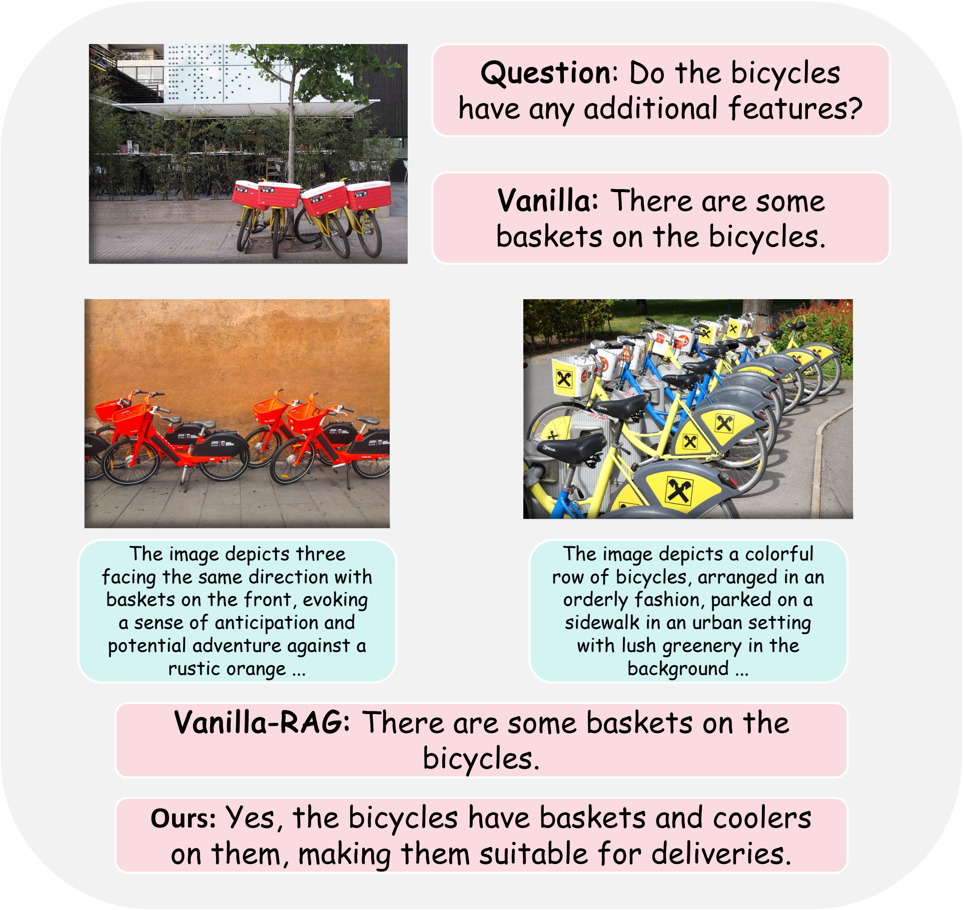
The image depicts a vibrant city street scene featuring a blue and yellow King County Metro Transit bus driving by a commercial building with a red awning, highlight...

The photo depicts a lively street scene in Vancouver, Canada, featuring a blue and white bus making a left turn at an intersection, surrounded by other vehicles ...

Vanilla-RAG: There is no wire above the vehicle.

Ours: Electrical.

Figure 10: Case for comparing our method with zero-shot and vanilla-RAG.



Question: Do the bicycles have any additional features?

Vanilla: There are some baskets on the bicycles.

The image depicts three facing the same direction with baskets on the front, evoking a sense of anticipation and potential adventure against a rustic orange ...

The image depicts a colorful row of bicycles, arranged in an orderly fashion, parked on a sidewalk in an urban setting with lush greenery in the background ...

Vanilla-RAG: There are some baskets on the bicycles.

Ours: Yes, the bicycles have baskets and coolers on them, making them suitable for deliveries.

Figure 11: Case for comparing our method with zero-shot and vanilla-RAG.