Look, Compare, Decide: Alleviating Hallucination in Large Vision-Language Models via Multi-View Multi-Path Reasoning

Xiaoye Qu[1*], **Jiashuo Sun**[2*], **Wei Wei**[1†], **Daizong Liu**[3], **Jianfeng Dong**[4], **Yu Cheng**[5]
[1] Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory,
School of Computer Science & Technology, Huazhong University of Science and Technology
[2] Xiamen University [3] Peking University
[4] Zhejiang Gongshang University [5]The Chinese University of Hong Kong
{xiaoye, weiw}@hust.edu.cn; gasolsun36@gmail.com;
dzliu@hust.edu.cn; dongjf24@gmail.com; chengyu@cse.cuhk.edu.hk

Abstract

Recently, Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in multi-modal context comprehension. However, they still suffer from hallucination problems referring to generating inconsistent outputs with the image content. To mitigate hallucinations, previous studies mainly focus on retraining LVLMs with custom datasets. Although effective, they inherently come with additional computational costs. In this paper, we propose a training-free framework, **MVP**, that aims to reduce hallucinations by making the most of the innate capabilities of the LVLMs via **M**ulti-**V**iew Multi-**P**ath Reasoning. Specifically, we first devise a multi-view information-seeking strategy to thoroughly perceive the comprehensive information in the image, which enriches the general global information captured by the original vision encoder in LVLMs. Furthermore, during the answer decoding, we propose multi-path reasoning for each information view to quantify and aggregate the certainty scores for each potential answer among multiple decoding paths and finally decide the output answer. By fully grasping the information in the image and carefully considering the certainty of the potential answers when decoding, our MVP can effectively reduce hallucinations in LVLMs. The extensive experiments verify that our proposed MVP significantly mitigates the hallucination problem across four well-known LVLMs. Furthermore, MVP is plug-and-play and can integrate with other decoding methods for more performance boosts. The source code is available at: https://github.com/GasolSun36/MVP.

1 Introduction

Large Vision-Language Models (LVLMs) have become indispensable and marked a significant milestone in the field of Artificial Intelligence. These

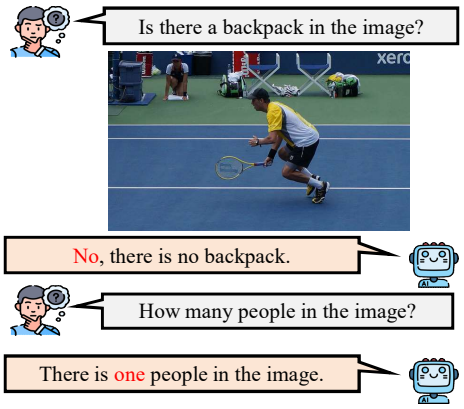

Figure 1: Given an image, LVLM fails to recognize objects or miscounts the quantity.

LVLM models, owing to their ability to generate contextually relevant textual renditions of visual inputs, are being extensively employed across a diverse spectrum of applications, such as healthcare (Liu et al., 2023b; Li et al., 2024; Bazi et al., 2023), autonomous systems (Cui et al., 2024; Tian et al., 2024; Park et al., 2024b), and robotics (Liu et al., 2024c; Shah et al., 2023; Kelly et al., 2024).

Despite substantial advancements (Sun et al., 2024; Liu et al., 2025), LVLMs suffer from a significant challenge termed "hallucination", whereby the models produce semantically plausible but factually inaccurate text, misaligned with the ground-truth content of the associated image. As shown in Figure 1, LVLMs fail to recognize "backpack" and incorrectly identify the number of people in the image. In applications where precision and reliability of generated content are paramount, such hallucinations can trigger a cascade of erroneous decisions. Consequently, addressing the hallucination issue is indispensable for strengthening the trustworthiness of LVLMs across practical applications.

To tackle hallucination, most recent studies focusing on retraining the LVLMs with constructed hallucination-related datasets by supervised fine-tuning (SFT) (Chen et al., 2023; Wang et al., 2024b;

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4428–4441
January 19–24, 2025.

Park et al., 2024a; Liu et al., 2023a), or Reinforcement Learning from Human Feedback (RLHF) (Yu et al., 2023; Yan et al., 2024; Sun et al., 2023). Although these methods for alleviating hallucination in LVLMs have shown effectiveness, they acquire a substantial number of high-quality examples for training and are quite time-consuming and labor-intensive. Recently, there are also works exploring training-free paradigms to mitigate the hallucinations. Woodpecker (Yin et al., 2023) pick out and correct hallucinations from the generated text. MARINE (Zhao et al., 2024) employs classifier-free guidance to incorporate the additional object grounding features to improve the precision of LVLMs' generations. However, most of them heavily rely on external complicated tools, such as Grounding DINO (Liu et al., 2023c), or BLIP-2-FlanT5X (Li et al., 2023a).

In this work, to alleviate hallucinations in LVLMs, we focus on maximizing the innate ability of LVLMs without introducing additional training costs or external tools. To this end, we propose a novel training-free framework MVP, namely **M**ulti-**V**iew Multi-**P**ath Reasoning. Different from previous works, our MVP is grounded in an analysis of the key factors underlying hallucination, including the incomplete comprehension of image content and low certainty when decoding answer tokens in original LVLMs. First, if the vision encoder of LVLMs can not fully capture the information from the input image, language models may generate outputs based on this incomplete content, thus resulting in hallucinatory descriptions. Second, during the answer decoding, hallucinations occur more frequently when the certainty of answer tokens is low. In this scenario, the model is uncertain about multiple candidate tokens, leading to potentially inaccurate outputs.

Thus, our MVP proposes to fully capture the information in the image and carefully consider the certainty of the potential answers when decoding. Specifically, we first devise a multi-view information-seeking strategy, which involves an exhaustive perception of the image from varying dimensions: a "top-down" look captures overarching scene context, a "regular" view addresses elementary visual information, and a "bottom-up" perspective zooms in on intricate details. Instead of using tools, the captured information from these diversified views is generated by the LVLMs, and effectively reinforces the global image context captured by the original vision encoder of LVLMs, thereby reducing the hallucinations from misunderstanding the image's information. In addition, during the answer decoding stage, we further introduce multi-path reasoning for each information view by explicitly quantifying the certainty score of the potential answers and then aggregating the overall certainty among multiple paths. Then, the answer with the highest certainty score will be chosen as the final answer, thus effectively alleviating the hallucinations caused by low certainty. To verify the effectiveness of MVP, we conduct experiments on four widely-used LVLMs. The promising results demonstrate that our framework significantly outperforms recent training-free methods.

To sum up, our contributions are summarized as:

- We propose a training-free framework to alleviate hallucinations with Multi-view Multi-path Reasoning. Our framework focuses on maximizing the innate ability of LVLMs without introducing additional training costs or external tools.
- To comprehensively grasp the image, we seek information from multi-view perspectives, including "bottom-up", "regular", and "top-down" views. During decoding, we introduce multi-path reasoning to quantify and compare the certainty of each potential answer.
- Through comprehensive experiments, we demonstrate the superior performance of our MVP in alleviating hallucinations across four LVLMs. Moreover, our framework is plug-and-play and can integrate with other decoding methods for further improvement.

2 Method

2.1 Overall of the MVP Framework

As shown in Figure 2, given that hallucinations commonly arise due to incomplete comprehension of image content, we propose to seek complementary information from the input image with three different views. Subsequently, the acquired information is leveraged to augment the global vision information from the vision encoder for LLM reasoning. For each view, considering different decoding paths have different certainty for potential answers, we introduce certainty-driven multi-path reasoning, which quantifies and aggregates the certainty score for each potential answer among multiple decoding paths. In this stage, we maximize the inherent

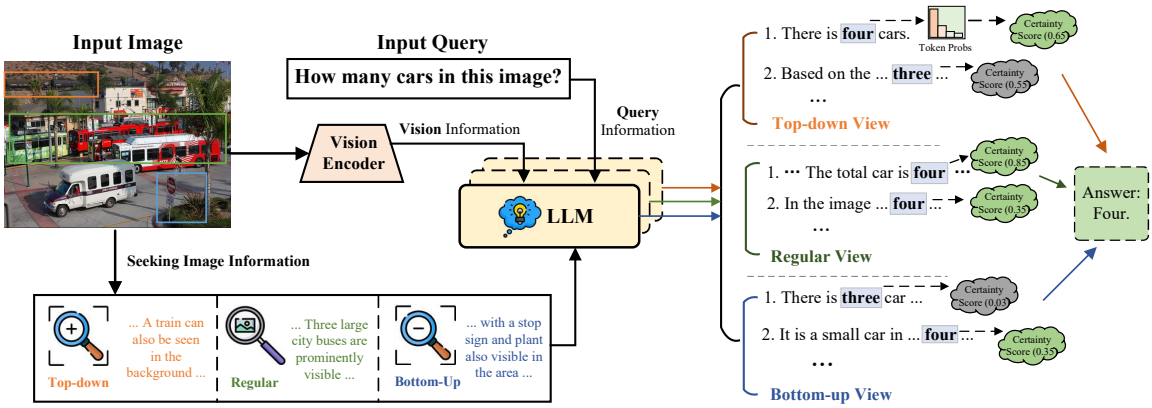

Figure 2: An overview of our Multi-View Multi-Path Reasoning. (1) Seeking image information from multiple perspectives including top-down, regular, and bottom-up views. (2) Augmenting the global vision information with each view information. (3) The certainty-driven decoding corresponding to each view quantifies and aggregates certainty scores for each potential answer among multiple decoding paths. The final results are obtained by comparing certainty scores among all candidates.

reasoning ability of the model. Finally, with the multi-view information and multi-path reasoning, we achieve superior performance for alleviating hallucinations.

2.2 LVLMs Input and Decoding

The input of LVLMs contains both image and text. The image is first processed by a vision encoder (e.g. CLIP (Radford et al., 2021), BLIP (Li et al., 2022)) to obtain visual tokens. Then, the image tokens are mapped to the input space of LLMs for decoding. We denote the visual tokens as $\mathbf{x}^v = \{x_1^v, x_2^v, \ldots, x_N^v\}$. Here N is the length of the visual tokens. Correspondingly, the input query is tokenized with the tokenizer. We denote it as $\mathbf{x}^q = \{x_1^q, x_2^q, \ldots, x_M^q\}$ with length M. The image and text tokens are concatenated as the final input sequence X with length $N + M$.

$$X = [x^v : x^q] = [x_1^v, x_2^v, \ldots, x_N^v, x_1^q, x_2^q, \ldots, x_M^q], \quad (1)$$

After feeding the input tokens X to the LVLMs, the model outputs answers in an auto-regressive manner which predicts the next token based on previous tokens, formally:

$$p(O_t|O_{<t}) = \text{SoftMax}[LVLM([\{O_i\}_{i=1}^{t-1})], \quad (2)$$

where we omit the input query X and $\{O_i\}_{i=1}^{t-1}$ are decoding tokens from the previous t-1 rounds and the first decoding token O_1 is decoded with the input X in Eq. 1. At time step t, the token with the highest probability is chosen from the vocabulary. During the decoding period, hallucinations arise when probabilities are improperly attributed to tokens that fail to correlate with the presented visual image.

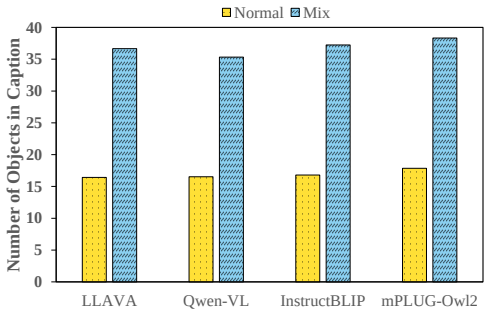

Figure 3: Comparison of the number of objects between regular and multi-view caption. The statistic is obtained in MSCOCO popular part of the POPE benchmark.

2.3 Multi-view Image Information Seeking

Previous LVLM research, utilizing a CLIP for global image representation, may neglect intricate, object-specific details and background components, consequently leading to hallucinations precipitated by a partial grasp of the input image (Zhang et al., 2024a,b). For instance, when querying the detailed information that is not captured by the video encoder, the LVLMs tend to hallucinate. Thus, it is imperative to master comprehensive information about the image before responding to the input query. Naturally, a wealth of visual information exists in images and can be located by various methods, such as invoking external visual detection tools

(Liu et al., 2023c; He et al., 2017).

In this paper, we resort to maximizing using the innate ability of the LVLMs and design three perspectives for extracting comprehensive information: "bottom-up", "regular", and "top-down". To accomplish it, we use the given LVLM to generate captions by designing dedicated prompts, thus eliminating the need for external tools or the design of specific networks. For example, to extract the information from a top-down perspective, we use the prompt: "*Given the overall scene depicted in the image, taking into account the context, environmental factors, and any relevant visual cues, describe this image in details.*" (Please see Section 4.3 for more prompts). To demonstrate the effectiveness of multi-view information-seeking strategy, we conduct a statistical analysis of the visual richness of multi-view captions. As shown in Figure 3, the LLaVA-1.5 model captures an average of 16.43 objects per image using only regular perspective caption, while an average of 36.66 objects can be recognized when three perspectives are adopted. Formally, the captions from a specific perspective can be tokenized and denoted as $\mathbf{x}^c = \{x_1^c, x_2^c, \ldots, x_K^c\}$ with length K and $c \in \{\text{Top-down}, \text{Bottom-up}, \text{Regular}\}$. Subsequently, the caption is integrated with the input for LLM decoding:

$$X' = [x_1^v, \ldots, x_N^v, x_1^c, \ldots, x_K^c, x_1^q, \ldots, x_M^q], \tag{3}$$

2.4 Multi-path Certainty-driven Reasoning

Decoding strategies are important in guiding how LVLMs produce textual answer. Previous decoding strategies commonly consider each output token with the same level of importance, thus ignoring the unique importance of the answer token. However, we observe that the answer tokens present different certainty during diverse decoding paths. In Figure 2, for the question (How many cars are in this image), the first decoding path of "Bottom-up" and "Regular" perspectives produce different answers "four" and "three" but their certainty is significantly different (0.65 and 0.03, respectively). This phenomenon indicates that hallucinations occur more frequently when the certainty of answer tokens is low and inspires us with certainty-driven reasoning to alleviate hallucinations. Formally, we quantify the *Centainty Score* S, which is the difference between the probabilities of the two tokens

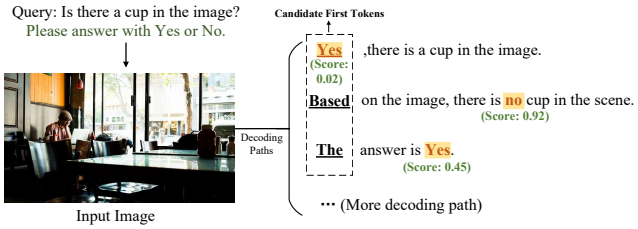

Figure 4: An illustration of certainty-driven multi-path reasoning. The correct answer is "No". "Score" denotes the certainty score of the answer token. "Yes", 'Based", "The" are candidate decoding tokens at first place. The three decoding paths are greedy decoding with these candidate tokens.

with the highest probabilities at time step t:

$$S = p(x_t^1 \mid v, x_{<t}) - p(x_t^2 \mid v, x_{<t}) \tag{4}$$

where x_t^1 and x_t^2 represent the post-softmax probabilities of top-two tokens at each decoding step t. It is worth noting that we only consider the probability disparity of the answer tokens.

2.4.1 Multi-Path Certainty-driven Reasoning

To illustrate certainty-driven reasoning, we first consider a basic situation where only one greedy decoding paths exist. As shown in Figure 5, given the input query, we observe that LVLMs tend to hallucinate when a low certainty score occurs, where greedy decoding mistakenly takes the bottle for a cup and outputs the wrong answer: "Yes, there is a cup in the image", while the certainty score of the answer token "Yes" is only 0.02. With further investigation, when decoding the first token, besides "Yes", there are many other candidates (i.e. "Based", "The"), which are displayed by underline in Figure 4 and sorted by probability from high to low. Instead of introducing complex methods to building multiple decoding paths, we simply inspect more top-K paths starting from relatively lower probability tokens, namely decoding from the second word "Based", the third word "The", and so on. Notably, the second path leads to the correct answer "no" with a significantly higher certainty score of 0.92.

Thus, we introduce a multi-path reasoning which explicitly considers the certainty of the answer tokens. Specifically, to build multiple paths, we consider the top-K candidates in the decoding process of the first token, and then continue decoding based on each candidate to generate the K paths with different answers. Formally, each path corresponds to an answer A_k. Here the answers can be identified

by the question type or specified prompt format. For instance, we can search for numbers in the output to answer the question in Figure 2, or identify "yes" or "no" in Figure 4. Then, we aggregate the certainty score for same $\hat{A}_k$ from K paths:

$$S_{\hat{A}_k} = \sum_{j=1}^{K} M_j(p(A_j^1 \mid v, A_{<t}) - p(A_j^2 \mid v, A_{<t})) \tag{5}$$

where $A_{<t}$ denotes the sequence before generating answer A_j. Here M_j is an optional parameter denoting the probability of the first token in the K-th path and we will explore it in the experiment. Thus, we can obtain the certainty score for each potential answer.

2.4.2 Multi-View Multi-Path Reasoning

In Section 2.3, we seek image information from three perspectives $c \in \{\text{Top-down}, \text{Bottom-up}, \text{Regular}\}$. Considering that each view captures different information from the input image, the corresponding reasoning paths also present specific preference, thus we can further aggregate certainty scores for multi-view multi-path:

$$S_{\hat{A}_{c,k}} = \sum_{i=1}^{c} \alpha_i \sum_{j=1}^{K} M_{ij}(p(A_{x_{ij}}^1 \mid v, A_{<t}) - p(A_{x_{ij}}^2 \mid v, A_{<t})) \tag{6}$$

where α_i is a hyperparameter denoting the importance of a specific perspective. Finally, the answer with the highest certainty score is selected as our final answer:

$$A_{final} = argmax(S_{\hat{A}_{c,k}}) \tag{7}$$

3 Experiment

3.1 Evaluation Benchmarks

Following previous works (Leng et al., 2023; Huang et al., 2023), we use the following two benchmarks POPE and MME.

POPE the Polling-based Object Probing Evaluation (Li et al., 2023b). In this benchmark, LVLMs are queried to determine whether a specific object is present in the provided image. It encompasses three distinct settings: random, popular, and adversarial, each differing in the construction of negative samples. The POPE benchmark aggregates data from three distinct sources: MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). It involves 500 images from each dataset under each sampling setting. The performance is gauged using four key metrics: Accuracy, Precision, Recall, and F1.

MME (Fu et al., 2024) acts as a comprehensive benchmark designed to evaluate LVLMs across a range of dimensions. It is composed of ten perception-related subtasks and four cognition-focused ones. In the experiments, we evaluate the full dataset. In addition, we take into account the existence and count subsets for the inspection of object-level hallucination, along with the position and color subsets for attribute-level hallucination evaluation. The combined metric of accuracy and accuracy+ is used to quantify the performance as per the official implementation.

3.2 Evaluation LVLM and Baselines

LVLMs. To comprehensively evaluate our model and have a fair comparison with previous works, we experiment with our proposed MVP on four state-of-the-art LVLMs, including LLaVA1.5, Qwen-VL, InstructBLIP, and mPLUG-Owl2. All four LVLMs are based on 7B LLM backbone models.

Baselines. To verify the effectiveness of our framework, we compare MVP with the vanilla LVLMs and two recent training-free methods, including VCD and OPERA. In our main experiments, for fair comparison, vanilla, VCD, and our MVP all adopt the decoding strategy of direct sampling. In addition, OPERA introduces a penalty term on the model logits during the beam-search decoding to mitigate the over-trust issue.

3.3 Experiment Results

Results on POPE. Table 1 summarizes the experimental results on the MSCOCO part of POPE benchmark, including experiments under random, popular, and adversarial settings. The results of A-OKVQA and GQA are presented in the Appendix. Specifically, under different settings, our method significantly surpasses the vanilla model's performance across all LVLMs. For example, with LLaVA1.5, MVP achieves an average improvement of 15.9 in Accuracy and 21.84 in F1 score across random, popular, and adversarial settings. For LLaVA1.5, Qwen-VL, and InstructBLIP, the improvement in F1 scores is mainly due to an increase in recall, while in mPLUG-Owl2, the increase comes from the simultaneous improvement of precision and recall. Furthermore, compared to VCD and OPERA, our method still achieves better results in most cases. These results demonstrate

Setting	Model	Decoding	Accuracy↑	Precision↑	Recall↑	F1 Score↑
Random	LLaVA1.5	Vanilla	$83.29_{(\pm 0.35)}$	$92.13_{(\pm 0.54)}$	$72.80_{(\pm 0.57)}$	$81.33_{(\pm 0.41)}$
		VCD	$87.73_{(\pm 0.40)}$	$91.42_{(\pm 0.55)}$	$83.28_{(\pm 0.42)}$	$87.16_{(\pm 0.41)}$
		OPERA	$89.17_{(\pm 0.15)}$	$93.21_{(\pm 0.21)}$	$85.20_{(\pm 0.37)}$	$89.03_{(\pm 0.11)}$
		Ours	**91.10**$_{(\pm 0.17)}$	$93.69_{(\pm 0.25)}$	$88.13_{(\pm 0.40)}$	**90.82**$_{(\pm 0.16)}$
	Qwen-VL	Vanilla	$84.36_{(\pm 0.48)}$	$95.65_{(\pm 0.43)}$	$72.00_{(\pm 0.32)}$	$82.16_{(\pm 0.51)}$
		VCD*	$86.03_{(\pm 0.13)}$	$95.92_{(\pm 0.35)}$	$75.26_{(\pm 0.13)}$	$84.34_{(\pm 0.10)}$
		OPERA	$86.13_{(\pm 0.21)}$	$97.54_{(\pm 0.37)}$	$74.13_{(\pm 0.18)}$	$84.24_{(\pm 0.22)}$
		Ours	**86.33**$_{(\pm 0.25)}$	$95.95_{(\pm 0.16)}$	$75.86_{(\pm 0.22)}$	**84.74**$_{(\pm 0.25)}$
	InstructBLIP	Vanilla	$80.71_{(\pm 0.73)}$	$81.67_{(\pm 0.67)}$	$79.19_{(\pm 1.14)}$	$80.41_{(\pm 0.80)}$
		VCD	$84.53_{(\pm 0.38)}$	$88.55_{(\pm 0.54)}$	$79.32_{(\pm 0.44)}$	$83.68_{(\pm 0.40)}$
		OPERA	$89.86_{(\pm 0.24)}$	$94.46_{(\pm 0.30)}$	$85.33_{(\pm 0.47)}$	$89.66_{(\pm 0.16)}$
		Ours	**90.30**$_{(\pm 0.41)}$	$92.54_{(\pm 0.28)}$	$87.66_{(\pm 0.31)}$	**90.04**$_{(\pm 0.19)}$
	mPLUG-Owl2	Vanilla	$86.70_{(\pm 0.18)}$	$91.73_{(\pm 0.45)}$	$80.66_{(\pm 0.33)}$	$85.84_{(\pm 0.56)}$
		VCD	$88.13_{(\pm 0.24)}$	$93.93_{(\pm 0.12)}$	$81.53_{(\pm 0.45)}$	$87.29_{(\pm 0.31)}$
		OPERA	$86.90_{(\pm 0.26)}$	$91.90_{(\pm 0.39)}$	$80.93_{(\pm 0.17)}$	$86.07_{(\pm 0.43)}$
		Ours	**91.13**$_{(\pm 0.26)}$	$92.49_{(\pm 0.14)}$	$89.53_{(\pm 0.36)}$	**90.98**$_{(\pm 0.24)}$
Popular	LLaVA1.5	Vanilla	$81.88_{(\pm 0.48)}$	$88.93_{(\pm 0.60)}$	$72.80_{(\pm 0.57)}$	$80.06_{(\pm 0.05)}$
		VCD	$85.38_{(\pm 0.38)}$	$86.92_{(\pm 0.53)}$	$83.28_{(\pm 0.42)}$	$85.06_{(\pm 0.37)}$
		OPERA	$86.00_{(\pm 0.33)}$	$84.09_{(\pm 0.18)}$	$88.80_{(\pm 0.44)}$	$86.38_{(\pm 0.17)}$
		Ours	**87.06**$_{(\pm 0.27)}$	$84.84_{(\pm 0.13)}$	$90.27_{(\pm 0.45)}$	**87.47**$_{(\pm 0.39)}$
	Qwen-VL	Vanilla	$84.06_{(\pm 0.18)}$	$94.20_{(\pm 0.43)}$	$72.60_{(\pm 0.45)}$	$82.00_{(\pm 0.23)}$
		VCD*	$85.80_{(\pm 0.07)}$	$94.82_{(\pm 0.10)}$	$75.73_{(\pm 0.19)}$	$84.21_{(\pm 0.09)}$
		OPERA	$85.73_{(\pm 0.21)}$	$96.52_{(\pm 0.15)}$	$74.13_{(\pm 0.11)}$	$83.86_{(\pm 0.14)}$
		Ours	**85.96**$_{(\pm 0.38)}$	$94.40_{(\pm 0.11)}$	$76.46_{(\pm 0.46)}$	**84.49**$_{(\pm 0.21)}$
	InstructBLIP	Vanilla	$78.22_{(\pm 0.84)}$	$77.87_{(\pm 1.03)}$	$78.85_{(\pm 0.52)}$	$78.36_{(\pm 0.76)}$
		VCD	$81.47_{(\pm 0.42)}$	$82.89_{(\pm 0.64)}$	$79.32_{(\pm 0.44)}$	$81.07_{(\pm 0.39)}$
		OPERA	**83.43**$_{(\pm 0.12)}$	$81.21_{(\pm 0.46)}$	$87.00_{(\pm 0.35)}$	**84.00**$_{(\pm 0.24)}$
		Ours	$79.93_{(\pm 0.23)}$	$76.01_{(\pm 0.25)}$	$87.46_{(\pm 0.37)}$	$81.34_{(\pm 0.49)}$
	mPLUG-Owl2	Vanilla	$83.66_{(\pm 0.37)}$	$85.51_{(\pm 0.25)}$	$81.06_{(\pm 0.48)}$	$83.23_{(\pm 0.19)}$
		VCD	$84.00_{(\pm 0.29)}$	$80.57_{(\pm 0.13)}$	$89.60_{(\pm 0.45)}$	$84.85_{(\pm 0.21)}$
		OPERA	$84.53_{(\pm 0.15)}$	$87.59_{(\pm 0.38)}$	$80.46_{(\pm 0.49)}$	$83.87_{(\pm 0.22)}$
		Ours	**86.30**$_{(\pm 0.03)}$	$90.12_{(\pm 0.28)}$	$81.53_{(\pm 0.14)}$	**85.61**$_{(\pm 0.47)}$
Adversarial	LLaVA1.5	Vanilla	$78.96_{(\pm 0.52)}$	$83.06_{(\pm 0.58)}$	$72.75_{(\pm 0.59)}$	$77.57_{(\pm 0.57)}$
		VCD	$80.88_{(\pm 0.33)}$	$79.45_{(\pm 0.29)}$	$83.29_{(\pm 0.43)}$	$81.33_{(\pm 0.34)}$
		OPERA	$79.13_{(\pm 0.31)}$	$74.41_{(\pm 0.23)}$	$88.80_{(\pm 0.41)}$	$80.97_{(\pm 0.14)}$
		Ours	**81.50**$_{(\pm 0.20)}$	$78.20_{(\pm 0.44)}$	$87.33_{(\pm 0.42)}$	**82.51**$_{(\pm 0.33)}$
	Qwen-VL	Vanilla	$82.66_{(\pm 0.30)}$	$90.49_{(\pm 0.33)}$	$73.00_{(\pm 0.50)}$	$80.81_{(\pm 0.37)}$
		VCD*	$83.30_{(\pm 0.39)}$	$89.17_{(\pm 0.45)}$	$75.80_{(\pm 0.39)}$	$81.94_{(\pm 0.39)}$
		OPERA	$83.93_{(\pm 0.45)}$	$92.55_{(\pm 0.22)}$	$73.80_{(\pm 0.37)}$	$82.12_{(\pm 0.18)}$
		Ours	**84.23**$_{(\pm 0.39)}$	$92.89_{(\pm 0.12)}$	$74.13_{(\pm 0.43)}$	**82.46**$_{(\pm 0.28)}$
	InstructBLIP	Vanilla	$75.84_{(\pm 0.45)}$	$74.30_{(\pm 0.63)}$	$79.03_{(\pm 0.68)}$	$76.59_{(\pm 0.40)}$
		VCD	$79.56_{(\pm 0.41)}$	$79.67_{(\pm 0.59)}$	$79.39_{(\pm 0.50)}$	$79.52_{(\pm 0.38)}$
		OPERA	$80.73_{(\pm 0.32)}$	$77.31_{(\pm 0.46)}$	$87.0_{(\pm 0.15)}$	$81.87_{(\pm 0.23)}$
		Ours	**80.82**$_{(\pm 0.24)}$	$81.23_{(\pm 0.18)}$	$82.91_{(\pm 0.38)}$	**82.06**$_{(\pm 0.15)}$
	mPLUG-Owl2	Vanilla	$81.73_{(\pm 0.44)}$	$82.82_{(\pm 0.33)}$	$80.06_{(\pm 0.18)}$	$81.42_{(\pm 0.29)}$
		VCD	$80.70_{(\pm 0.30)}$	$75.94_{(\pm 0.17)}$	$89.87_{(\pm 0.41)}$	$82.33_{(\pm 0.24)}$
		OPERA	$82.43_{(\pm 0.35)}$	$83.48_{(\pm 0.28)}$	$80.86_{(\pm 0.41)}$	$82.15_{(\pm 0.14)}$
		Ours	**84.40**$_{(\pm 0.08)}$	$86.49_{(\pm 0.37)}$	$81.53_{(\pm 0.19)}$	**83.94**$_{(\pm 0.61)}$

Table 1: Results on MSCOCO source of POPE. The best performances are **bolded**. VCD and OPERA are two recently proposed training-free methods in CVPR24. * denotes our reproduced VCD results with Qwen-VL.

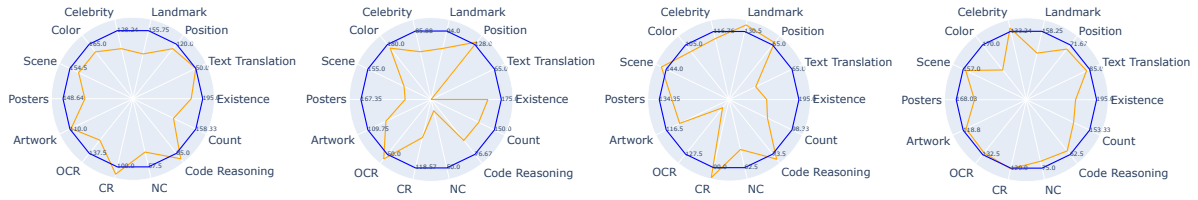

Figure 5: MME full set results on LLaVA-1.5, Qwen-VL, InstructBLIP, and mPLUG-Owl2 on 14 subtasks. The orange lines represent the vanilla model and the blue lines denotes our MVP model.

Model	Decoding	Object-level		Attribute-level		Total Scores↑
		Existence↑	*Count*↑	*Position*↑	*Color*↑	
LLaVA1.5	Vanilla	175.67	124.67	114.00	151.00	565.33
	VCD	184.66	138.33	**128.67**	153.00	604.66
	OPERA	185.00	108.33	111.67	145.00	550.00
	Ours	**195.00**	**158.33**	120.00	**165.00**	**638.33**
Qwen-VL	Vanilla	155.00	127.67	131.67	173.00	587.33
	VCD	156.00	131.00	128.00	**181.67**	596.67
	OPERA	165.00	145.00	**133.33**	180.00	623.33
	Ours	**175.00**	**150.00**	128.00	180.00	**633.00**
InstructBLIP	Vanilla	141.00	75.33	**66.67**	97.33	380.33
	VCD	168.33	92.33	64.00	**123.00**	447.67
	OPERA	156.00	78.33	55.00	95.00	384.33
	Ours	**195.00**	**98.33**	65.00	105.0	**456.67**
mPLUG-Owl2	Vanilla	160.00	130.00	68.33	123.33	481.66
	VCD	170.00	**155.00**	71.67	141.67	538.34
	OPERA	173.33	150.00	**85.00**	138.33	546.66
	Ours	**195.00**	153.33	71.67	**170.00**	**589.99**

Table 2: Results on the hallucination subset of MME. The best performances within each setting are **bolded**.

Regular	Bottom	Top	Accuracy	Precision	Recall	F1
-	-	-	79.33	74.88	88.26	81.02
✓			80.36	78.09	84.40	81.13
	✓		80.60	78.15	84.93	81.40
		✓	80.73	78.35	84.93	81.51
✓	✓		80.60	78.06	85.13	81.44
✓		✓	80.86	78.26	85.47	81.71
	✓	✓	80.90	77.80	86.46	81.91
✓	✓	✓	**81.50**	78.20	87.33	**82.51**

Table 3: Performance comparison of different views. Here "Bottom" means bottom-up perspective, and "Top" indicates top-down view. The experiments are conducted on "Adversarial" MSCOCO part of POPE using LLaVA1.5 model.

the effectiveness of our MVP.

Results on MME Hallucination Subset. We further evaluate our method on a subset of MME, which includes object-level hallucinations and attribute-level hallucinations. The results in Table 2 demonstrate that our method significantly improves the performance of all LVLMs in addressing object-level and attribute-level hallucinations. Additionally, we compare our method with the recent strong methods VCD and OPERA, and our approach still exhibits better overall performance.

Results on MME Full Set. As depicted in Figure 5, we test our method on the complete MME set to assess the overall capability of models. Four models with our MVP (blue lines) present significant improvement compared to the vanilla models on most evaluation subsets. This can be attributed to our method's introduction of multi-view information and multi-path reasoning, allowing LVLMs to comprehensively understand visual elements in the images and carefully consider the certainty of potential answers, thereby enhancing the LVLMs' capabilities in downstream tasks.

3.4 Ablation Study

3.4.1 Effectiveness of Multi-view Caption

Table 3 presents the performance from different perspectives. The first row presents the performance without using any additional caption information, while rows 2-4 respectively use a single perspective. The improvement is more pronounced with more perspectives involved. These results have confirmed that multi-view information can contribute to more comprehensive image understanding, thus mitigating the hallucinations in LVLMs. Notably, more views lead to greater performance improve-

Model	Caption	Accuracy	Precision	Recall	F1
LLaVA1.5	✗	91.10	93.69	88.13	90.82
	✓	92.36	94.04	90.46	92.22
		(+1.26)	(+0.35)	(+2.33)	(+1.40)
Qwen-VL	✗	86.33	95.95	75.86	84.74
	✓	86.56	95.97	76.33	85.03
		(+0.23)	(+0.02)	(+0.47)	(+0.29)
InstructBLIP	✗	90.30	92.54	87.66	90.04
	✓	91.33	92.80	89.73	91.24
		(+1.03)	(+0.26)	(+2.07)	(+1.20)
mPLUG-Owl2	✗	91.13	92.49	89.53	90.98
	✓	91.26	92.98	89.85	91.39
		(+0.13)	(+0.49)	(+0.32)	(+0.41)

Table 4: With captions from LLaVA1.6, our models achieve further improvement.

ment, with the cost of efficiency. Actually, our multi-view information-seeking strategy is flexible, indicating the number of views can be adaptively chosen in practical scenarios, thus achieving a balance between performance and efficiency.

3.4.2 The transferability of captions.

Intuitively, the quality of the captions has a direct impact on the model performance. Therefore, in this study, we explore the transferability of captions. Specifically, we employ a more powerful open-source model LLaVA1.6 (Liu et al., 2024b) to generate three-perspective captions for images in the random part of POPE MSCOCO, and use these captions for our model. As depicted in Table 4, with better captions, our method provides further improvements across four LVLMs. These results also confirm the significance of the multi-view information for alleviating hallucinations, as well as the plug-and-play flexibility of our method.

3.4.3 Multi-path Reasoning

To investigate the multi-path decoding, in this study, we only adopt the regular perspective. We experiment on the random and adversarial part of the POPE MSCOCO benchmark, as shown in Figure 6. We first conduct experiments on Top-K in Equation 5. As K increases from 1 to 5, peak performance is observed at K equals 3. When K becomes larger, the performance does not improve. This is due to the decoding path extending from the first tokens with minuscule probabilities do not provide any beneficial information.

Secondly, we explore a new aggregation strategy MVP-Max. Instead of accumulating the certainty scores in Equation 5, the potential answer with maximum certainty score among all paths is chosen as the final answer. It can be seen that after using MVP-Max, the final performance of the model decreases significantly. This demonstrates

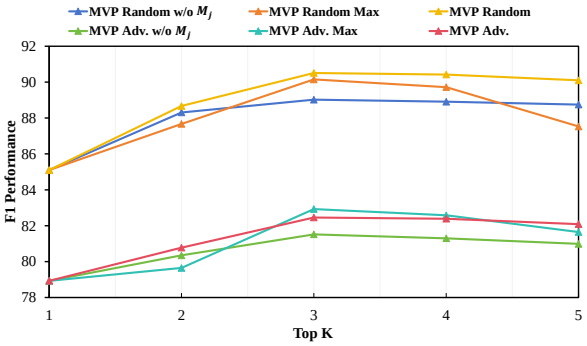

Figure 6: The ablation study of multi-path variants.

Setting	Decoding	Accuracy↑	Precision↑	Recall↑	F1 Score↑
Random	Greedy	91.16	93.63	88.31	90.89
	Nucleus	91.10	93.69	88.13	90.82
	Beam Search	**92.36**	94.04	90.46	**92.21**
	VCD	89.73	95.77	83.13	89.01
	OPERA	89.90	96.15	83.17	89.19
Adversarial	Greedy	81.63	78.34	87.39	82.62
	Nucleus	81.50	78.20	87.33	82.51
	Beam Search	81.86	77.63	89.53	**83.16**
	VCD	82.56	82.20	83.13	82.66
	OPERA	**82.65**	82.31	83.22	82.86

Table 5: Ablation of different decoding strategies using LLaVA1.5. Experiments are performed on the MSCOCO source of the POPE benchmark.

the effectiveness of our aggregation strategy.

Finally, we explore removing M_j in Equation 5, we found that relying solely on the certainty of tokens will damage the stability and effectiveness of our model.

3.5 Decoding Strategy

In this section, we analyze the impact of different decoding strategies on our method. Specifically, we investigate five decoding methods. Notably, nucleus sampling is used in our main experiments for a fair comparison with recent methods. Our MVP can further enhance the performance with training-free decoding methods such as VCD and OPERA, as shown in Table 5. We can observe that using beam search as the decoding strategy performs the best accuracy on the random setting, while OPERA achieves the most significant accuracy on the adversarial part. These results also imply that our method is a novel plug-and-play approach, which can be flexibly integrated with other techniques.

4 Conclusion

In this paper, we propose a novel training-free framework MVP to reduce hallucinations in LVLMs through Multi-View Multi-Path Reasoning. Specifically, we devise a multi-view information-seeking strategy to perceive the intricate details of

the image information. Furthermore, we propose multi-path reasoning to quantify and aggregate the certainty scores for each potential answer and finally decide the output answer. With the multi-view multi-path reasoning, our method effectively alleviates hallucinations in LVLMs.

Limitations

There are still some limitations in our work, which are listed below:

Latency Due to the introduction of the multi-view image information seeking, our method may have a higher reasoning time than trivial methods. However, the number of views can be flexibly decided, thus achieving a balance between performance and efficiency.

Applicability Our method mainly applies to question answering tasks and is not specifically designed for image captioning tasks. However, it is possible to extend our method to image captioning. We have detailed additional experiments in the Appendix.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110. The authors would also like to thank the anonymous reviewers or their comments on improving the quality of this paper.

References

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2022. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking micro-action recognition: Dataset, method, and application. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

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

Chris Kelly, Luhui Hu, Bang Yang, Yu Tian, Deshun Yang, Cindy Yang, Zaoshan Huang, Zihao Li, Jiayin Hu, and Yuexian Zou. 2024. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *arXiv preprint arXiv:2403.09027*.

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*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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

Tsung-Yi Lin, Michael Maire, Serge 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 13*, pages 740–755. Springer.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021a. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244.

Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078.

Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. 2021b. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2100–2108.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. 2025. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Wei Hu, and Yu Cheng. 2024a. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.

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, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023b. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

Kangcheng Liu, Xinhu Zheng, Chaoqun Wang, Hesheng Wang, Ming Liu, and Kai Tang. 2024c. Online robot navigation and and manipulation with distilled vision-language models. *arXiv preprint arXiv:2401.17083*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. *arXiv preprint arXiv:2406.15479*.

Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. 2024a. Mitigating dialogue hallucination for large multi-modal models via adversarial instruction tuning. *arXiv preprint arXiv:2403.10492*.

SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. 2024b. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 980–987.

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, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR.

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

Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024. Surf:

Teaching large vision-language models to selectively utilize retrieved information. *arXiv preprint arXiv:2409.14083*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

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.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *arXiv preprint arXiv:2402.06118*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

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. 2024a. Avoiding feature suppression in contrastive learning: Learning what has not been learned before. *arXiv preprint arXiv:2402.11816*.

Jihai Zhang, Xiaoye Qu, Tong Zhu, and Yu Cheng. 2024b. Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling. *arXiv preprint arXiv:2409.19291*.

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*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

A Implementation Details

In our paper, we adopt three different views to capture the information from the image. For "Bottom-up" perspective, we use following prompt: *"Through a systematic examination of the image at the pixel level and by analyzing various visual features, such as shape, color, and texture, along with employing object detection techniques, describe this image in details."* In addition, we use *"Describe this image in details"* for the regular caption. The prompt for top-down perspective has been described in Section 3.3. In addition, to generate these multi-view captions, a temperature of 0.9 and a top-p parameter of 0.95 are set to guarantee diversity. K in Equation 5 is set to 3. α in Equation 6 is set to 0.4, 0.2, and 0.4 for "bottom-up", "regular", and "top-down" perspectives, considering that the "bottom-up" and "top-down" perspectives bring more beneficial information. We do not tune α much as the aggregation from multiple decoding paths inherently presents certain robustness.

B Related Work

B.1 Hallucination in LVLMs

Recently, the potential hallucination (Qu et al., 2024a,b; Liu et al., 2024a; Lu et al., 2024; Su et al., 2024) problem in large models has garnered considerable attention, mainly due to the direct impact on the downstream applications (Liu et al., 2021a,b, 2020; Qu et al., 2020; Dong et al., 2022; Guo et al., 2024). In LVLMs, the term "hallucination" refers to models that generate seemingly plausible outputs inclusive of objects, attributes, or relations that do not correspond with the images.

Regarding hallucination mitigation, the primary focus of most current methods has been to enhance the quality of the supervised fine-tuning or reinforcement learning data. VIGC (Wang et al., 2024a) presents a component to correct visual instructions with the aim to minimize hallucinations generated from lengthy sequences. LRV (Liu et al., 2023a) attempt to alleviate hallucinations by developing expansive and diverse SFT data. For methods based on reinforcement learning, LLaVA-RLHF (Sun et al., 2023) is the pioneer in applying Reinforcement Learning with Human Feedback (RLHF) to mitigate hallucination in LVLMs. RLHF-V (Yu et al., 2023) further develops a fine-grained correctional human feedback. Considering the instability and training difficulty of RLHF, Zhao et al. (2023) employ Direct Preference Optimization (DPO) and build a hallucination-aware dataset for alleviating hallucination. Although these methods have achieved significant improvements, they inevitably introduce a large training cost and are prone to overfitting to the training data. Instead, there are also training-free works aiming to solve the hallucination without introducing training cost. VCD (Leng et al., 2023) contrasts the output distributions derived from original and distorted visual inputs, aiming to recalibrate the model's excessive dependence on unimodal priors and statistical biases. OPERA (Huang et al., 2023) introduces a penalty term to the model logits during the beam-search decoding to alleviate the overconfidence problem, complemented by a rollback strategy. In this paper, we focus on the training-free paradigm for mitigating hallucination in LVLMs.

B.2 Training-free Decoding Strategy

As the recent training-free methods for mitigating hallucination focus on the decoding process, we describe several decoding strategies here. Decoding strategies in language models are instrumental in guiding how these models produce text. They are a significant factor in influencing the quality, relevance, and coherence of the output. Greedy decoding takes the simplest approach, choosing the most probable next word at every step. Despite its speed and computational efficiency, this method often results in repetitive and monotonous text. Conversely, beam search offers a more advanced technique that maintains a predetermined number of hypotheses at each step, elaborating on them to identify a more optimum sequence. In nucleus sampling, a flexible range of words is considered, which accumulate to achieve the given probability p. More recently, there are two methods specifically proposed for mitigating hallucinations. VCD adopts contrastive decoding to calibrate the model's output distribution. OPERA introduces a penalty term on the model logits during the beam-search decoding to mitigate the over-trust issue. In this paper, we focus on the certainty of the answer token during the decoding process, which does not conflict with designing different decoding paths. Thus, our MVP framework is plug-and-play and can further combine with the above decoding strategies.

	LLaVA-1.5		InstructBLIP	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Nucleus	48.8	14.2	54.6	24.8
Ours	45.4	12.9	50.4	20.5

Table 6: CHAIR hallucination evaluation results.

	MS-COCO				
	BLUE-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑	SPICE ↑
Base	22.3	28.0	50.9	75.3	22.1
Ours	22.6	28.4	52.3	79.6	22.5

Table 7: Text quality evaluation on MSCOCO dataset.

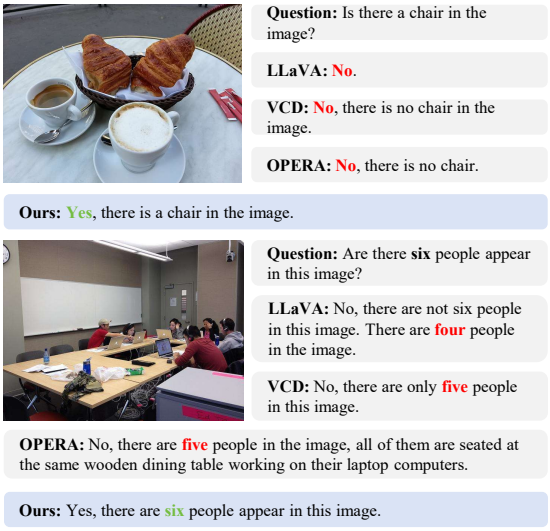

Figure 7: Two representative examples from POPE and MME datasets. The qualitative results of LLaVA 1.5, VCD, OPERA, and our proposed MVP.

C Additional Experiments

C.1 Results on POPE benchmark

To further demonstrate the effectiveness of our proposed MVP, we conduct experiments on POPE based on AOKVQA and GQA with random, popular, and adversarial settings, respectively. The experiment settings are the same as the main experiment in the MSCOCO. The results are shown in Tables 8 and 9. It is obvious that our proposed MVP has greatly improved from these two tables compared with the baseline models. Specifically, under different settings, our method significantly surpasses the vanilla model's performance across all LVLMs. For example, with LLaVA1.5, MVP achieves an average improvement of 4.38 in Accuracy and 6.29 in F1 score across random, popular, and adversarial settings on AOKVQA. For LLaVA1.5, InstructBLIP and mPLUG-Owl2, the improvement in F1 and Accuracy scores is mainly due to an increase in precision, while in Qwen-VL, the increase comes from the simultaneous improvement of precision and recall. These results demonstrate the effectiveness of our MVP in alleviating hallucinations.

C.2 Results on Image Caption

In this paper, we mainly focus on the question-answer task. However, it is possible to extend our method to image captioning. For each view (eg. bottom-up), we can perceive different kinds of objects and generate multiple captions with the top-K path. Finally, all generated captions from different views alongside the input image can be summarized by the LVLM to obtain the final caption. In this way, we can completely grasp the image information and output accurate captions. We report the image caption results of CHAIR metric (Rohrbach et al., 2018) in Table 6. The maximum new tokens of this experiment are 512. In the same setting, we implement nucleus sampling and our MVP. As shown in this table, our MVP obtains better performance than the base method. Furthermore, we use the official COCO evaluation toolkit and perform textual quality evaluation on LLaVA 1.5. As shown in Table 7, our method can also effectively improve the textual quality of the generation.

D Qualitative Results

To qualitatively verify the effectiveness of our method on downstream tasks, we presented two examples from the MSCOCO POPE and MME datasets. As illustrated in Figures 7, in both examples, MVP is able to accurately address questions. In the top figure where the chair is in the top right corner and not fully visible, our MVP comprehensively captures multi-view information, thus contributing to identifying this chair. In the bottom figure where distant people appear blurry, our MVP effectively decides the accurate number of people with multi-path reasoning. Through these two direct comparisons, our method answers questions more precisely than the currently existing strong baselines, significantly reducing hallucinations in LVLM models.

Setting	Model	Decoding	Accuracy↑	Precision↑	Recall↑	F1 Score↑
Random	LLaVA1.5	Vanilla	$83.45_{(\pm 0.48)}$	$87.24_{(\pm 0.68)}$	$78.36_{(\pm 0.54)}$	$82.56_{(\pm 0.50)}$
		Ours	$\mathbf{91.20}_{(\pm 0.48)}$	$89.06_{(\pm 0.68)}$	$93.93_{(\pm 0.54)}$	$\mathbf{91.43}_{(\pm 0.50)}$
	Qwen-VL	Vanilla	$86.67_{(\pm 0.48)}$	$93.16_{(\pm 0.55)}$	$79.16_{(\pm 0.59)}$	$85.59_{(\pm 0.53)}$
		Ours	$\mathbf{88.16}_{(\pm 0.45)}$	$94.34_{(\pm 0.26)}$	$81.20_{(\pm 0.19)}$	$\mathbf{87.28}_{(\pm 0.06)}$
	InstructBLIP	Vanilla	$80.91_{(\pm 0.34)}$	$77.97_{(\pm 0.59)}$	$86.16_{(\pm 0.88)}$	$81.86_{(\pm 0.32)}$
		Ours	$\mathbf{88.60}_{(\pm 0.16)}$	$90.95_{(\pm 0.32)}$	$85.73_{(\pm 0.50)}$	$\mathbf{88.26}_{(\pm 0.07)}$
	mPLUG-Owl2	Vanilla	$77.63_{(\pm 0.14)}$	$70.46_{(\pm 0.29)}$	$95.13_{(\pm 0.42)}$	$80.96_{(\pm 0.11)}$
		Ours	$\mathbf{90.30}_{(\pm 0.33)}$	$88.92_{(\pm 0.44)}$	$92.07_{(\pm 0.11)}$	$\mathbf{90.47}_{(\pm 0.28)}$
Popular	LLaVA1.5	Vanilla	$79.90_{(\pm 0.33)}$	$80.85_{(\pm 0.31)}$	$78.36_{(\pm 0.54)}$	$79.59_{(\pm 0.37)}$
		Ours	$\mathbf{84.60}_{(\pm 0.23)}$	$79.19_{(\pm 0.37)}$	$93.87_{(\pm 0.08)}$	$\mathbf{85.91}_{(\pm 0.49)}$
	Qwen-VL	Vanilla	$85.56_{(\pm 0.35)}$	$90.44_{(\pm 0.56)}$	$79.53_{(\pm 0.84)}$	$84.63_{(\pm 0.42)}$
		Ours	$\mathbf{87.30}_{(\pm 0.15)}$	$92.48_{(\pm 0.33)}$	$81.20_{(\pm 0.27)}$	$\mathbf{86.47}_{(\pm 0.40)}$
	InstructBLIP	Vanilla	$76.19_{(\pm 0.80)}$	$72.16_{(\pm 0.69)}$	$85.28_{(\pm 0.79)}$	$78.17_{(\pm 0.73)}$
		Ours	$\mathbf{78.16}_{(\pm 0.80)}$	$74.90_{(\pm 0.69)}$	$84.73_{(\pm 0.79)}$	$\mathbf{79.51}_{(\pm 0.73)}$
	mPLUG-Owl2	Vanilla	$72.06_{(\pm 0.15)}$	$65.16_{(\pm 0.30)}$	$94.80_{(\pm 0.05)}$	$77.24_{(\pm 0.33)}$
		Ours	$\mathbf{83.30}_{(\pm 0.28)}$	$78.56_{(\pm 0.47)}$	$91.60_{(\pm 0.13)}$	$\mathbf{84.58}_{(\pm 0.22)}$
Adversarial	LLaVA1.5	Vanilla	$74.04_{(\pm 0.34)}$	$72.08_{(\pm 0.53)}$	$78.49_{(\pm 0.38)}$	$75.15_{(\pm 0.23)}$
		Ours	$\mathbf{74.70}_{(\pm 0.12)}$	$67.75_{(\pm 0.45)}$	$94.27_{(\pm 0.36)}$	$\mathbf{78.84}_{(\pm 0.09)}$
	Qwen-VL	Vanilla	$79.57_{(\pm 0.31)}$	$79.77_{(\pm 0.34)}$	$79.23_{(\pm 0.73)}$	$79.50_{(\pm 0.38)}$
		Ours	$\mathbf{81.60}_{(\pm 0.22)}$	$81.85_{(\pm 0.38)}$	$81.20_{(\pm 0.07)}$	$\mathbf{81.53}_{(\pm 0.54)}$
	InstructBLIP	Vanilla	$70.71_{(\pm 0.76)}$	$65.91_{(\pm 0.74)}$	$85.83_{(\pm 0.80)}$	$75.56_{(\pm 0.57)}$
		Ours	$\mathbf{75.03}_{(\pm 0.16)}$	$73.75_{(\pm 0.31)}$	$85.25_{(\pm 0.08)}$	$\mathbf{79.08}_{(\pm 0.29)}$
	mPLUG-Owl2	Vanilla	$55.13_{(\pm 0.14)}$	$53.26_{(\pm 0.42)}$	$83.73_{(\pm 0.24)}$	$65.11_{(\pm 0.33)}$
		Ours	$\mathbf{73.43}_{(\pm 0.19)}$	$67.32_{(\pm 0.48)}$	$91.07_{(\pm 0.11)}$	$\mathbf{77.42}_{(\pm 0.37)}$

Table 8: Results on AOKVQA source of POPE benchmark. The best performances are **bolded**.

Setting	Model	Decoding	Accuracy↑	Precision↑	Recall↑	F1 Score↑
Random	LLaVA1.5	Vanilla	$83.73_{(\pm 0.27)}$	$87.16_{(\pm 0.39)}$	$79.12_{(\pm 0.35)}$	$82.95_{(\pm 0.28)}$
		Ours	$\mathbf{90.20}_{(\pm 0.17)}$	$87.64_{(\pm 0.29)}$	$93.60_{(\pm 0.38)}$	$\mathbf{90.52}_{(\pm 0.07)}$
	Qwen-VL	Vanilla	$80.97_{(\pm 0.32)}$	$88.07_{(\pm 0.34)}$	$71.64_{(\pm 0.57)}$	$79.01_{(\pm 0.40)}$
		Ours	$\mathbf{83.57}_{(\pm 0.24)}$	$91.10_{(\pm 0.35)}$	$74.40_{(\pm 0.10)}$	$\mathbf{81.91}_{(\pm 0.44)}$
	InstructBLIP	Vanilla	$79.65_{(\pm 0.24)}$	$77.14_{(\pm 0.43)}$	$84.29_{(\pm 0.36)}$	$80.56_{(\pm 0.18)}$
		Ours	$\mathbf{85.67}_{(\pm 0.28)}$	$90.65_{(\pm 0.33)}$	$79.53_{(\pm 0.19)}$	$\mathbf{84.73}_{(\pm 0.22)}$
	mPLUG-Owl2	Vanilla	$80.43_{(\pm 0.26)}$	$75.01_{(\pm 0.30)}$	$91.26_{(\pm 0.15)}$	$82.34_{(\pm 0.39)}$
		Ours	$\mathbf{89.00}_{(\pm 0.23)}$	$90.18_{(\pm 0.41)}$	$87.53_{(\pm 0.12)}$	$\mathbf{88.84}_{(\pm 0.45)}$
Popular	LLaVA1.5	Vanilla	$78.17_{(\pm 0.17)}$	$77.64_{(\pm 0.26)}$	$79.12_{(\pm 0.35)}$	$78.37_{(\pm 0.18)}$
		Ours	$\mathbf{79.43}_{(\pm 0.32)}$	$72.94_{(\pm 0.25)}$	$93.61_{(\pm 0.40)}$	$\mathbf{81.99}_{(\pm 0.07)}$
	Qwen-VL	Vanilla	$75.99_{(\pm 0.33)}$	$78.62_{(\pm 0.41)}$	$71.40_{(\pm 0.38)}$	$74.84_{(\pm 0.34)}$
		Ours	$\mathbf{79.80}_{(\pm 0.29)}$	$83.40_{(\pm 0.37)}$	$74.40_{(\pm 0.12)}$	$\mathbf{78.65}_{(\pm 0.45)}$
	InstructBLIP	Vanilla	$73.87_{(\pm 0.58)}$	$69.63_{(\pm 0.54)}$	$84.69_{(\pm 0.68)}$	$76.42_{(\pm 0.52)}$
		Ours	$\mathbf{73.98}_{(\pm 0.26)}$	$78.08_{(\pm 0.33)}$	$78.08_{(\pm 0.18)}$	$\mathbf{78.08}_{(\pm 0.23)}$
	mPLUG-Owl2	Vanilla	$71.96_{(\pm 0.20)}$	$66.01_{(\pm 0.40)}$	$90.53_{(\pm 0.15)}$	$76.35_{(\pm 0.48)}$
		Ours	$\mathbf{77.87}_{(\pm 0.35)}$	$73.46_{(\pm 0.47)}$	$87.27_{(\pm 0.10)}$	$\mathbf{79.77}_{(\pm 0.22)}$
Adversarial	LLaVA1.5	Vanilla	$75.08_{(\pm 0.33)}$	$73.19_{(\pm 0.49)}$	$79.16_{(\pm 0.35)}$	$76.06_{(\pm 0.24)}$
		Ours	$\mathbf{77.53}_{(\pm 0.21)}$	$67.81_{(\pm 0.34)}$	$93.13_{(\pm 0.43)}$	$\mathbf{78.48}_{(\pm 0.08)}$
	Qwen-VL	Vanilla	$75.46_{(\pm 0.63)}$	$77.92_{(\pm 0.73)}$	$71.07_{(\pm 0.97)}$	$74.33_{(\pm 0.71)}$
		Ours	$\mathbf{79.20}_{(\pm 0.25)}$	$82.30_{(\pm 0.36)}$	$74.40_{(\pm 0.14)}$	$\mathbf{78.15}_{(\pm 0.47)}$
	InstructBLIP	Vanilla	$70.56_{(\pm 0.53)}$	$66.12_{(\pm 0.32)}$	$84.33_{(\pm 1.05)}$	$74.12_{(\pm 0.58)}$
		Ours	$\mathbf{74.71}_{(\pm 0.28)}$	$70.37_{(\pm 0.32)}$	$85.31_{(\pm 0.17)}$	$\mathbf{77.12}_{(\pm 0.23)}$
	mPLUG-Owl2	Vanilla	$54.86_{(\pm 0.22)}$	$53.36_{(\pm 0.39)}$	$77.13_{(\pm 0.13)}$	$63.08_{(\pm 0.41)}$
		Ours	$\mathbf{75.26}_{(\pm 0.30)}$	$71.17_{(\pm 0.33)}$	$84.93_{(\pm 0.16)}$	$\mathbf{77.44}_{(\pm 0.37)}$

Table 9: Results on GQA source of POPE benchmark. The best performances are **bolded**.