PLANE: Plug-and-Play Fine-Grained Visual Experts for Hallucination Mitigation in Large Vision-Language Models

Jianghao Yin[1], **Lang Yu**[1], **Qianxi Yu**[1], **Qin Chen**[1*], **Jie Zhou**[1], **Zhikai Lei**[2], **Liang He**[1]
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Qiji Zhifeng Co., Ltd.

jhyin@stu.ecnu.edu.cn, qchen@cs.ecnu.edu.cn

Abstract

Large Vision-Language Models (LVLMs) commonly suffer from hallucinations, which appear as fabricating non-existent objects or outputting incorrect attributes and spatial relationships in images. Previous works empirically indicate that CLIP-style vision encoders tend to focus on primary objects while ignoring image details, which is one of the causes of hallucinations. To address this issue, we propose plug-and-play fine-grained visual experts (PLANE) comprising two components: LoRA visual experts operating on CLIP for multi-aspect image encoding, and an instruction-aware gate network for adaptive image feature fusion. During training, we construct a fine-grained image-caption pairs dataset, where images are described from three aspects: objects, attributes, and relations between objects. Image-caption pairs from different aspects are used to train the three fine-grained LoRA experts respectively. Aiming at comprehensively evaluate our method's capabilities, we manually annotate a fine-grained hallucination evaluation benchmark for LVLMs which is GPT4-independent during evaluation phase (GIFT). We conduct experiments both on popular hallucination evaluation benchmarks and GIFT. The results show that our method effectively reduces hallucinations across models of different types and sizes while introducing minimal inference latency. Compared to advanced baselines, our model also demonstrates superior performance.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional generation and reasoning capabilities (Achiam et al., 2023; Dubey et al., 2024; Yang et al., 2024a; Jiang et al., 2024; Yang et al., 2023; GLM et al., 2024). Building on this foundation, Large Vision-Language Models (LVLMs) have rapidly developed, typically consisting of a CLIP-style vision encoder (Radford et al., 2021), a projection layer, and an LLM (Zhang et al., 2024; Jin et al., 2025; Liu et al., 2024b; Bai et al., 2023; Dai et al., 2023; Li et al., 2025a; Wei et al., 2024). They have shown excellent performance in tasks such as image captioning (Stefanini et al., 2022; Li et al., 2023a) and visual question answering (VQA) (Antol et al., 2015; Schwenk et al., 2022). However, similar to LLMs (Chen et al., 2025a; Huang et al., 2025), LVLMs also suffer from severe hallucination issues (Liu et al., 2024a; Bai et al., 2024; Leng et al., 2024; Yin et al., 2024; Yue et al., 2024; Yuan et al., 2024; Xing et al., 2024; Yin et al., 2026a; Chen et al., 2025b). As illustrated in Figure 1, previous work categorizes multimodal hallucinations into three types: object hallucinations, attribute hallucinations, and relation hallucinations (Liu et al., 2024a). These hallucinations undermine the credibility of LVLMs and hinder their application in various fields.

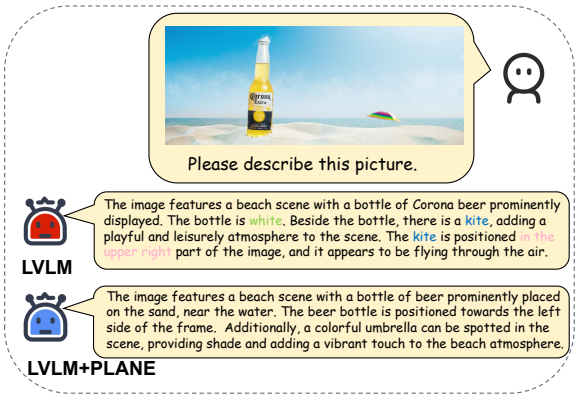


Figure 1: A hallucination case in LVLMs, with different colors representing object hallucinations, attribute hallucinations, and relation hallucinations.

Many methods have been proposed to mitigate hallucinations in LVLMs, and these can generally fall into three types. The first type is decoding-based methods, such as VCD (Leng et al., 2024)

*Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37559–37576
October 24-29, 2026 ©2026 Association for Computational Linguistics

and OPERA (Huang et al., 2024), which introduce specific strategies during the inference stage to influence the model's final output. While these methods typically do not require additional training, they will increase the inference time. The second type involves post-hoc methods, such as Volcano (Lee et al., 2024) and LURE (Zhou et al., 2024) which use a revisor model to correct hallucinations in the generated text. However, these approaches often increase the time for obtaining the final output and are unable to address the inner cause of hallucinations within the model. The third type is fine-tuning-based methods, such as LRV (Liu et al., 2023). These methods generally create more robust datasets to fine-tune the LVLMs. However, a drawback of this approach is that it usually requires retraining when improving different LVLMs, consuming significant computational resources. Given the shortcomings of existing multimodal hallucination mitigation methods, can we mitigate hallucinations from the perspective of the internal causes that trigger hallucinations in LVLMs and propose an efficient plug-and-play solution that can be directly integrated into multiple LVLMs without requiring retraining?

Previous work has pointed out that CLIP-style vision encoders (Radford et al., 2021) tend to focus on primary objects while neglecting the details of images, which can also lead to hallucinations in LVLMs (Jain et al., 2024; Cho et al., 2022; Liu et al., 2024a). Based on these insights, this work aims to address the hallucination issue through enhancement of the vision encoder in LVLMs. An intuitive idea is to train a vision encoder by aligning it with highly detailed captions. However, the CLIP text encoder can only encode 77 tokens (Urbanek et al., 2024), which significantly limits the level of detail that captions can provide about the image. Therefore, we decompose a detailed caption into multiple aspects to train several expert encoders. Based on the above analysis, we propose **p**lug-and-**l**ay fine-gr**a**i**n**ed visual **e**xperts (PLANE) which consists of three fine-grained LoRA (Hu et al., 2022) visual experts operating on CLIP and an instruction-aware gate network. The former encodes images at object-aspect, attribute-aspect, and relation-aspect, while the latter fuses different features at the token-level according to instructions. The fused features are then used for LVLMs inference. To train the model, we constructed a fine-grained image-caption pairs dataset, which consists of three types of image-caption pairs: image-object caption pairs, image-attribute caption pairs, and image-relation caption pairs. The object captions primarily describe the objects present in the image, the attribute captions describe various attributes of the objects in the image, such as color, texture, shape, size, etc., and the relation captions describe the spatial relationships between objects in the image. Then, we use the constructed three types image-caption pairs to train three different LoRA visual expert, with each expert specializing in capturing different aspects of the image features.

Currently, existing multimodal generative hallucination evaluation benchmarks have limitations: some only focus on object-level hallucination (Rohrbach et al., 2018; Liu et al., 2023; Wang et al., 2023), while others, although considering fine-grained hallucination, rely on ChatGPT or GPT-4 for assessment (Jing et al., 2024), making the evaluation costly and difficult to scale. To comprehensively evaluate model performance, we have constructed a **G**PT4-**i**ndependent (during evaluation phase) **f**ine-grained hallucina**t**ion evaluation benchmark (GIFT) which is manually annotated. This benchmark enables the evaluation of LVLMs' hallucinations at the object-level, attribute-level, and relation-level in open-ended generation tasks. In summary, the main contributions of our paper are:

- We propose **p**lug-and-**l**ay fine-gr**a**i**n**ed visual **e**xperts (PLANE) for hallucination mitigation, which is trained on our constructed fine-grained image-caption pairs dataset covering object-aspect, attribute-aspect, and relation-aspect. Once trained, PLANE can be directly integrated into other LVLMs that utilize the same vision encoder without requiring retraining.

- We construct a **G**PT4-**i**ndependent (during evaluation phase) **f**ine-grained hallucina**t**ion evaluation benchmark (GIFT) that differs from previous object-level focused generative benchmarks. Our benchmark evaluates model hallucinations at object, attribute, and relation levels.

- We conduct detailed experiments on different task benchmarks. The results show that our method effectively reduces hallucinations across models of different types and sizes while introducing minimal inference latency.

Compared to advanced baselines, our model also demonstrates superior performance.

2 Related Work

2.1 Hallucinations in LVLMs

To mitigate hallucinations in LVLMs, numerous methods have been proposed. VCD (Leng et al., 2024) and OPERA (Huang et al., 2024) modify decoding strategies but increase inference time. Volcano (Lee et al., 2024) and Woodpecker (Yin et al., 2024) use post-hoc correction with additional models, which doesn't address the root cause and also increases inference time. LRV (Liu et al., 2023) constructs more robust training datasets but requires retraining for different models. Recently, some work has focused on image encoders. Liu et al. (Liu et al., 2024d) explored object hallucinations in CLIP, using fine-grained object-level contrastive loss to fine-tune the CLIP model and reduce hallucinations in LVLMs. However, they only focus object-level hallucinations. Our goal is to propose a vision encoder optimization method that can mitigate hallucinations at multiple levels.

2.2 Hallucination Benchmarks for LVLMs

Existing LVLM hallucination evaluation benchmarks assess hallucination through discriminative and generative tasks (Liu et al., 2024a; Bai et al., 2024). Discriminative tasks like POPE (Li et al., 2023b) and MME (Fu et al., 2026) require binary Yes/No answers, with POPE focusing on object-level hallucination and MME covering objects, attributes, and relations. Generative tasks evaluate hallucinations in image descriptions using benchmarks like GAVIE (Liu et al., 2023) and Faithscore (Jing et al., 2024), which rely on GPT-4 or ChatGPT as judges. While enabling fine-grained multi-level evaluation, their dependence on GPT-4/ChatGPT limits scalability. Alternative approaches like CHAIR (Rohrbach et al., 2018) and AMBER (Wang et al., 2023) avoid GPT-4 but are restricted to object-level assessment. To address this limitation, we propose GIFT, a GPT-4-independent benchmark (during evaluation phase) for comprehensive fine-grained hallucination evaluation.

3 Method

The model architecture of PLANE is demonstrated in Figure 3. The main idea is to encode images with multiple visual experts that focus on different aspects, obtaining diverse visual features, which are then fused based on the instruction, and finally, the fused features are used for LVLMs inference. In this section, we will explain PLANE's training data construction, model architecture, and training methods.

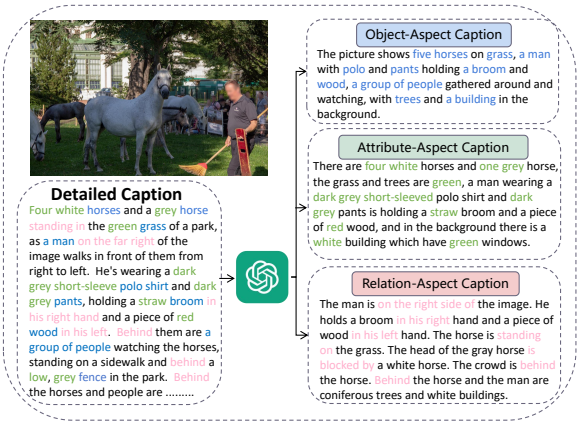


Figure 2: The construction process for the fine-grained image-caption pairs dataset, where different colors indicate descriptions related to object, attribute, and relation.

3.1 Dataset Construction

Urbanek et al. (Urbanek et al., 2024) develop the Densely Captioned Images (DCI) dataset, a manually annotated, dense and highly-aligned image-text pair dataset. They also demonstrate that fine-tuning vision-language models (VLMs) using a summarized version of this small dataset improved VLM performance. Inspired by this work, we build our fine-grained image-caption dataset based on the DCI dataset. As shown in Figure 2, we use ChatGPT to decompose the dense image captions into three categories: object captions, attribute captions, and relation captions. The object caption describes what is present in the image, the attribute caption focuses on the characteristics of each object, such as color, texture, and material, while the relation caption highlights the spatial relationships between objects in the image. We ensure that the length of each caption is kept within 77 tokens, and together these three captions aim to cover the content of the original dense caption as comprehensively as possible. For captions exceeding 77 tokens in length, ChatGPT will be called again to summarize them to within 77 tokens. To increase data diversity, we filter out a set of image-caption pairs related to attributes and position from the MSCOCO dataset (Lin et al., 2014) and incorporate them into our dataset. Finally, we construct a fine-grained image-

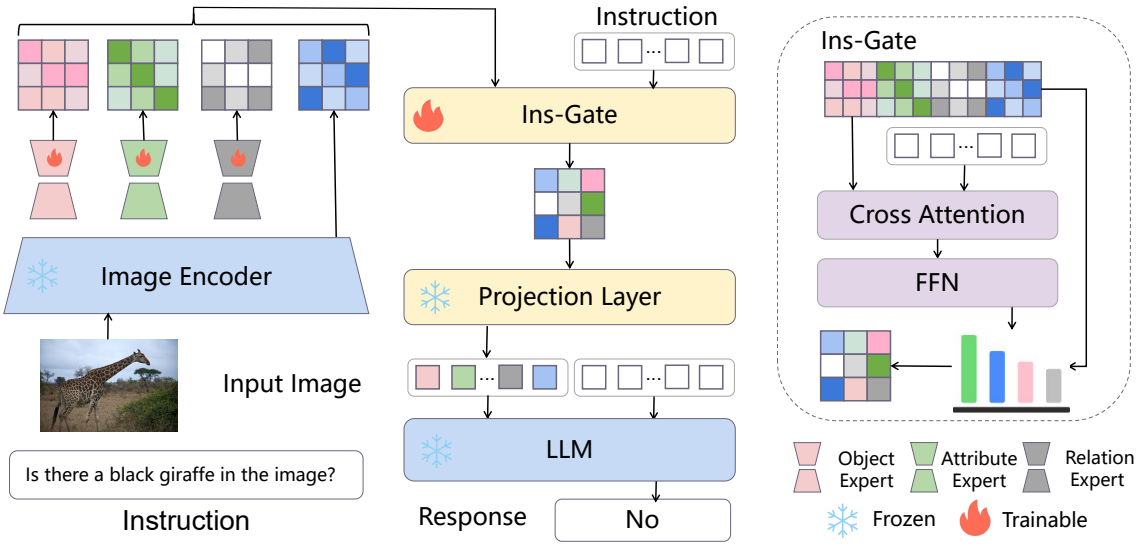


Figure 3: Overview of the PLANE. The input image is encoded by three LoRA visual experts and the original LVLM encoder, generating multi-level image features. These features are fused by Ins-Gate based on the instruction for LVLM inference.

caption pairs dataset with a total of 38,375 samples, which consists of 7,800 image-object caption pairs, 18,121 image-attribute caption pairs, and 12,454 image-relation caption pairs.

Previous studies show that hard negative captions during CLIP training can improve model performance (Yuksekgonul et al., 2022; Urbanek et al., 2024). Therefore, we design hallucination captions that maintain similar structure but differ semantically from original captions. Using ChatGPT, we generate three types of hallucinations for different fine-grained captions : object hallucinations by replacing or adding non-existent objects, attribute hallucinations by modifying descriptive terms (e.g., 'wooden spoon' to 'glass spoon'), and relation hallucinations by altering spatial relationships (e.g., 'birds on the tree' to 'birds under the tree').

Formally, our dataset $D = \{D_i\}$, $i \in \{obj, attr, rela\}$, where i indicates the specific aspect of the caption-image pairs, with obj, $attr$, and $rela$ respectively representing object, attribute, and relation. Each subset D_i corresponds to a fine-grained dataset, which consists of multiple triplets (I, C^i, C^{i-}), where I is the image, C^i is the fine-grained caption, and C^{i-} is the hallucination caption.

3.2 Model Architecture

Our model consists of two parts: fine-grained visual experts and an instruction-aware gate network (Ins-Gate). The former will apply to LVLMs vision encoder and capture different aspects of image representations, while the latter fuses these aspects into a new representation based on instruction.

Fine-grained LoRA visual experts. These three LoRA visual experts are trained on object caption-image pairs, attribute caption-image pairs, and relation caption-image pairs, which are constructed in last section.

Given an input (I, T), with image I and instruction T. The image I is encoded by the original vision encoder from LVLMs into $z_{ori} \in \mathbb{R}^{n \times d}$. At the same time, the image will also be encoded by the other three expert encoders, as shown in Equation 1:

$$z_i = E_i(I), \quad i \in \{obj, attr, rela\} \tag{1}$$

where E_i represents the vision encoder after applying the respective LoRA expert to the original encoder, and $z_i \in \mathbb{R}^{n \times d}$ is the associated fine-grained visual embedding.

Ins-Gate. Ins-Gate contains a cross-attention layer and a feed-forward network (FFN) layer. The instruction embedding t, which is encoded from instruction T, and the concatenated features z_c (representations from three aspects and the original encoder's representations) are used as inputs. The instruction embeddings and image features interact fully in the cross-attention layer, extracting image features relevant to the user's instruction from the four types of image representations. Then

$gate \in \mathbb{R}^{n \times 4}$ is computed using Equation 2.

$$gate = Softmax(FFN(Attention(z_c, t))) \tag{2}$$

The $gate$ provides token-level fusion weights, where each row represents the weights assigned to each image token from the four types of image representations. The feature fusion process is described in Equation 3:

$$f_{fusion} = \sum_{j=1}^{4} gate[:, j] \odot z_j \tag{3}$$

f_{fusion} represents the fused image features, which are concatenated with the text tokens to continue the forward process of the LVLMs.

3.3 Two-Stage Training

The training process is divided into two stages. Throughout the entire training process, we only train PLANE, while all original parameters of the LVLM were frozen. In the training stage I, different fine-grained LoRA visual experts are trained following the CLIP training paradigm (Radford et al., 2021). In the training stage II, PLANE will be integrated into the LVLM, keeping only the Ins-Gate trainable, following the end-to-end LLaVA training paradigm (Liu et al., 2024b).

Training Stage I At this stage, the fine-grained image-caption pairs dataset D is used to train the LoRA visual experts. Each of the three LoRA experts is trained on D_{obj}, D_{attr}, and D_{rela}, respectively, to enhance the image encoder's ability to capture aspect-specific features. Following prior work on VLMs (Radford et al., 2021; Urbanek et al., 2024), we employ contrastive learning loss $\mathcal{L}_{CL}$ to train fine-grained image encoders. For a batch of N image-caption pairs from dataset D_i, $\mathcal{L}_{CL}$ combines bidirectional image-to-text and text-to-image contrastive losses using cosine similarity sim and temperature parameter τ.

Inspired by previous work (Yuksekgonul et al., 2022; Urbanek et al., 2024), we incorporate a negative loss $\mathcal{L}_{NL}$ to enhance the vision-language model's discriminative capability, where hallucination captions serve as carefully designed negative

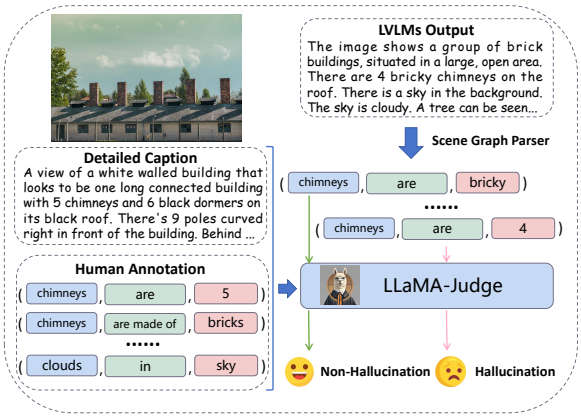


Figure 4: Pipeline of GIFT. The LVLM output is fed into a scene graph parser to extract triplets, which are then input into Llama-Judge along with the detailed caption and manually annotated triplets for hallucination detection.

samples for each aspect:

$$\mathcal{L}_{NL} = -\frac{1}{N} \sum_{k=1}^{N} \log \Bigg(\exp\left(sim(I_k^i, C_k^i)/\tau\right) \times \Big[\sum_{C^{i*} \in \{C^i, C^{i-}\}} \exp\left(sim(I_k^i, C^{i*})/\tau\right) \Big]^{-1} \Bigg) \tag{4}$$

Here, (I_k^i, C_k^i) represents an image-caption pair from D_i, and C_k^{i-} is the corresponding hallucination caption. The final training loss combines both objectives:

$$\mathcal{L}_{final} = \alpha \mathcal{L}_{CL} + \beta \mathcal{L}_{NL} \tag{5}$$

where α and β are hyperparameters.

Training Stage II At this stage, only the Ins-Gate is trainable, while all other modules remain frozen. We use the LRV-instruction dataset (Liu et al., 2023) to train the Ins-Gate with the autoregressive loss (Liu et al., 2024b).

4 GIFT : GPT4-Independent Fine-Grained Hallucination Evaluation Benchmark

As discussed in related work, there is a lack of fine-grained hallucination evaluation benchmarks for generation tasks that can function independently of GPT-4. Therefore, we propose a GPT4-independent fine-grained benchmark (during evaluation phase) to comprehensively assess our models. We build our benchmark based on the DCI dataset (Urbanek et al., 2024), which consists of manually

annotated detailed text-image pairs. We select 206 images from its validation and test sets, and manually annotate 2,120 scene graph triplets (including object, attribute,and relation) aiming to cover as many image details as possible.

The evaluation pipeline of GIFT is shown in Figure 4, LVLMs are first prompted with 'Please describe the image in as much detail as possible' to generate captions for each image. Then, scene graph parser (Li et al., 2023c) is used to extract triplets contained in the captions. To determine whether these triples are hallucinating compared to manually annotated triples, we train a model called Llama-Judge based on Llama 3 (Dubey et al., 2024). These hallucinations may include fabricating non-existent objects, incorrect attributes, or relationships, which help determine if hallucination occurs. We introduce two metrics, Hallu Rate and Cover Rate, to measure the extent of hallucination and level of detail in model-generated content. The definition of Hallu Rate is shown in Equation 6:

$$\text{Hallu Rate} = \frac{H}{M} \times 100\% \tag{6}$$

H represents the number of hallucinated triplets detected by Llama-Judge, M represents the total number of triplets extracted from LVLMs' generation. The definition of Cover Rate is shown in Equation 7:

$$\text{Cover Rate} = \frac{NH}{K} \times 100\% \tag{7}$$

NH represents the number of triplets judged as non-hallucinated, and K represents the number of manually annotated triplets. More details about GITF are presented in the Appendix A.4.

5 Experimental Setup

5.1 Datasets and Implementation Details

We evaluate our model on discriminative tasks and open-ended generation tasks. For discriminative tasks, we select **MME** (Fu et al., 2026) and **POPE** (Li et al., 2023b) (the experimental results for POPE are provided in the Appendix A.6). For open-ended generation tasks, we select **AMBER** (Wang et al., 2023) and **GIFT**. More details about the benchmarks and metrics are provided in the Appendix A.2.

5.2 Baselines and Implementation Details

We evaluate PLANE on several LVLMs of different sizes: LLaVA-v1.5 7B (Liu et al., 2024b), Mini-Gemini 7B (Li et al., 2025c), Parrot 7B (Sun et al., 2025), LLaVA 13B (Liu et al., 2024c), Otter 7B (Li et al., 2025a), and Vary Toy 1.8B (Wei et al., 2024). Our model is trained on LLaVA-v1.5 7B, and after training, it is directly applied to other LVLMs in a plug-and-play manner. Furthermore, we conduct comprehensive comparisons between our proposed method and current advanced approaches in both categories: training-free models and those requiring fine-tuning. Among training-free methods, we benchmark against VCD (Leng et al., 2024), DROPOUT DECODING (Fang et al., 2026) and VAF (Yin et al., 2025). For approaches requiring training, we compare with OHD-Caps (Liu et al., 2024d), HADPO (Zhao et al., 2023), and HALVA (Sarkar et al., 2025) to evaluate our method's effectiveness. All experiments are conducted on two NVIDIA A800(80GB) GPUs.

In training stage I, we train visual experts on two vision encoders CLIP-ViT-L-224 (which is used for Otter, Vary Toy, LLaVA) and CLIP-ViT-L-336 (which is used for LLaVA-1.5, Mini-Gemini, Parrot). We set the LoRA rank to 32, the batch size to 32, and both alpha and beta to 1, running for up to 10 epochs. In training stage II, we integrate our approach into LLaVA-1.5 to train the Ins-Gate. The batch size is 128, running for up to 2 epochs. For evaluation, we set the temperature to 0.5 for all methods. For a more convincing comparison, we reproduce all the baselines based on the officially released code and run each experiment over 3 times, reporting the average results.

6 Experiment

6.1 Results on MME

Table 1 presents the results on MME, showing a significant improvement in the overall score after incorporating our method compared to the original models. For each subtask, our approach effectively mitigates different types of hallucinations, which can be attributed to use of fine-grained image-caption pairs to train visual experts that capture different aspects of the information. Compared to existing hallucination mitigation methods, PLANE shows superior performance and remains effective across models of varying sizes and types. Notably, we train PLANE only on LLaVA-v1.5 and it serves as a plug-and-play module for other LVLMs using a frozen CLIP-ViT-L encoder, effectively mitigating hallucinations across diverse tasks. As evidenced by the MME full set results in Appendix A.5, PLANE further improves overall performance

Model	Method	Existence↑	Count ↑	Position↑	Color↑	Total↑
LLaVA-1.5 7B	Original	179.00$_{(\pm 2.24)}$	115.00$_{(\pm 15.76)}$	104.67$_{(\pm 20.26)}$	156.71$_{(\pm 4.88)}$	555.38
	OHD-Caps	180.00$_{(\pm 3.54)}$	120.67$_{(\pm 9.47)}$	101.67$_{(\pm 17.36)}$	157.00$_{(\pm 9.08)}$	559.34
	HA-DPO	183.33$_{(\pm 2.89)}$	118.89$_{(\pm 7.88)}$	112.78$_{(\pm 3.47)}$	149.44$_{(\pm 5.85)}$	564.44
	HALVA	181.67$_{(\pm 2.89)}$	116.33$_{(\pm 14.43)}$	**122.33$_{(\pm 1.92)}$**	160.00$_{(\pm 10.00)}$	580.33
	VCD	182.00$_{(\pm 2.74)}$	128.00$_{(\pm 7.58)}$	115.00$_{(\pm 15.76)}$	147.66$_{(\pm 20.26)}$	572.66
	DROPOUT DEC.	183.00$_{(\pm 5.70)}$	121.33$_{(\pm 12.38)}$	119.67$_{(\pm 3.41)}$	148.67$_{(\pm 6.61)}$	572.67
	VAF	183.33$_{(\pm 5.77)}$	**128.33$_{(\pm 7.58)}$**	120.56$_{(\pm 3.47)}$	149.44$_{(\pm 5.09)}$	581.66
	PLANE	**184.00$_{(\pm 4.18)}$**	119.67$_{(\pm 5.05)}$	112.00$_{(\pm 13.10)}$	**166.67$_{(\pm 12.19)}$**	**582.34**
	Δ	**↑5.00**	**↑4.67**	**↑7.33**	**↑9.96**	**↑26.96**
Mini-Gemini 7B	Original	179.44$_{(\pm 11.34)}$	132.22$_{(\pm 11.09)}$	109.44$_{(\pm 7.87)}$	139.44$_{(\pm 9.62)}$	560.54
	VCD	167.77$_{(\pm 11.09)}$	126.11$_{(\pm 23.11)}$	105.00$_{(\pm 2.89)}$	141.67$_{(\pm 17.64)}$	540.55
	PLANE	**180.00$_{(\pm 5.00)}$**	**137.22$_{(\pm 11.09)}$**	**119.45$_{(\pm 13.47)}$**	**149.44$_{(\pm 5.85)}$**	**586.11**
	Δ	**↑0.56**	**↑5.00**	**↑10.01**	**↑10.00**	**↑25.57**
Parrot 7B	Original	183.33$_{(\pm 2.89)}$	138.89$_{(\pm 2.55)}$	137.78$_{(\pm 0.96)}$	166.11$_{(\pm 7.88)}$	626.12
	VCD	**185.00$_{(\pm 0.00)}$**	143.67$_{(\pm 17.04)}$	133.30$_{(\pm 8.72)}$	168.33$_{(\pm 5.77)}$	630.30
	PLANE	**185.00$_{(\pm 5.00)}$**	**144.45$_{(\pm 9.48)}$**	**145.00$_{(\pm 2.89)}$**	**173.33$_{(\pm 2.89)}$**	**647.78**
	Δ	**↑1.67**	**↑5.56**	**↑7.22**	**↑7.22**	**↑21.66**
Otter 7B	Original	163.00$_{(\pm 4.47)}$	102.67$_{(\pm 7.51)}$	66.67$_{(\pm 7.07)}$	113.33$_{(\pm 3.54)}$	445.67
	VCD	170.00$_{(\pm 3.54)}$	96.33$_{(\pm 10.89)}$	67.33$_{(\pm 6.73)}$	114.67$_{(\pm 8.11)}$	448.33
	PLANE	**174.00$_{(\pm 2.23)}$**	**105.00$_{(\pm 2.63)}$**	**70.67$_{(\pm 3.65)}$**	**115.00$_{(\pm 4.57)}$**	**464.67**
	Δ	**↑11.00**	**↑2.33**	**↑4.00**	**↑1.67**	**↑19.00**
Vary Toy 1.8B	Original	176.33$_{(\pm 2.74)}$	86.00$_{(\pm 2.74)}$	84.33$_{(\pm 2.74)}$	87.00$_{(\pm 9.76)}$	433.66
	VCD	176.67$_{(\pm 5.00)}$	93.34$_{(\pm 2.74)}$	85.67$_{(\pm 2.74)}$	94.00$_{(\pm 8.04)}$	449.68
	PLANE	**183.67$_{(\pm 3.20)}$**	**99.00$_{(\pm 2.74)}$**	**85.67$_{(\pm 2.74)}$**	**103.00$_{(\pm 1.90)}$**	**471.34**
	Δ	**↑7.34**	**↑13.00**	**↑1.34**	**↑16.00**	**↑37.68**
LLaVA 13B	Original	116.67$_{(\pm 8.50)}$	68.67$_{(\pm 7.94)}$	56.00$_{(\pm 0.91)}$	58.67$_{(\pm 3.61)}$	300.01
	VCD	117.20$_{(\pm 21.14)}$	61.33$_{(\pm 5.45)}$	50.00$_{(\pm 3.73)}$	**65.33$_{(\pm 3.98)}$**	293.86
	PLANE	**125.00$_{(\pm 11.73)}$**	**72.00$_{(\pm 9.82)}$**	**59.67$_{(\pm 3.98)}$**	63.00$_{(\pm 5.70)}$	**319.67**
	Δ	**↑8.33**	**↑3.33**	**↑3.67**	**↑4.33**	**↑19.66**

Table 1: Results on MME. The best results are shown in bold, and Δ indicates the change in performance between our method compared to the original model. Standard deviation is shown in parentheses.

beyond hallucination reduction.

6.2 Results on GIFT and AMBER

The evaluation results on the GIFT and AMBER benchmarks are shown in Table 2. After integrating PLANE with different LVLMs, we observe a decrease in the Hallu Rate and an increase in the Cover Rate on GIFT, indicating that our method reduces fine-grained hallucinations without sacrificing the level of detail in the generated content. This enhancement is attributed to the three LoRA visual experts, each focusing on distinct visual aspects. Moreover, PLANE achieves a plug-and-play effect in generation tasks—trained only once on LLaVA1.5, it can effectively mitigate hallucinations when incorporated into other models. Experiments on AMBER further demonstrate that our approach achieves superior performance across various hallucination metrics compared to existing training-based and training-free methods. Overall, PLANE effectively reduces hallucinations across a variety of model architectures in generation tasks, demonstrating superior performance in alleviating fine-grained hallucinations compared to other training and training-free methods, which further corroborates the findings on GIFT.

6.3 Further Analysis

6.3.1 Inference Speed

We compare inference time on LLaVA1.5, with results shown in Figure 5. Our method's inference time is close to the original model when generating

Model	Method	GIFT		AMBER		
		Hallu↓	Cover↑	Hal↓	CHAIR↓	Cog↓
LLaVA-1.5	Original	23.39	52.96	36.5	8.1	3.4
	OHD-Caps	22.77	53.06	36.6	7.8	3.5
	HA-DPO	23.43	52.78	35.0	7.7	3.4
	VCD	25.25	53.55	38.1	8.4	4.2
	VAF	23.05	53.74	37.9	8.0	3.9
	PLANE	**20.30**	**54.07**	**33.5**	**7.5**	**3.3**
Mini-Gemini	Original	18.64	83.81	54.3	9.7	4.9
	VCD	22.12	83.86	58.0	10.7	4.9
	PLANE	**17.27**	**83.86**	**51.2**	**8.8**	**4.0**
Otter	Original	21.81	25.56	19.8	8.9	1.5
	VCD	26.97	25.79	24.7	11.3	1.7
	PLANE	**20.18**	**28.24**	**18.5**	**7.8**	**1.3**
Vary Toy	Original	29.71	29.79	39.9	10.0	4.5
	VCD	28.67	**31.69**	40.4	10.2	4.1
	PLANE	**28.31**	30.12	**37.8**	**9.6**	**4.0**

Table 2: Results on GIFT and AMBER.

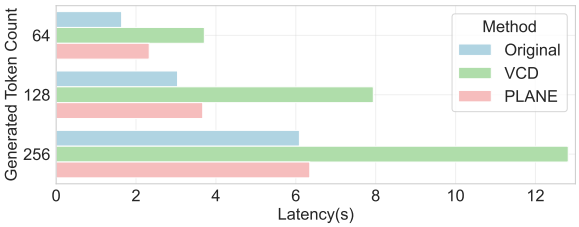


Figure 5: Inference time comparison.

content of different lengths, introducing minimal latency. VCD requires nearly twice the inference time due to its contrastive decoding. This demonstrates our model's superiority in effectively mitigating hallucinations with almost no additional inference overhead.

6.3.2 Ablation Studies

We conduct ablation studies on MME using Vary Toy to evaluate each component's effectiveness. We test three expert encoders (object, attribute, and relation) by incorporating them individually into LVLMs to assess their impact on reducing specific types of multimodal hallucinations. We verify the effect of the Ins-Gate by removing it and taking the average of the image representations as the final image embedding. As shown in Table 3, each specialized encoder improves performance on its corresponding subtask, while also enhancing overall model performance. While direct feature averaging shows notable improvements, optimal results are achieved with the Ins-Gate integration.

Method	Existence↑	Count↑	Position↑	Color↑	Total↑
Original	176.33	86.00	84.33	87.00	433.66
w Object	**186.00**	81.67	84.33	94.67	446.67
w Relation	179.00	97.67	**88.67**	102.67	468.01
w Attribute	175.33	92.00	76.67	**108.33**	452.33
w/o Ins-Gate	180.66	89.00	87.33	106.00	462.99
PLANE	183.67	**99.00**	85.67	103.00	**471.34**

Table 3: Ablation study.

6.3.3 Human Evaluation on GIFT

To validate the reliability of GIFT in evaluating fine-grained hallucinations, we conduct experiments on Vary Toy using greedy decoding. We employ both LLaMA-Judge, GPT4 and human evaluators to determine whether the generated triplets contain hallucinations, measuring both Hallu Rate and Cover Rate (in Table 4). We also report the consistency of LLaMA-Judge and GPT4 with human evaluations, calculated as the proportion of triplets where judgment results align with human annotations. Our trained LLaMA-Judge achieves 83.70% consistency with human evaluators when assessing hallucinations in triplets, significantly outperforming GPT4's 71.81%, which further confirms GIFT's reliability in assessing multimodal hallucinations. It's also worth noting that regardless of which model is used as the judge, PLANE consistently reduces hallucination rates and improves coverage, demonstrating PLANE's excellent performance in mitigating hallucinations.

	Human		LLaMA-Judge			GPT4		
Method	Hallu↓	Cover↑	Hallu↓	Cover↑	CWH↑	Hallu↓	Cover↑	CWH↑
Original	33.99	28.04	32.78	28.56	**83.70**	54.82	19.20	71.81
PLANE	**30.66**	**28.09**	**28.97**	**28.84**		**50.41**	**20.14**	

Table 4: Human Evaluation on GIFT. CWH (Consistency with Human) measures the percentage of triplets where model judgments match human annotations.

7 Conclusion

This paper proposes a plug-and-play fine-grained vision encoder to mitigate multimodal hallucination, consisting of fine-grained LoRA visual experts and an instruction-aware gate network. The former encodes images at object-aspect, attribute-aspect, and relation-aspect respectively, while the latter fuses multiple image features based on user instructions. We also construct a fine-grained benchmark which is GPT4- independent during

evaluation phase for LVLMs. The experimental results on popular benchmarks and GIFT demonstrate the effectiveness of our proposed method. Additionally, Appendix evaluation shows most mainstream LVLMs exhibit high hallucination and low coverage rates on GIFT, demonstrating it as a challenging benchmark that can advance the LVLMs community.

Limitations

The proposed fine-grained open-ended hallucination benchmark (GIFT) remains limited in scale. Given that fine-grained image annotation is highly labor-intensive and time-consuming, scaling up this benchmark is designated as a key objective for future work.

Current mitigation methods primarily focus on single-image inputs. However, multi-image scenarios (Yin et al., 2026b; Li et al., 2025b) are also prevalent in real-world applications; therefore, future work will focus on mitigation strategies for multi-image inputs. Beyond that, we will also investigate hallucination in a wider range of application scenarios (Zhu et al., 2025; Yang et al., 2024b).

Acknowledgments

This research is funded by the National Nature Science Foundation of China (No.62477010, No.62577022 and No.62307028), Shanghai General AI Foundation Models Program (No.2025SHZDZX026D24) and Shanghai Science and Technology Innovation Action Plan (No.24YF2710100).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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*, 1(2):3.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, and Zheng Feng. 2025a. Enhancing uncertainty modeling with semantic graph for hallucination detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23586–23594.

Qin Chen, Jianghao Yin, Lang Yu, Jie Zhou, and Liang He. 2025b. Multi-melo: Unified multimodal model editing with dynamic lora. *Expert Systems with Applications*, 273:126766.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

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

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. 2026. Enhancing vision-language model reliability with uncertainty-guided dropout decoding. *Advances in Neural Information Processing Systems*, 38:149193–149218.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2026. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. 2025. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Seongyun Lee, Sue Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557.

Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. 2025b. Mihbench: Benchmarking and mitigating multi-image hallucinations in multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3143–3152.

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.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2025c. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023c. Factual: A benchmark for faithful and consistent textual scene graph parsing. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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.

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

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. 2024d. Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18288–18301.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2025. Data-augmented phrase-level alignment for mitigating object hallucination. In *The Thirteenth International Conference on Learning Representations*.

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.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. 2025. Parrot: Multilingual visual instruction tuning. In *Forty-second International Conference on Machine Learning*.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Small language model meets with reinforced vision vocabulary. *arXiv preprint arXiv:2401.12503*.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37:92012–92035.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Quanwei Yang, Jiazhi Guan, Kaisiyuan Wang, Lingyun Yu, Wenqing Chu, Hang Zhou, ZhiQiang Feng, Haocheng Feng, Errui Ding, Jingdong Wang, et al. 2024b. Showmaker: Creating high-fidelity 2d human video via fine-grained diffusion modeling. *Advances in Neural Information Processing Systems*, 37:51039–51062.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14625–14634.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. 2026a. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2602.21704*.

Jianghao Yin, Qingbin Li, Kun Sun, Cheng Ding, Jie Wang, Qin Chen, Jie Zhou, Nan Wang, Changqing Li, Pei Wu, et al. 2026b. Mimic human cognition, master multi-image reasoning: A meta-action framework for enhanced visual understanding. *arXiv preprint arXiv:2601.07298*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Fan Yuan, Chi Qin, Xiaogang Xu, and Piji Li. 2024. Helpd: Mitigating hallucination of lvlms by hierarchical feedback learning with vision-enhanced penalty decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1768–1785.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11766–11781.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov,

Louis-Philippe Morency, Katia P Sycara, and Yaqi Xie. 2025. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Findings of the association for computational linguistics: ACL 2024*, pages 12401–12430.

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*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

A Appendix

A.1 Fine-Grained Image-Caption Training Dataset

The statistical information of the fine-grained image-caption pairs we constructed is shown in 5.

Aspect	Count
Image-Object Caption Pairs	7800
Image-Attribute Caption Pairs	18121
Image-Relation Caption Pairs	12454
Total	38375

Table 5: Fine-grained image-caption dataset statistical information.

A.2 Details about benchmarks

We conduct a comprehensive evaluation of our approach on both discriminative and generative tasks. In this section, we will introduce the datasets, metrics, training details, and baselines used in our experiments.

MME (Fu et al., 2026) is a comprehensive multimodal LVLMs benchmark consisting of 14 subtasks which are discriminative task, the LVLMs need to respond with either 'yes' or 'no'. Following previous work (Yin et al., 2024; Leng et al., 2024), we use a subset of these tasks, including *existence*, *count*, *color*, and *position* subsets, as the hallucination evaluation set, referred to as MME-Hallu. We evaluate the hallucination mitigation performance of our model on MME-Hallu, as well as its overall capabilities on the MME-Full set. In line with (Fu et al., 2026), we employ the sum of accuracy and accuracy+ as the evaluation metrics.

POPE (Li et al., 2023b) is a VQA benchmark for evaluating object hallucination. The benchmark incorporates three sampling strategies of varying difficulty levels. The first is the random strategy, which involves randomly sampling nonexistent objects. The second is the popular strategy, focusing on selecting frequently occurring objects. The third is the adversarial strategy, which entails choosing objects that frequently co-occur with those present in the image. We report Accuracy, Precision, Recall, F1 Score as the evaluation metrics.

AMBER (Wang et al., 2023) includes both discriminative and generative tasks. We select the generative task for evaluation, in which the LVLMs is required to generate descriptions for images, and evaluated using three metrics: *CHAIR*, *Hal*, and *Cog*. *CHAIR* measures frequency of non-existent objects in responses, *Hal* calculates the proportion of responses with hallucinations, *Cog* compares model hallucinations to human cognitive patterns.

GIFT is a GPT4-independent fine-grained benchmark hallucination evaluation introduced in this paper.

A.3 Details about baselines

We conduct comprehensive comparisons between our proposed method and current state-of-the-art approaches in both categories: training-free models and those requiring fine-tuning. Among training-free methods, we benchmark against VCD (Leng et al., 2024), DROPOUT DECODING (Fang et al., 2026) and VAF (Yin et al., 2025). For approaches requiring training, we compare with OHD-Caps (Liu et al., 2024d), HADPO (Zhao et al., 2023), and HALVA (Sarkar et al., 2025) to thoroughly evaluate our method's effectiveness.

- VCD (Leng et al., 2024) is a contrastive decoding method that mitigates hallucinations by determining the final output distribution through the combined outputs of both the perturbed image and the original image.
- DROPOUT DECODING (Fang et al., 2026) measures uncertainty in visual inputs and selectively dropping out visual tokens where uncertainty is high.
- VAF (Yin et al., 2025) reduces hallucinations by amplifying visual features during modality fusion, decreasing reliance on language priors while maintaining inference speed and generation quality, ensuring better consistency between visual inputs and generated content.
- OHD-Caps (Liu et al., 2024d) uses the OHD-Caps training set to train a CLIP model with reduced hallucination as the vision encoder for LVLMs. While this method was not given a specific name in the paper, we refer to it as OHD-Caps in this work.
- HA-DPO (Zhao et al., 2023) mitigates hallucinations by training LVLMs to prefer non-

hallucinatory outputs through direct preference optimization.

- HALVA (Sarkar et al., 2025) uses data-augmented phrase-level alignment loss to train LVLMs to reduce the likelihood of hallucinated phrases compared to the correct ones.

A.4 Details about GIFT

A.4.1 Evaluating LVLMs on GIFT

We evaluate mainstream LVLMs on the GIFT benchmark in Table 7. Our experimental results show that most models exhibit high hallucination rates and low coverage scores. Even the state-of-the-art GPT-4V demonstrates significant hallucination tendencies. These findings suggest that GIFT serves as a challenging hallucination evaluation benchmark, providing valuable insights and directions for the future development of the LVLMs research community.

A.4.2 Comparison with Existing Benchmarks

Table 8 shows the comparison between our proposed GIFT and existing generative hallucination evaluation benchmarks. We compare four dimensions: whether object-level, attribute-level, and relation-level hallucinations are considered, and whether ChatGPT or GPT-4 assistance is required for judgment. We can see that CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023) (here we only discuss the generation task in AMBER), and GAVIE (Liu et al., 2023) mainly evaluate object-level hallucinations. Faithscore (Jing et al., 2024) can evaluate hallucinations at multiple levels but requires ChatGPT assistance for judgment, which increases costs and affects its scalability. Our GIFT can detect fine-grained hallucinations while not requiring assistance from ChatGPT/GPT-4 during the evaluation phase.

A.4.3 Prompt for LLaMA-Judge

The prompt for LLaMA-Judge is shown in Figure 8, where the detailed caption in DCI (Urbanek et al., 2024) and manually annotated scene graph triples are concatenated to 'Detailed caption' and 'Scene graph' respectively. The triples extracted from LVLMs output are concatenated after 'Triples to be judged' .

A.4.4 LLaMA-Judge Training Details

To train a model that can detect hallucinations in extracted triples without relying on ChatGPT/GPT-4, we train Llama-Judge based on Llama 3 (Dubey et al., 2024). First, we construct training data by instructing GPT-4o to generate triples both with and without hallucinations. We use uniform sampling with a positive-to-negative sample ratio close to 1:1. For training, we use LoRA fine-tuning with the rank of 8 and a learning rate of 5.0×10^{-5}.

A.5 Results on MME Full Set

We conduct experiments on the MME-Full set using LLaVA-1.5 and Vary Toy, with results shown in Figure 6. After incorporating PLANE, different models show varying trends across different subtasks, which may be attributed to differences in model architectures and training data. However, the implementation of PLANE leads to performance enhancements across the majority of subtasks compared to the baseline. Overall, our model not only excels at hallucination mitigation but also enhances performance on general tasks compared to the original model.

A.6 Results on POPE

We evaluate our method on POPE using LLaVA v1.5, and the experimental results are presented in Table 6. Our method yields improvements under the three settings of Random, Popular, and Adversarial. Compared with existing methods, it achieves the best Accuracy and, in most cases, the best F1 score as well.

A.7 Parameters Efficiency Analysis

In this section, we compare the trainable parameter counts with methods that require training. The experimental results are shown in Table 9. PLANE has fewer parameters than the other two training-required methods, HA-DPO and HALVA, while achieving better performance (Table 1).

A.8 Response Length Analysis

To demonstrate that the hallucination reduction on GIFT does not arise from shortening the model's output, we evaluate the output length at the triplet level. We conduct experiments on LLaVA-1.5 and Vary Toy, using GIFT's scene graph parser to extract the triplets from the model outputs. The experimental results are shown in Table 11. We can observe that PLANE barely changes the model's output length, with only negligible differences from the original model. This confirms that the reduction in hallucinations does not come from shortening the model's output length.

Dataset	Setting	Method	Accuracy ↑	Precision	Recall	F1 Score ↑
A-OKVQA	Random	Regular	83.45	87.24	78.36	82.56
		VCD (Leng et al., 2024)	86.15	85.18	87.53	86.34
		RITUAL (Woo et al., 2024)	85.17	79.79	-	86.40
		DeGF (Zhang et al., 2025)	86.93	84.28	-	87.42
		Ours	**88.93**	95.13	82.07	**88.11**
	Popular	Regular	79.90	80.85	78.36	79.59
		VCD (Leng et al., 2024)	81.85	78.60	87.53	82.82
		RITUAL (Woo et al., 2024)	78.83	71.99	-	81.68
		DeGF (Zhang et al., 2025)	80.90	75.68	-	82.66
		Ours	**85.3**	87.79	82.00	**84.79**
	Adversarial	Regular	74.04	72.08	78.49	75.15
		VCD (Leng et al., 2024)	74.97	70.01	87.36	77.73
		RITUAL (Woo et al., 2024)	68.57	62.26	-	74.99
		DeGF (Zhang et al., 2025)	72.70	66.70	-	76.86
		Ours	**78.63**	76.79	82.07	**79.34**
GQA	Random	Regular	83.73	87.16	79.12	82.95
		VCD (Leng et al., 2024)	86.65	84.85	89.24	86.99
		RITUAL (Woo et al., 2024)	86.10	80.30	-	87.31
		DeGF (Zhang et al., 2025)	87.40	83.51	-	**88.09**
		Ours	**87.93**	94.31	80.73	87.00
	Popular	Regular	78.17	77.64	79.12	78.37
		VCD (Leng et al., 2024)	80.73	76.26	89.24	**82.24**
		RITUAL (Woo et al., 2024)	74.80	67.50	-	79.15
		DeGF (Zhang et al., 2025)	78.10	71.56	-	80.98
		Ours	**81.80**	82.63	80.53	81.57
	Adversarial	Regular	75.08	73.19	79.16	76.06
		VCD (Leng et al., 2024)	76.09	70.83	88.75	78.78
		RITUAL (Woo et al., 2024)	68.23	61.75	-	75.10
		DeGF (Zhang et al., 2025)	74.07	67.42	-	78.22
		Ours	**79.43**	78.72	80.67	**79.68**

Table 6: Results on POPE. Best results are in bold, and second-best values are underlined.

Model	Hallu Rate↓	Cover Rate↑
Vary Toy 1.8B	29.7	29.8
LLaVA 7B	19.0	53.2
LLaVA 13B	19.4	53.4
LLaVA-1.5 7B	23.4	53.0
QwenVL 7B	14.3	21.4
Cambrian 8B	19.3	66.0
LLaVA-OneVision 7B	17.7	72.3
GPT4V	20.9	81.1

Table 7: Results on GIFT. The best results are shown in bold.

Benchmark	Object	Attribute	Relation	GPT Free
CHAIR	✓	✗	✗	✓
AMBER	✓	✗	✗	✓
GAVIE	✓	✗	✗	✗
Faithscore	✓	✓	✓	✗
GIFT	✓	✓	✓	✓

Table 8: Comparison between GIFT and existing generative hallucination benchmarks. Object, Attribute, and Relation denote coverage of corresponding hallucination types, while GPT-free indicates whether the evaluation runs without relying on ChatGPT/GPT-4.

A.9 Necessary of Separate Experts

To verify the necessity of the separate experts design in PLANE, we combine all training tree cap-

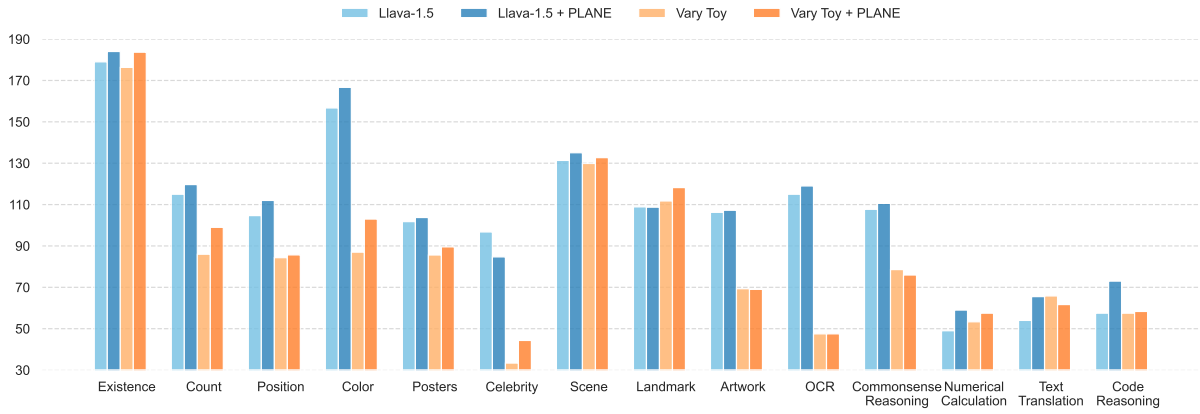


Figure 6: Results on MME-Full set.

Method	Trainable Parameters
HA-DPO	340.8 M
HALVA	340.8 M
PLANE	**332.4 M**

Table 9: Trainable Parameters Comparison.

Method	Existence↑	Count↑	Position↑	Color↑	Total↑
Original	176.33	86.00	84.33	87.00	433.66
All in one	165.00	90.00	86.67	**103.89**	445.56
PLANE	**183.67**	**99.00**	**85.67**	103.00	**471.34**

Table 10: Ablation on the necessary of Separate Experts.

tion types into a single expert (denoted as "All in one"). The experimental results are shown in Table 10. "All in one" underperforms compared to PLANE across most subtasks, which demonstrates the necessity of the separate experts design.

A.10 Extended Ablation Study

In this section, we will explore how different fusion strategies in Ins-Gate affect the results. We compare two fusion methods: one is to fuse the image feature as a whole feature, where $gate \in \mathbb{R}^{1\times4}$, which we denote as Image-Level; the other is to fuse the image feature by token, where $gate \in \mathbb{R}^{n\times4}$ (n represents the number of tokens contained in the image feature), which we denote as Token-Level. We conduct experiments using LLaVA-1.5 and Vary Toy on MME-Hallu, and the experimental results are shown in Table 12. We find that both fusion strategies can mitigate hallucinations compared to the original model, with Token-Level achieving better results.

A.11 Case Study

To more intuitively demonstrate the effects of PLANE, we conducted a case study on GIFT. Figure 7 shows three cases, displaying the output of Vary Toy and the output after incorporating PLANE. And the sentences where the hallucinations occurred are marked in red. We can observe that the original model easily outputs object-level hallucinations and relation-level hallucinations. After incorporating PLANE, there is a great degree of mitigation of fine-grained hallucinations, while also preserving the semantic integrity and coherence. These cases also verify the advantages of PLANE in mitigating hallucinations.

Model	Method	Length (Avg. Triplets)	Hallu Rate (%)	Cover Rate (%)
LLaVA-1.5	Original	7.18	23.39	52.96
	PLANE	7.03	**20.30**	**54.07**
Vary-Toy	Original	5.02	29.71	29.79
	PLANE	5.08	**28.31**	**30.12**

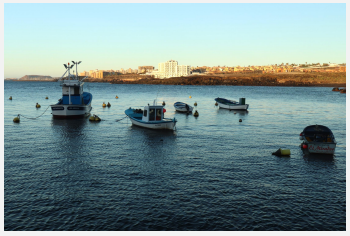
Table 11: Response Length Analysis.

Original model:
The image depicts a man standing among a large pile of garbage, which is piled up on top of a hillside. The man appears to be inspecting or examining the trash, possibly assessing its condition or deciding how to dispose of it. There are several other people present in the scene, scattered around the area, possibly involved in collecting or sorting the garbage. Some of them can be seen near the top of the pile, while others are positioned further down. A bicycle is also visible in the background, likely belonging to one of the individuals responsible for collecting the garbage.

With PLANE:
The image depicts a man standing among a large pile of garbage, which is piled up on top of a hillside. The pile consists of various items such as plastic bottles, bags, and other waste materials. The scene gives off an impression of a landfill or recycling area, where people have been throwing their unwanted belongings into the large bin. In addition to the trash, there are birds flying around the area, adding some life to the otherwise desolate environment. The man appears to be inspecting the pile of garbage, possibly assessing its condition or deciding how to dispose of it.

Original model:
The image features a tall Aldi store sign prominently displayed on the side of its building, with trees in the background. The store is located near a street corner, where there are two traffic lights visible in the scene. These elements create an interesting contrast between the modern signage and the natural surroundings.

With PLANE:
The image features a large, white Aldi sign on the side of a building. The sign is prominently displayed, with the letters \"Aldi\" displayed above the logo. The sign's size and prominent placement make it easily visible to passersby.



Original model:
The image depicts a serene scene of several boats floating on a large body of water, likely a lake or a bay. The boats are anchored close to one another, creating a sense of community among them. Some of the boats are small, while others are larger, varying in size and design. In addition to the boats, there are two birds visible in the sky, adding life to the scene. The overall atmosphere is that of relaxation and leisure, as people enjoy their time on the water with their boats.

With PLANE:
The image depicts a serene scene of several boats floating on a large body of water, likely a lake or a bay. The boats are anchored close to one another, creating a small community of vessels. Some boats are closer to the foreground, while others can be seen further away. There is a mix of different-sized boats, including some very large and others quite small. These boats come in various shapes and sizes, adding diversity to the scene. The overall atmosphere suggests that it could be a popular spot for boating enthusiasts or a relaxing destination for locals seeking a peaceful escape from their daily routines.

Figure 7: Case study on GIFT. The sentences that experience hallucinations are marked in red.

Prompt for LLaMA-Judge

You are an expert at discovering multimodal hallucinations. Multimodal hallucinations refer to the phenomenon that the output of a large multimodal model is inconsistent with the ground truth. This inconsistency may be the fabrication of non-existent objects, or wrong attributes, or wrong positional relationships, or fabricating non-existent triples. I will give you a detailed caption of a picture and the scene graph triple of the picture. Then I will give you a triplet generated by a model, you need to combine the caption and scene graph to determine whether the triplet you need to judge has hallucinations. If so, answer Yes, and if not, answer No. The singular or plural form of a word is not the basis for your judgment. If the triples to be judged have the same meaning but different words, it is not considered as hallucination. Just output the final result "Yes" or "No".
Detailed caption:
Scene graph:
Triples to be judged:
Judgment result:

Figure 8: Prompt for LLaMA-Judge.

Model	Method	Existence↑	Count ↑	Position↑	Color↑	Total↑
LLaVA-1.5 7B	Original	$179.00_{(\pm 2.24)}$	$115.00_{(\pm 15.76)}$	$104.67_{(\pm 20.26)}$	$156.71_{(\pm 4.88)}$	555.38
	Image-Level	$\mathbf{184.00}_{(\pm \mathbf{2.24})}$	$113.34_{(\pm 1.67)}$	$\mathbf{112.33}_{(\pm \mathbf{16.01})}$	$151.33_{(\pm 15.20)}$	561.00
	Token-Level	$\mathbf{184.00}_{(\pm \mathbf{4.18})}$	$\mathbf{119.67}_{(\pm \mathbf{5.05})}$	$112.00_{(\pm 13.10)}$	$\mathbf{166.67}_{(\pm \mathbf{12.19})}$	**582.34**
Vary Toy 1.8B	Original	$176.33_{(\pm 2.74)}$	$86.00_{(\pm 2.74)}$	$84.33_{(\pm 2.74)}$	$87.00_{(\pm 9.76)}$	433.66
	Image-Level	$174.67_{(\pm 3.41)}$	$90.00_{(\pm 5.27)}$	$\mathbf{90.67}_{(\pm \mathbf{5.22})}$	$89.00_{(\pm 2.24)}$	444.34
	Token-Level	$\mathbf{183.67}_{(\pm \mathbf{3.20})}$	$\mathbf{99.00}_{(\pm \mathbf{2.74})}$	$85.67_{(\pm 2.74)}$	$\mathbf{103.00}_{(\pm \mathbf{1.90})}$	**471.34**

Table 12: Comparison of different fusion strategies. The best results are shown in bold. Standard deviation is shown in parentheses.