FADE: Mitigating Hallucinations by Reducing Language Priors Dominance in Large Vision-Language Models

Anonymous ACL submission

Abstract

Despite the impressive capabilities of Large Vision-Language Models (LVLMs), they remain susceptible to hallucination—generating content inconsistent with the input image. Recent studies attribute this to the dominance of language priors over visual inputs and employ contrastive decoding methods to mitigate this dominance, but the mechanistic origin remains unexplored. We investigate the information flow through each transformer layer and find that attention modules consistently aggregate visual evidence, while FFN modules at critical layers act as the source of language priors. These priors can override visual evidence, causing correct predictions in intermediate layers to drift toward incorrect outputs. Based on this insight, we propose **FADE** (**F**FN **A**ttenuation for **DE**coding), a training-free method that attenuates FFN outputs to reduce language dominance. Evaluations on POPE, CHAIR and MME benchmarks across LLaVA-1.5, mPLUG-Owl2 and InstructBLIP show that FADE effectively mitigates hallucinations while preserving inference efficiency.

1 Introduction

Large Vision-Language Models (LVLMs) have achieved remarkable progress in recent years, bridging the gap between vision and language through effective multimodal alignment (Radford et al., 2021; Li et al., 2022, 2023a; Liu et al., 2023b; Chen et al., 2024b; Bai et al., 2023; Wang et al., 2024b). These models have achieved significant success across diverse applications including visual question answering (VQA) and image captioning. However, a persistent challenge remains: LVLMs often generate text that is not consistent with the visual content of the input image, known as hallucination (Li et al., 2023b; Rohrbach et al., 2018; Huang et al., 2024a; Guan et al., 2024). This phenomenon can cause serious risks in critical applications, including medical diagnosis (Liu et al., 2024c), autonomous driving (Cui et al., 2023), and embodied agents (Driess et al., 2023), where precision and reliability are essential.

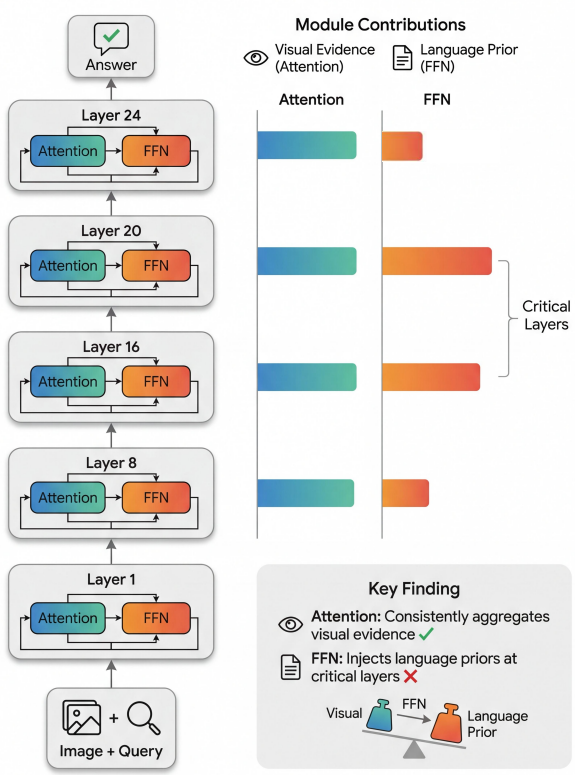


Figure 1: Analyzing information flow through transformer layers. Attention consistently aggregates visual evidence, while FFN at critical layers (16–22) introduces language priors that can override visual evidence.

Recent research on mitigating hallucinations can be divided into two categories. Training-based approaches employ instruction tuning (Liu et al., 2023a), RLHF (Sun et al., 2024a) or DPO (Zhao et al., 2023) to reduce hallucinations at the source, but they require expensive data collection and retraining. Training-free methods intervene during inference without modifying model parameters. Attention modification approaches (Liu et al., 2024d; Huang et al., 2024b) amplify visual token weights to enhance visual grounding. Layer-wise

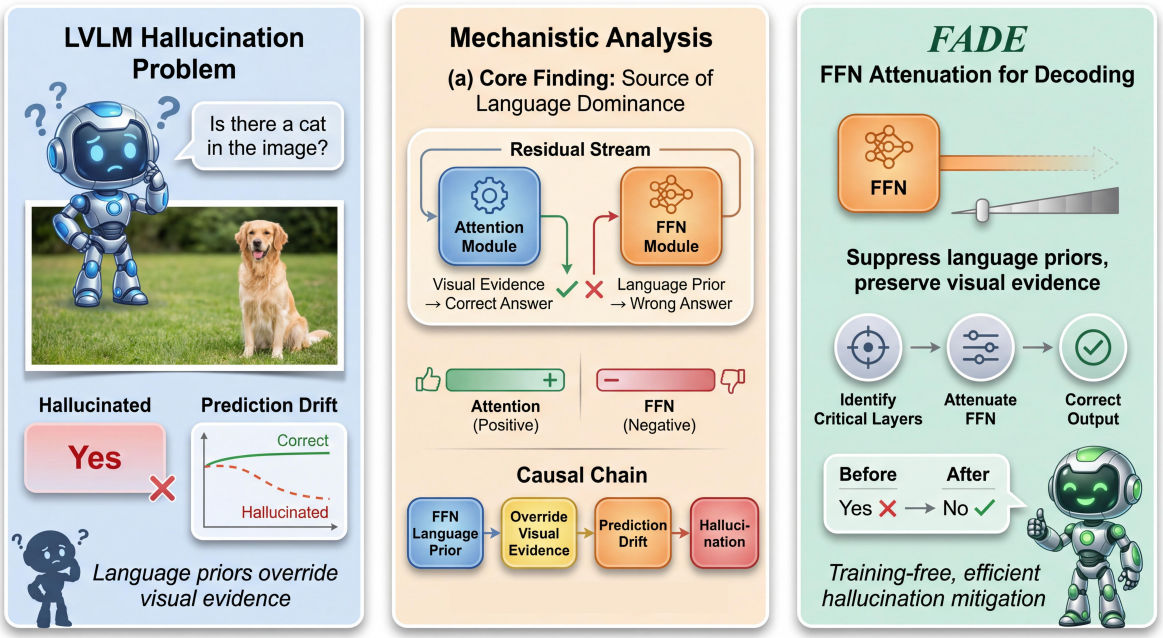


Figure 2: Overview of our approach. **Left:** LVLMs suffer from hallucinations where language priors override visual evidence, causing prediction drift from correct to incorrect outputs. **Middle:** Our mechanistic analysis reveals that attention modules aggregate visual evidence toward correct answers, while FFN modules at critical layers introduce language priors that can override visual evidence. **Right:** FADE attenuates FFN outputs at critical layers to suppress language priors while preserving visual evidence, enabling training-free hallucination mitigation.

intervention methods (Chuang et al., 2024; Wang et al., 2025) exploit cross-layer differences to improve output quality. Contrastive decoding methods (Leng et al., 2024; Manevich and Tsarfaty, 2024) attribute hallucination to the dominance of language priors over visual inputs and attempt to suppress this dominance by contrasting output distributions. However, these methods operate at the output level without understanding where language priors originate within the model. Understanding this origin is crucial for developing more targeted and efficient solutions.

In this work, we investigate the mechanistic origin of language dominance. We decompose transformer computations into attention and FFN contributions using the residual stream perspective (Elhage et al., 2021), and measure their effects on predictions through logit lens projections (Geva et al., 2022; Belrose et al., 2023). As illustrated in Figure 1, our analysis reveals two key findings: (1) *Attention Aggregates Visual Evidence.* Attention mechanisms consistently aggregate visual features to generate correct predictions. (2) *FFN Introduces Language priors.* FFN modules at critical layers act as the source of language priors that can override visual evidence, causing hallucinations.

Based on this insight, we propose **FADE** (**F**FN **A**ttenuation for **DE**coding), a training-free method that attenuates FFN outputs at critical layers to reduce language dominance (Figure 2). By weakening FFN contributions, FADE preserves the visual evidence while suppressing the language priors that cause hallucination. Unlike contrastive decoding methods, FADE operates within a single pass with minimal overhead.

Our contributions can be summarized as follows:

- We conduct a mechanistic analysis revealing the origin of language priors dominance in LVLMs: attention aggregates visual evidence, while FFN at critical layers acts as the source of language priors that can override it.
- We propose FADE, a training-free method that attenuates FFN outputs at critical layers to reduce language dominance while preserving visual evidence.
- Extensive experiments on POPE and CHAIR benchmarks across LLaVA-1.5, mPLUG-Owl2, InstructBLIP and Qwen-VL demonstrate that FADE effectively mitigates hallucinations while maintaining inference efficiency.

2 Related Work

2.1 Large Vision-Language Models

Large Vision-Language Models (LVLMs) have evolved from early BERT-based decoders (Chen et al., 2020; Li et al., 2020; Zhang et al., 2021; Li et al., 2021; Wang et al., 2022; Li et al., 2022) designed to integrate visual and textual information into a paradigm driven by large language models (LLMs) (Touvron et al., 2023a,b; Jiang et al., 2023; Dubey et al., 2024; Yang et al., 2024). The emergence of LLMs has sustainably enhanced the capabilities and performance of LVLMs. In this process, supported by end-to-end training techniques (Alayrac et al., 2022; Dai et al., 2023), LVLMs have achieved unified decoding of visual and textual tokens, indicating that both their expressiveness and adaptability have significantly improved. Recent works, such as LLaVA (Liu et al., 2023b,a, 2024b) and InstructBLIP (Dai et al., 2023), have further refined these models through visual instruction tuning, enhancing their performance in various vision-language tasks. More recently, models such as the Qwen-VL series (Bai et al., 2023; Wang et al., 2024b) and InternVL series (Chen et al., 2024b,a) have further scaled up through improved alignment strategies and large-scale joint training.

2.2 Hallucination Mitigation in LVLMs

Hallucination in LVLMs refers to generating content that is linguistically plausible but inconsistent with visual input (Rohrbach et al., 2018; Li et al., 2023b). Training-based approaches involve additional training to align models with ground truth, including robust instruction tuning (Liu et al., 2024a), post-hoc revision (Zhou et al., 2024), reinforcement learning from human feedback (Yu et al., 2024), and direct preference optimization (Zhao et al., 2023; Wang et al., 2024a). While effective, these methods suffer from extensive training costs and data requirements.

In contrast, training-free methods can be applied directly during inference without modifying model parameters. *Attention-based methods* modify attention patterns to enhance visual grounding. PAI (Liu et al., 2024d) amplifies attention weights on image tokens to prioritize visual information during generation. Other approaches penalize overtrust on summary tokens (Huang et al., 2024b) or fuse global-local attention features (An et al., 2025; Qian et al., 2025). *Contrastive decoding methods* suppress hallucinated content by contrasting output distributions. VCD, LCD (Leng et al., 2024; Manevich and Tsarfaty, 2024) suppresses language priors via output contrast. Similar strategies have been applied at the instruction level (Wang et al., 2024c) or using self-generated descriptions (Kim et al., 2024). *Layer-wise intervention methods* exploit the hierarchical structure of transformers. DAMO (Wang et al., 2025) accumulates activation momentum across layers to stabilize predictions. Related approaches contrast logits from different layers (Chuang et al., 2024) or aggregate representations to enforce inter-layer consistency (Huo et al., 2025; Li et al., 2025a; Tang et al., 2025). *Representation engineering methods* directly manipulate hidden representations using pre-computed steering vectors. VISTA (Li et al., 2025b) introduces visual steering vectors combined with self-logits augmentation from early layers. VTI (Liu et al., 2025) applies intervention vectors computed via PCA on contrastive pairs, while FlexAC (Yuan et al., 2025) controls associative reasoning through middle-layer representations.

Our work takes a mechanistic perspective, investigating where language priors originate within the model. We propose **FADE** (**F**FN **A**ttenuation for **DE**coding), which attenuates FFN outputs at critical layers to reduce language priors dominance.

3 Method

3.1 Preliminaries

A transformer-based LVLM processes inputs through L decoder layers. Each layer l applies attention and FFN with residual connections:

$$\tilde{\mathbf{h}}^{(l)} = \mathbf{h}^{(l)} + \mathrm{Attn}^{(l)}(\mathbf{h}^{(l)}) \quad (1)$$

$$\mathbf{h}^{(l+1)} = \tilde{\mathbf{h}}^{(l)} + \mathrm{FFN}^{(l)}(\tilde{\mathbf{h}}^{(l)}) \quad (2)$$

From the residual stream perspective (Elhage et al., 2021), attention aggregates information across positions while FFN performs per-position transformations. Priors work shows FFN layers function as key-value memories storing factual knowledge (Geva et al., 2021; Meng et al., 2022).

3.2 Motivation: Prediction Drift in LVLMs

We begin by examining how predictions evolve across layers in LVLMs. Using logit lens projections on LLaVA-1.5-7B, we track the probability of correct answer tokens at each layer for samples from POPE-Adversarial.

Figure 3 reveals a striking pattern: for hallucinated samples, predictions drift from high to low P(Correct Answer) in later layers, while correct samples maintain stable high probability throughout. This observation raises a critical question: *what causes this prediction drift?* We address this through mechanistic analysis in the following sections.

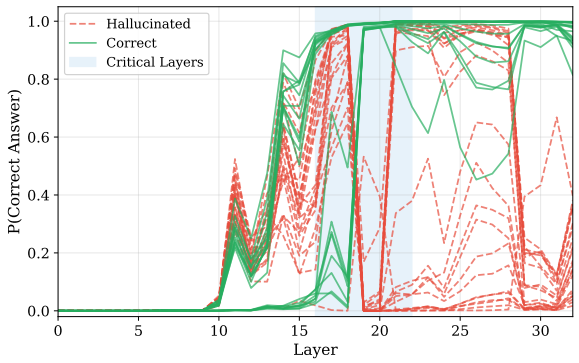


Figure 3: P(Correct Answer) trajectories across layers for hallucinated (red, dashed) and correct (green, solid) samples. Correct samples maintain high probability throughout, while hallucinated samples drift to low probability in later layers. The shaded region indicates critical layers (16–22).

3.3 Mechanistic Analysis

To understand what causes the prediction drift observed in Figure 3, we decompose the contributions of attention and FFN modules at each layer. We analyze LLaVA-1.5-7B on 50 samples from POPE-Adversarial.

Contribution Analysis. To measure each component's contribution, we use a differential logit lens approach. For attention at layer l:

$$\Delta^{(l)}_{\text{Attn}}(t) = \text{LM}_{\text{head}}(\tilde{\mathbf{h}}^{(l)})_t - \text{LM}_{\text{head}}(\mathbf{h}^{(l)})_t \quad (3)$$

where t is the target token. We compute $\Delta^{(l)}_{\text{FFN}}$ analogously. This differential approach accounts for the nonlinearity of layer normalization.

Correct-Direction Metric. To enable comparison across samples with different ground truths, we define a *correct-direction* metric:

$$C^{(l)} = \Delta^{(l)}(t_{\text{correct}}) - \Delta^{(l)}(t_{\text{incorrect}}) \quad (4)$$

Under this metric, $C^{(l)} > 0$ indicates the component pushes toward the correct answer, while $C^{(l)} < 0$ indicates it pushes toward the wrong answer.

Prediction	Attn	$\text{FFN}_{\text{total}}$	$\text{FFN}_{16\text{-}22}$	FFN_{L18}
Correct (n=40)	+1.2	+1.7	**+8.4**	+6.0
Wrong (n=10)	+0.8	−2.0	**−3.5**	−2.4

Table 1: Mean contributions toward correct answer (correct-direction metric). Values are summed across layers and averaged across samples. Positive values indicate pushing toward ground truth. FFN at layers 16–22 shows the largest difference between correct and wrong predictions.

OBS-1: Attention Aggregates Visual Evidence. Attention contributions are positive and comparable for both correct (+1.2) and hallucinated (+0.8) samples (Table 1). This indicates that attention consistently aggregates visual features toward correct predictions across all samples.

OBS-2: FFN Introduces Language priors. In contrast, FFN at layers 16–22 shows a striking difference: +8.4 for correct predictions and −3.5 for wrong predictions. For correct samples, FFN reinforces the prediction; for hallucinated samples, FFN actively pushes toward the wrong answer. We identify FFN as the source of language priors—when these priors conflict with visual evidence, they can override attention's correct predictions.

This directly explains the drift in Figure 3: attention establishes correct predictions in intermediate layers, but language priors from FFN at layers 16–22 override the visual evidence, causing the prediction to drift toward incorrect outputs.

3.4 FADE: FFN Attenuation for Decoding

Based on our analysis, we propose **FADE**, which attenuates FFN outputs at critical layers to reduce language dominance:

$$\mathbf{h}^{(l+1)} = \tilde{\mathbf{h}}^{(l)} + (1-\alpha) \cdot \text{FFN}^{(l)}(\tilde{\mathbf{h}}^{(l)}) \quad (5)$$

where $\tilde{\mathbf{h}}^{(l)}$ is the post-attention hidden state and $\alpha \in [0, 1]$ is the attenuation strength. FADE is applied at layers 16–22 with $\alpha = 0.5$ by default. The method is training-free, requires no additional parameters, and introduces negligible overhead. By reducing FFN contributions, FADE suppresses language priors while preserving visual evidence aggregated by attention.

4 Experiments

4.1 Experimental Setup

Models. We evaluate FADE on three representative LVLMs spanning diverse architectures: **LLaVA-1.5-7B** (Liu et al., 2023a), which uses

Table 2: POPE benchmark results across three VLMs. We evaluate across three sampling strategies (Random, Popular, Adversarial) and three datasets (MSCOCO, A-OKVQA, GQA). Best results are in **bold**, second best are underlined.

		LLaVA-1.5-7B						mPLUG-Owl2-7B						InstructBLIP-7B					
		MSCOCO		A-OKVQA		GQA		MSCOCO		A-OKVQA		GQA		MSCOCO		A-OKVQA		GQA	
Setting	**Method**	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Random	Greedy	88.5	87.3	91.0	90.7	89.3	89.0	88.2	87.4	88.5	88.4	86.9	86.1	87.2	85.8	88.6	88.4	87.4	87.1
	PAI *ECCV'24*	88.5	87.4	91.0	90.7	89.2	88.8	**88.5**	**87.8**	88.4	88.2	86.6	85.7	69.0	73.1	66.2	71.8	64.4	70.6
	VCD *CVPR'24*	**89.9**	**90.0**	88.5	89.4	88.1	89.0	87.9	**88.1**	85.0	85.9	85.2	85.6	89.1	88.6	87.1	87.7	86.3	87.0
	DAMO *ICLR'25*	86.4	84.6	89.0	88.2	86.4	85.2	87.9	87.0	**88.6**	**88.4**	86.2	85.2	88.2	88.1	83.7	85.2	84.4	85.8
	VISTA *ICML'25*	88.8	87.8	**91.4**	**91.3**	**90.0**	**89.8**	88.2	87.4	88.5	88.4	86.7	85.8	88.0	86.9	88.5	88.6	87.7	87.8
	FADE	89.2	88.3	91.2	91.1	89.8	89.7	88.5	87.8	88.6	88.5	87.0	86.2	84.9	83.7	82.1	83.1	80.3	81.3
Popular	Greedy	87.2	86.1	87.6	87.6	**84.5**	84.7	86.2	85.5	84.6	85.0	80.0	80.2	84.8	83.6	81.3	82.2	77.1	78.8
	PAI *ECCV'24*	87.4	86.3	**87.8**	**87.8**	**84.5**	84.7	**86.4**	**85.9**	84.6	85.0	79.9	80.0	65.0	70.6	59.2	66.8	55.7	65.9
	VCD *CVPR'24*	86.6	**87.1**	82.3	84.5	76.6	80.4	83.6	84.4	81.2	82.9	77.5	79.4	**85.2**	85.2	79.9	82.0	**77.1**	80.0
	DAMO *ICLR'25*	85.4	83.7	87.2	86.6	84.2	83.2	**86.4**	85.6	**84.9**	**85.2**	**80.2**	80.0	82.4	83.3	74.7	78.8	71.5	76.8
	VISTA *ICML'25*	87.4	86.5	86.6	87.1	83.2	84.1	86.2	85.6	84.7	85.1	80.0	80.1	85.5	84.6	80.3	81.9	75.8	78.5
	FADE	87.7	86.9	87.0	87.4	84.1	84.8	86.4	85.9	84.5	85.1	79.9	80.2	84.9	83.7	82.1	83.1	**80.3**	81.3
Adversarial	Greedy	85.1	84.2	80.4	81.8	**81.5**	82.3	84.0	83.6	77.6	79.7	78.5	79.0	**83.0**	82.0	74.6	77.3	75.1	77.3
	PAI *ECCV'24*	**85.3**	84.3	80.7	81.9	**81.5**	82.2	**84.2**	**83.9**	77.7	79.7	78.4	78.8	62.8	69.3	55.7	63.0	55.4	65.7
	VCD *CVPR'24*	81.2	82.7	73.3	78.4	72.0	77.5	79.9	81.6	73.7	77.9	75.0	77.9	82.0	**82.5**	72.0	76.6	72.9	**77.2**
	DAMO *ICLR'25*	83.9	82.3	**82.1**	**82.2**	81.4	80.8	**84.2**	83.6	**78.2**	**80.0**	**78.6**	78.7	79.4	81.0	67.3	74.2	68.1	74.7
	VISTA *ICML'25*	85.2	84.5	79.2	81.2	80.4	81.9	84.0	83.6	77.7	79.7	78.5	**78.9**	**83.0**	**82.5**	**73.3**	76.9	**73.6**	76.9
	FADE	85.2	84.6	79.5	81.4	81.1	82.5	84.2	83.9	77.4	79.7	78.3	79.0	**84.9**	**83.7**	**82.1**	**83.1**	**80.3**	**81.3**

a two-stage training with visual instruction tuning; **mPLUG-Owl2-7B** (Ye et al., 2024), which employs modality-adaptive modules for vision-language alignment; and **InstructBLIP-7B** (Dai et al., 2023), which introduces instruction-aware visual feature extraction via Q-Former. This selection covers the major LVLM design paradigms and enables comprehensive evaluation of our method's generalizability.

Benchmarks. We adopt three widely-used benchmarks: **POPE** (Li et al., 2023b) probes object hallucination via binary (Yes/No) questions across three sampling strategies (Random, Popular, Adversarial) on MSCOCO, A-OKVQA, and GQA; **CHAIR** (Rohrbach et al., 2018) measures hallucination in image captioning, where CHAIR_S and CHAIR_I denote sentence-level and instance-level hallucination rates (lower is better) and Recall measures coverage (higher is better); **MME** (Fu et al., 2023) evaluates perception and cognition across 14 subtasks, and we report the perception score focusing on existence, count, position, and color.

Baselines. We compare against representative training-free methods from each category: **PAI** (Liu et al., 2024d) amplifies attention on image tokens; **VCD** (Leng et al., 2024) contrasts outputs from original and distorted images; **DAMO** (Wang et al., 2025) applies momentum-based activation stabilization; and **VISTA** (Li et al., 2025b) steers representations using pre-computed visual vectors. All baselines use official implementations with recommended hyperparameters.

Implementation. All experiments use greedy decoding on 8 NVIDIA H100 80GB GPUs. For FADE, we set attenuation strength $\alpha = 0.6$ and intervene at layer 18 for LLaVA-1.5; model-specific configurations and all baseline hyperparameters are provided in Appendix A.

4.2 Main Results

4.2.1 Results on POPE

Table 2 presents results under random, popular, and adversarial settings. FADE consistently outperforms baseline methods across both LLaVA-1.5 and mPLUG-Owl2, achieving 82.5% F1 on LLaVA-1.5 and 79.0% on mPLUG-Owl2 under the challenging adversarial setting on GQA. On LLaVA-1.5, FADE surpasses VCD by 5.0% F1 and DAMO by 1.7% F1 on the GQA adversarial subset. Notably, VCD shows limited generalization with severe degradation on GQA (72.0% accuracy under adversarial), likely because its contrastive decoding with noisy images disrupts fine-grained spatial reasoning required for GQA's scene graph questions.

Table 3: CHAIR benchmark results across three VLMs. C_S/C_I: sentence/instance-level hallucination rates (lower is better). Rec: recall (higher is better). Best results are in **bold**, second best are underlined.

	LLaVA-1.5-7B				mPLUG-Owl2-7B				InstructBLIP-7B			
Method	$C_S\downarrow$	$C_I\downarrow$	Rec↑	Len	$C_S\downarrow$	$C_I\downarrow$	Rec↑	Len	$C_S\downarrow$	$C_I\downarrow$	Rec↑	Len
Greedy	49.8	14.8	80.6	101.2	57.8	17.1	78.6	105.6	63.4	38.5	**73.7**	101.6
PAI *ECCV'24*	35.6	9.8	74.8	107.6	57.4	**14.5**	**79.7**	105.7	**48.6**	37.9	58.1	68.8
VCD *CVPR'24*	58.6	16.5	**82.1**	105.5	64.4	18.1	77.9	110.4	57.2	40.1	63.5	87.8
DAMO *ICLR'25*	56.6	16.7	81.6	106.7	58.6	17.5	77.3	106.4	65.6	39.5	73.5	104.7
VISTA *ICML'25*	**19.2**	**6.5**	62.6	86.2	69.8	42.1	78.7	105.5	54.0	41.3	60.4	85.6
FADE	46.6	14.1	78.7	98.6	**55.0**	16.6	76.3	105.4	49.2	**14.0**	72.9	99.8

DAMO and VISTA improve over greedy decoding but exhibit inconsistent behavior—DAMO gains on A-OKVQA but plateaus on GQA, while VISTA shows marginal improvements that do not consistently exceed the baseline across all settings.

4.2.2 Results on CHAIR

Table 3 reports image captioning results. Among existing methods, we observe a clear accuracy-coverage trade-off: VISTA achieves the lowest $CHAIR_S$ (19.2%) on LLaVA-1.5 but sacrifices Recall significantly (62.6% vs. 80.6% for greedy), indicating over-aggressive suppression of generation. Conversely, VCD and DAMO increase hallucination rates on most models—VCD raises $CHAIR_S$ from 49.8% to 58.6% on LLaVA-1.5, suggesting that their uniform intervention strategies disrupt fluent generation.

FADE demonstrates consistent improvements across all three models. On mPLUG-Owl2, FADE achieves the lowest $CHAIR_S$ (55.0%) among all methods. Most notably, on InstructBLIP, FADE dramatically reduces instance-level hallucination with $CHAIR_I$=14.0%, compared to 37.9% for PAI and 38.5% for greedy—a 2.7× reduction while maintaining competitive Recall (72.9%). This suggests that FFN attenuation is particularly effective for models with Q-Former-based visual encoding.

4.2.3 Results on MME

Table 5 reports MME perception scores across 10 subtasks. On LLaVA-1.5, FADE achieves 1519.0 total perception score, improving over greedy decoding (1505.7) by +13.3 points and outperforming all baselines including PAI (1508.9). The improvement is particularly notable on counting (+5.0 over greedy) and celebrity recognition (+1.7), subtasks that require precise object grounding. Interestingly, different methods show architecture-dependent behavior: PAI improves LLaVA-1.5 (+3.2) but slightly degrades mPLUG-Owl2 (−15.6), while DAMO gains significantly on mPLUG-Owl2's counting subtask (+10.0) but loses on LLaVA-1.5 (−6.7). This architecture sensitivity suggests that attention-based and contrastive methods may interact differently with each model's vision-language alignment mechanism. FADE's FFN-level intervention provides a more architecture-agnostic approach by targeting the representation drift phenomenon that is common across transformer-based LVLMs.

Table 4: MMHal-Bench results on LLaVA-1.5-7B. Scores range 0-4 (higher is better). GPT-4 evaluates both hallucination rate and informativeness.

Method	Attr.	Adv.	Affd.	Count	Spat.	Scene	OCR	Celeb.	Overall
Greedy	2.25	**1.33**	2.92	1.75	1.92	3.25	**1.58**	1.42	2.05
PAI	1.83	0.75	2.33	**2.00**	1.92	2.17	1.25	**2.42**	1.83
VCD	1.83	0.83	1.83	1.17	1.67	**4.17**	**1.58**	2.25	1.92
DAMO	**2.58**	**1.33**	**3.00**	1.75	1.92	3.42	1.17	1.42	2.07
FADE	2.42	1.25	2.83	1.75	**2.08**	3.25	1.42	1.75	**2.09**

4.2.4 Results on MMHal-Bench

We further evaluate on MMHal-Bench, where GPT-4 judges open-ended responses across eight question categories including object attributes, counting, spatial relations, and environment description. This benchmark tests whether hallucination mitigation methods can generalize beyond binary Yes/No questions to free-form generation. priors work (Sun et al., 2024b) shows that methods effective on POPE may not transfer to open-ended settings, as the generation dynamics differ significantly. Table 4 shows that the relative ranking of methods is largely consistent with POPE, though the absolute improvements are more modest due to the increased task complexity.

4.3 Efficiency Study

We analyze FADE's computational efficiency compared to existing methods.

Table 5: MME perception scores across 10 subtasks on LLaVA-1.5-7B and mPLUG-Owl2-7B. Higher is better. **Bold**: best per model. Underline: second best.

Model	Method	Exist.	Count	Pos.	Color	Poster	Celeb.	Scene	Land.	Art	OCR	Total
LLaVA-1.5	Greedy	**190.0**	155.0	128.3	**170.0**	**147.6**	136.8	158.0	163.0	119.5	137.5	1505.7
	PAI *ECCV'24*	**190.0**	155.0	**133.3**	**170.0**	145.6	136.5	157.8	163.0	117.8	**140.0**	1508.9
	VISTA *ICML'25*	**190.0**	150.0	**133.3**	165.0	144.6	134.4	156.0	163.3	115.0	125.0	1476.6
	DAMO *ICLR'25*	**190.0**	148.3	**133.3**	160.0	136.4	131.8	**159.0**	162.0	113.5	**140.0**	1474.3
	FADE	**190.0**	**160.0**	**133.3**	**170.0**	**147.6**	**138.5**	158.0	**163.8**	**120.3**	137.5	**1519.0**
mPLUG-Owl2	Greedy	**185.0**	160.0	**85.0**	150.0	160.2	163.5	152.8	163.3	137.3	**102.5**	1459.5
	PAI *ECCV'24*	**185.0**	155.0	81.7	145.0	158.2	**163.8**	154.0	160.3	138.5	**102.5**	1443.9
	VISTA *ICML'25*	**185.0**	155.0	80.0	150.0	158.2	161.8	153.5	159.5	**140.5**	**102.5**	1445.9
	DAMO *ICLR'25*	**185.0**	**170.0**	78.3	150.0	**164.6**	160.9	**156.0**	**170.5**	130.5	95.0	1460.8
	FADE	**185.0**	160.0	**85.0**	**155.0**	160.2	163.5	153.5	161.8	137.3	**102.5**	**1463.7**

Table 6 compares inference efficiency. FADE adds only 3% latency overhead compared to greedy decoding (122ms vs 118ms), while achieving significant speedups over all comparison methods: 19% faster than DAMO, 34% faster than PAI, 57% faster than VCD, and 73% faster than VISTA. VCD requires a second forward pass with distorted images, resulting in 2.4× total latency. VISTA incurs the highest overhead (3.9×) due to steering vector computation during inference. FADE's efficiency stems from: (1) FFN attenuation requiring only element-wise scaling at a single layer, not additional forward passes; and (2) no memory overhead (14.5 GB, identical to greedy decoding).

Table 6: Inference efficiency comparison on LLaVA-1.5-7B. Measured on POPE (500 samples) with H100 GPU.

Method	Prefill (ms/tok)	Decode (ms/tok)	Latency (ms)	Memory (GB)
Greedy	0.08	67.72	118	14.5
VCD *CVPR'24*	0.15	188.47	285	14.0
PAI *ECCV'24*	0.11	111.71	184	14.5
DAMO *ICLR'25*	0.10	88.24	150	14.6
VISTA *ICML'25*	0.34	239.28	459	14.5
FADE	**0.08**	**70.86**	**122**	14.5

4.4 Case Study

Figure 4 presents qualitative examples demonstrating FADE's effectiveness in correcting hallucinations. In Case 1, when asked which cat opens its mouth, greedy decoding incorrectly answers "the middle cat," while FADE correctly identifies "the cat on the right." This illustrates FADE's ability to correct spatial reasoning errors caused by language priors dominance.

In Case 2, greedy decoding hallucinates a dog in a skiing scene where no animal is present, responding "Yes, there is a dog running beside the skier." FADE correctly answers "No, there is no dog in this image," demonstrating its effectiveness in suppressing object hallucinations. These examples illustrate how FFN attenuation reduces language priors dominance, enabling more visually grounded responses.

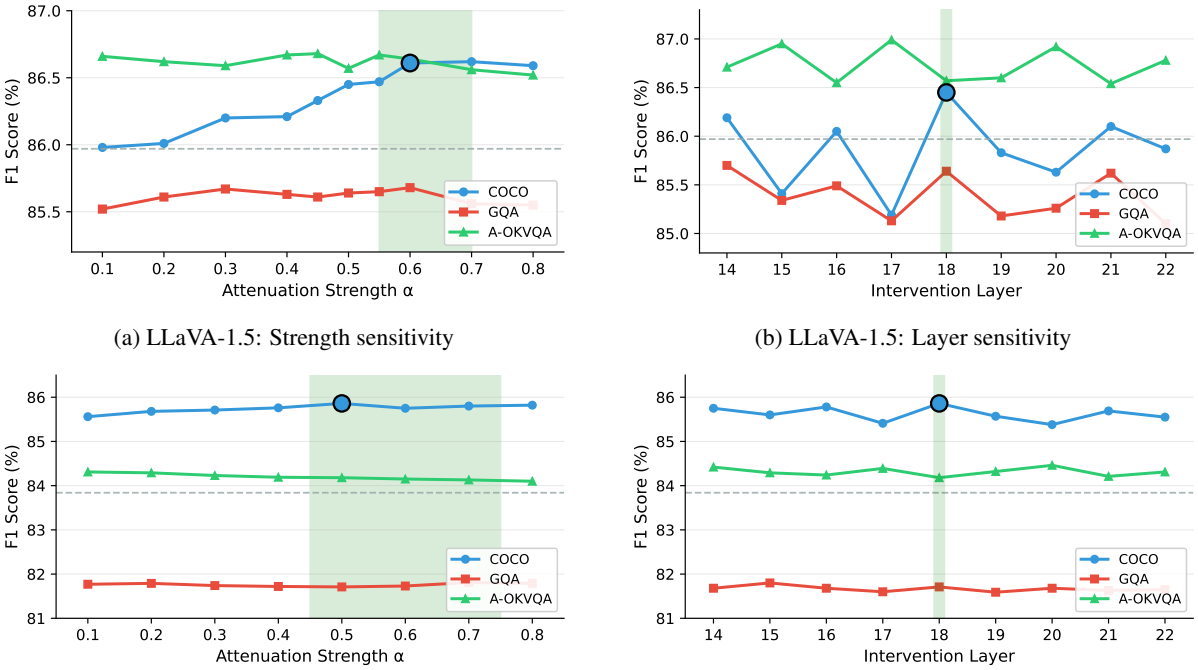
Figure 4: Qualitative comparison of hallucination correction. Case Study 1: Greedy decoding incorrectly identifies which cat opens its mouth, while FADE provides the correct answer. Case Study 2: Greedy decoding hallucinates a non-existent dog in the skiing scene, while FADE correctly denies its presence.

4.5 Ablation Study

We conduct ablations on POPE to analyze hyperparameter sensitivity across different model architectures.

Strength Sensitivity. We vary attenuation strength α from 0.1 to 0.8 on POPE (Figure 5a, 5c). For LLaVA-1.5, FADE achieves optimal performance at α=0.6 with F1 variation within 0.3% across [0.55, 0.7]. mPLUG-Owl2 shows similar patterns with optimal α=0.5–0.7, demonstrating consistent behavior across architectures.

Layer Selection. We test intervention layers 14–22 on POPE (Figure 5b, 5d). Both models show



(a) LLaVA-1.5: Strength sensitivity

(b) LLaVA-1.5: Layer sensitivity

(c) mPLUG-Owl2: Strength sensitivity

(d) mPLUG-Owl2: Layer sensitivity

Figure 5: Ablation study on POPE benchmark. (a)(c) Strength sensitivity: optimal range is [0.5, 0.7] for both models. (b)(d) Layer sensitivity: Layer 18 achieves the best trade-off across architectures. Shaded regions indicate the recommended hyperparameter ranges.

Layer 18 as optimal, consistent with our analysis that mid-to-late layers exhibit the highest directional drift. Notably, mPLUG-Owl2 shows more dataset-dependent variation: A-OKVQA prefers earlier layers (14, 20) while COCO/GQA favor Layer 18.

Task-Specific Tuning. Different tasks require dramatically different hyperparameters: MME optimal α=0.02 versus 0.6 for POPE—a 30$\times$ difference (see Appendix B). CHAIR exhibits the opposite pattern, preferring stronger attenuation (α=1.0) at later layers (Layer 20). These findings suggest that discriminative tasks (POPE) and diverse reasoning tasks (MME) have different tolerance to FFN intervention, informing practical deployment strategies.

5 Conclusion

We presented a mechanistic perspective on LVLM hallucination, investigating the origin of language priors dominance during inference. Our analysis reveals that while attention consistently aggregates visual features toward correct predictions, FFN modules at critical layers (16–22) act as the source of language priors that can override visual evidence and cause hallucination. Based on this insight, we introduced FADE (FFN Attenuation for DEcoding), a training-free method that attenuates FFN outputs at critical layers to reduce language dominance while preserving visual evidence. Unlike contrastive decoding methods that require additional forward passes, FADE operates within a single pass with minimal overhead. Experiments across LLaVA-1.5, mPLUG-Owl2, and InstructBLIP on POPE, CHAIR, and MME benchmarks demonstrate that FADE effectively mitigates hallucinations while maintaining inference efficiency. Our work opens new directions for understanding the mechanistic origins of hallucination in multimodal generation.

Limitations

Our work has several limitations. First, we only evaluate FADE on 7B-scale models; extending to larger-scale models would strengthen generalizability. Second, our experiments focus on hallucination-specific benchmarks (POPE and CHAIR); evaluation on general-purpose VQA benchmarks could provide broader insights. Third, the critical layer selection is fixed and sensitive to model architecture; exploring adaptive layer selection methods is a promising direction for future work.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 23716–23736.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, and 1 others. 2023. A survey on multimodal large language models for autonomous driving. *arXiv preprint arXiv:2311.12320*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

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. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Zechen Guan, Weihong Liu, Zhisong Wang, Jiexin Feng, Kai Li, Yongxin Yang, and Fei-Yue Wang. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haoyuan Wang, Qianglong Chen, Weihua Peng, Xiaochou Feng, Bing Qin, and 1 others. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024b. 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 (CVPR), pages 13418–13427. Highlight Paper.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Zhicheng Chen, Haozhao Wang, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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 (CVPR)*, pages 13872–13882. Highlight Paper.

Chenxi Li, Yichen Guo, Benfang Qian, Jinhao You, Kai Tang, Yaosong Du, Zonghao Zhang, and Xiande Huang. 2025a. Map: Mitigating hallucinations in large vision-language models with map-level attention processing. *arXiv preprint arXiv:2508.01653*.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–137.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025b. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *International Conference on Machine Learning (ICML)*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations (ICLR)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

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. *arXiv preprint arXiv:2404.13823*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jing Liu, Hongbin Wang, Yiyang Zhang, Hao Chen, and Xiaoming Li. 2024c. Medical report generation based on vision-language models. *Medical Image Analysis*, 89:102925.

Sheng Liu, Haotian Ye, and James Zou. 2025. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations (ICLR)*. Spotlight Paper.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision (ECCV)*.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). *arXiv preprint arXiv:2408.04664*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. 2025. Intervene-all-paths: Unified mitigation of lvlm hallucinations across alignment formats. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024a. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024b. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Kai Tang, Jinhao You, Xiuqi Ge, Hanze Li, Yichen Guo, and Xiande Huang. 2025. Mitigating hallucinations via inter-layer consistency aggregation in large vision-language models. *arXiv preprint arXiv:2505.12343*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Heming Wang, Yucong Liu, Xinyao Li, Jiayuan Luo, Shijian Lu, Yu Qiao, and Jing Wang. 2025. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *International Conference on Learning Representations (ICLR)*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024c. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13807–13816.

Shengming Yuan, Xinyu Lyu, Shuailong Wang, Beitao Chen, Jingkuan Song, and Lianli Gao. 2025. Flexac: Towards flexible control of associative reasoning in multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.

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 (ICLR)*.

A Detailed Experimental Settings

A.1 Model Descriptions

LLaVA-1.5 (Liu et al., 2023a). An improved version of LLaVA that achieves strong performance through simple architectural modifications and better training recipes. It uses a two-stage training process with visual instruction tuning on high-quality data.

InstructBLIP (Dai et al., 2023). A vision-language model that leverages instruction tuning on top of the BLIP-2 architecture. It uses a Q-Former to bridge frozen image encoders and LLMs with instruction-aware visual feature extraction.

mPLUG-Owl2 (Ye et al., 2024). Introduces modality collaboration through a shared module that enables better interaction between visual and textual modalities, achieving strong performance on various multimodal benchmarks.

A.2 Benchmark Descriptions

POPE (Li et al., 2023b). The Polling-based Object Probing Evaluation is designed to evaluate object hallucination in LVLMs. It contains 27,000 Yes/No questions about object existence in MSCOCO images, where the task is to judge whether the given object is present in the image. The benchmark includes three sampling strategies: random, popular, and adversarial. We compute accuracy, precision, recall, and F1 score for comprehensive evaluation.

CHAIR (Rohrbach et al., 2018). Caption Hallucination Assessment with Image Relevance quantifies object hallucinations in image captions by comparing generated objects to ground-truth annotations. We randomly select 500 images from the MSCOCO dataset and use three metrics: CHAIR_I (instance-level hallucination rate), CHAIR_S (sentence-level hallucination rate), and Recall (coverage of ground-truth objects).

MMHal-Bench (Sun et al., 2024b). This benchmark evaluates LVLMs beyond simple object hallucination and contains eight diverse question types: object attributes, adversarial objects, comparisons, counting, spatial relations, environment, holistic description, and others. We evaluate both the hallucination rate and response informativeness using GPT-4 as the judge.

MME (Fu et al., 2023). A comprehensive evaluation benchmark covering both perception and cognition abilities across 14 subtasks. The perception tasks include existence, count, position, color, poster, celebrity, scene, landmark, artwork, and OCR. The cognition tasks cover commonsense reasoning, numerical calculation, text translation, and code reasoning.

A.3 Comparison Method Descriptions

PAI (Liu et al., 2024d). Pays more attention to image tokens by amplifying the attention weights on visual features during decoding, ensuring that generated content is more grounded in the actual image content.

VCD (Leng et al., 2024). Visual Contrastive Decoding contrasts the output logits from original visual inputs with those from distorted visual inputs (e.g., Gaussian noise), suppressing hallucinated content that appears regardless of visual quality.

DAMO (Wang et al., 2025). Applies momentum-based activation stabilization to reduce hallucination by smoothing hidden state transitions during autoregressive generation.

VISTA (Li et al., 2025b). Introduces visual steering vectors combined with self-logits augmentation. It computes steering directions from contrastive image pairs and applies them during decoding to enhance visual grounding.

A.4 Implementation Details

All experiments are conducted on 8 NVIDIA H100 80GB GPUs. We use greedy decoding (temperature=0) for all methods to ensure reproducibility. The detailed hyperparameters for each comparison method are listed in Table 7.

For our method FADE, we use the following hyperparameters:

Note: MME requires significantly smaller attenuation strength (α=0.02–0.05) compared to POPE/CHAIR (α=0.5–0.7), as shown in Section B. This is because MME's diverse question types are more sensitive to FFN modifications.

B Detailed Ablation Study

We provide comprehensive ablation analysis on the FFN attenuation strength (α) and layer selection on POPE benchmark across three datasets (COCO, GQA, A-OKVQA).

Table 7: Hyperparameter settings for comparison methods. All hyperparameters follow the official implementations.

Method	Parameter	Value
PAI	α (attention amplification)	0.5
	γ (CFG guidance scale)	1.1
	CFG enabled	True
	Start/End layer	2 / 32
VCD	α (contrastive weight)	1.0
	β (plausibility threshold)	0.1
	Noise step (POPE)	999
	Noise step (CHAIR)	500
DAMO	α (exponential decay)	0.7
	β_1 (current hidden weight)	0.20
	β_2 (aggregated hidden weight)	0.40
	τ (similarity threshold)	-0.30
VISTA	vsv-λ (POPE)	0.01
	vsv-λ (CHAIR)	0.17
	SLA α	0.3
	SLA layers	25, 30

Table 8: Hyperparameter settings for FADE across different models.

Model	Strength α	Layer	Task
LLaVA-1.5-7B	0.6	18	POPE
	1.0	20	CHAIR
	0.02	17	MME
mPLUG-Owl2-7B	0.5	18	POPE-COCO
	0.7	18	POPE-GQA
	0.5	14	POPE-A-OKVQA
	0.6	20	CHAIR
	0.05	1	MME
InstructBLIP-7B	0.5	14	POPE

B.1 Strength Ablation (Fixed at Layer 18)

Table 9 shows the sensitivity analysis of the attenuation strength α while fixing the intervention layer at 18. We test 10 different strength values ranging from 0.1 to 0.8.

Key Findings: The optimal strength range is 0.6–0.7, achieving +0.34% to +0.44% improvement over greedy baseline. Weaker attenuation (α<0.5) provides insufficient correction, while stronger attenuation (α>0.7) shows diminishing returns. The sweet spot at α=0.6 suggests that moderate FFN suppression is sufficient to mitigate directional noise without over-correcting.

B.2 Layer Ablation (Fixed Strength at 0.5)

Table 10 analyzes the impact of layer selection while fixing α=0.5. We test 8 layers around the critical region identified in our analysis (layers 14–22).

Table 9: FFN attenuation strength ablation on POPE benchmark. Layer is fixed at 18. F1 scores are averaged across Random/Popular/Adversarial settings.

Strength	COCO F1	GQA F1	A-OKVQA F1	Avg F1
0.1	85.98	85.52	86.66	86.05
0.2	86.01	85.61	86.62	86.08
0.3	86.20	85.67	86.59	86.15
0.4	86.21	85.63	86.67	86.17
0.45	86.33	85.61	86.68	86.21
0.5	86.45	85.64	86.57	86.22
0.55	86.47	85.65	86.67	86.26
0.6	**86.61**	**85.68**	86.64	**86.31**
0.7	86.62	85.56	86.56	86.25
0.8	86.59	85.55	86.52	86.22

Baseline: Greedy decoding achieves 85.97% average F1

Table 10: Layer selection ablation on POPE benchmark. Attenuation strength is fixed at 0.5. F1 scores are averaged across Random/Popular/Adversarial settings.

Layer	COCO F1	GQA F1	A-OKVQA F1	Avg F1
14	86.19	85.70	86.71	86.20
15	85.41	85.34	86.95	85.90
16	86.05	85.49	86.55	86.03
17	85.19	85.13	86.99	85.77
18	**86.45**	**85.64**	86.57	**86.22**
19	85.83	85.18	86.60	85.87
20	85.63	85.26	86.92	85.94
21	86.10	85.62	86.54	86.09
22	85.87	85.10	86.78	85.92

Baseline: Greedy decoding achieves 85.97% average F1

Key Findings: Layer 18 consistently provides the best results across all three datasets. Layers 15 and 17 show significant degradation, suggesting these layers may serve different functional roles where FFN outputs should not be attenuated. The localized effectiveness around layer 18 validates our mechanistic analysis that directional noise is concentrated in specific critical layers rather than distributed uniformly.

B.3 MME Ablation Results

Table 11 and Table 12 show ablation results on MME Perception benchmark. Note that MME requires much smaller attenuation strength compared to POPE.

Key Findings: MME requires dramatically different hyperparameters compared to POPE: optimal strength is 0.02–0.05 (vs 0.5–0.7 for POPE), representing 2–5% attenuation vs 50–70%. This 10×–35× difference suggests that the diverse question types in MME are more sensitive to FFN modification, requiring gentler intervention. Layer 17 is optimal for MME (vs Layer 18 for POPE), indicating task-dependent critical layers.

Table 11: MME Perception: Strength ablation with Layer=18 fixed.

Strength	Perception	Δ vs Baseline	Cognition
0.01	**1512.63**	**+6.91**	363.21
0.02	1506.58	+0.86	363.21
0.03	1499.58	−6.14	367.50
0.04	1495.08	−10.64	368.21
0.05	1494.04	−11.68	360.71
0.1	1493.70	−12.02	363.57
0.2	1483.97	−21.75	355.71
0.3	1464.46	−41.26	328.21
0.5	1431.43	−74.29	290.71

Baseline: Greedy achieves 1505.72 Perception score

Table 12: MME Perception: Layer ablation with Strength=0.02 fixed.

Layer	Perception	Δ vs Baseline	Cognition
14	1504.83	−0.89	348.21
15	1518.10	+12.38	355.71
16	1505.88	+0.16	360.00
17	**1518.98**	**+13.26**	348.21
18	1506.58	+0.86	363.21
19	1508.38	+2.66	363.21
21	1508.08	+2.36	357.86
22	1505.88	+0.16	355.71

Baseline: Greedy achieves 1505.72 Perception score

B.4 mPLUG-Owl2 Ablation on POPE

We also conduct ablation studies on mPLUG-Owl2 to validate the generalizability of our findings across different model architectures.

Table 13: mPLUG-Owl2: FFN attenuation strength ablation on POPE benchmark. Layer is fixed at 18.

Strength	COCO F1	GQA F1	A-OKVQA F1	Avg F1
0.1	85.56	81.77	84.31	83.88
0.2	85.68	81.79	84.29	83.92
0.3	85.71	81.74	84.23	83.89
0.4	85.76	81.72	84.19	83.89
0.5	**85.86**	81.71	84.18	83.92
0.6	85.75	81.73	84.15	83.88
0.7	85.80	**81.81**	84.13	83.91
0.8	85.82	81.79	84.10	83.90

Baseline: Greedy decoding achieves 83.84% average F1

Key Findings for mPLUG-Owl2: Unlike LLaVA-1.5 which has a clear optimal configuration, mPLUG-Owl2 shows dataset-dependent optimal hyperparameters: (1) COCO benefits most from Layer 18 with strength 0.5; (2) GQA achieves best results at Layer 18 with strength 0.7; (3) A-OKVQA prefers Layer 14 or 20 with strength 0.5. This suggests that different model architectures may have different critical layer distributions, and dataset-specific tuning can further improve performance. The overall improvement is more modest (+0.08–0.11%) compared to LLaVA-1.5 (+0.34%), indicating that mPLUG-Owl2's modality collaboration mechanism may already partially address the directional noise issue.

Table 14: mPLUG-Owl2: Layer selection ablation on POPE benchmark. Attenuation strength is fixed at 0.5.

Layer	COCO F1	GQA F1	A-OKVQA F1	Avg F1
14	85.75	81.68	**84.42**	83.95
15	85.60	81.80	84.29	83.90
16	85.78	81.68	84.24	83.90
17	85.41	81.60	84.39	83.80
18	**85.86**	81.71	84.18	83.92
19	85.57	81.59	84.32	83.83
20	85.38	81.68	84.46	83.84
21	85.69	81.63	84.21	83.85
22	85.55	81.65	84.31	83.84

Baseline: Greedy decoding achieves 83.84% average F1

B.5 mPLUG-Owl2 MME Ablation Results

Table 15 shows the MME ablation results for mPLUG-Owl2, revealing notably different optimal layers compared to POPE.

Table 15: mPLUG-Owl2: Best configurations per layer on MME Perception benchmark.

Layer	Best α	Perception	Δ vs Baseline
1	**0.05**	**1463.73**	**+4.25**
7	0.02	1461.12	+1.64
8	0.028	1460.98	+1.50
10	0.2	1461.70	+2.22
14	0.01	1460.23	+0.75
17	0.01	1460.23	+0.75
18	0.005	1459.48	+0.00
20	0.005	1460.23	+0.75
28	0.02	1460.37	+0.89

Baseline: Greedy achieves 1459.48 Perception score

Key Findings for mPLUG-Owl2 on MME: Unlike POPE where middle layers (14–20) are optimal, MME benefits most from early layer intervention. Layer 1 with α=0.05 achieves the best result (+4.25), followed by Layer 10 (+2.22) and Layer 7 (+1.64). This suggests that for diverse question types in MME, suppressing language priors at the earliest layers is most effective. Notably, the optimal strength for early layers (0.02–0.05) is higher than for middle layers (0.005–0.01).

B.6 InstructBLIP Ablation on POPE

We validate FADE's effectiveness on InstructBLIP, which uses a Q-Former architecture with 32 visual

tokens (vs 576 for LLaVA). Table 16 shows the ablation results.

Table 16: InstructBLIP: Best configurations on POPE benchmark. Results averaged across Random/Popular/Adversarial settings.

Layer	Strength	COCO F1	A-OKVQA F1	GQA F1	Avg F1
14	**0.5**	83.7	**83.1**	**81.3**	**82.7**
17	0.5	**84.4**	82.5	81.0	82.6

Baseline (Greedy): COCO=83.8, A-OKVQA=82.6, GQA=81.1, Avg=82.5

Key Findings for InstructBLIP: The optimal configuration is layer 14 with α=0.5, achieving modest improvement on A-OKVQA (+0.5% F1) and GQA (+0.2% F1). Layer 17 achieves slightly better COCO performance but worse on other datasets. The smaller improvement compared to LLaVA-1.5 suggests that InstructBLIP's Q-Former architecture may already provide some robustness against hallucination through its learnable query-based visual feature extraction. Notably, the optimal layer (14) is earlier than LLaVA's optimal layer (18), possibly due to architectural differences in how visual information is integrated.

B.7 CHAIR Ablation Results

We provide comprehensive ablation analysis on the CHAIR benchmark, which evaluates caption hallucination through object-level metrics.

B.7.1 LLaVA-1.5 Layer Ablation on CHAIR

Table 17 shows the impact of layer selection on CHAIR metrics while fixing the attenuation strength at α=1.0. We test layers 10–22 to identify the optimal intervention point.

Table 17: LLaVA-1.5: Layer ablation on CHAIR benchmark. Strength is fixed at α=1.0. C_S/C_I: lower is better. Rec: higher is better.

Layer	$C_S\downarrow$	$C_I\downarrow$	Rec↑	Len
10	49.4	14.83	80.23	91.9
11	58.0	18.30	83.17	95.4
12	58.2	17.96	83.05	101.0
13	57.4	16.57	81.70	97.7
14	57.8	16.37	82.02	98.5
15	57.0	17.85	82.73	99.5
16	60.0	17.67	81.13	103.5
17	53.4	15.60	81.00	101.9
18	54.0	16.96	79.14	101.4
19	51.2	15.35	79.40	100.0
20	**46.6**	**14.08**	78.69	98.6
21	48.2	14.01	79.46	100.5

Baseline (Greedy): C_S=49.8, C_I=14.8, Rec=80.6, Len=101.2

Key Findings: Layer 20 achieves the lowest hallucination rates (C_S=46.6, C_I=14.08) with only a modest decrease in recall (78.69 vs 80.6 baseline). Earlier layers (10–16) either provide insufficient correction or increase hallucination. This differs from POPE where layer 18 is optimal, suggesting that discriminative and generative tasks have slightly different critical layers.

B.7.2 LLaVA-1.5 Strength Ablation on CHAIR

Table 18 shows the sensitivity to attenuation strength while fixing the intervention at layer 20.

Table 18: LLaVA-1.5: Strength ablation on CHAIR benchmark. Layer is fixed at 20.

Strength α	$C_S\downarrow$	$C_I\downarrow$	Rec↑	Len
0.1	51.2	14.94	80.23	101.2
0.2	51.6	14.89	79.85	101.0
0.3	51.0	15.18	79.72	100.3
0.4	50.4	14.76	79.91	100.2
0.5	49.0	14.59	79.78	99.4
0.6	48.8	14.98	78.95	99.2
0.7	47.4	14.54	78.57	98.5
0.8	47.4	14.70	78.82	97.9
0.9	46.8	14.28	78.69	98.0
1.0	**46.6**	**14.08**	78.69	98.6

Baseline (Greedy): C_S=49.8, C_I=14.8, Rec=80.6, Len=101.2

Key Findings: Unlike POPE where α=0.6 is optimal, CHAIR benefits from stronger attenuation (α=1.0), achieving C_S=46.6 (-3.2 vs baseline). This suggests that caption generation tasks require more aggressive FFN suppression to reduce hallucinated objects. The recall-hallucination trade-off is favorable: C_S drops by 6.4% while recall only decreases by 2.4%.

B.7.3 mPLUG-Owl2 Ablation on CHAIR

Table 19 presents ablation results for mPLUG-Owl2, showing layer and strength combinations.

Table 19: mPLUG-Owl2: Ablation on CHAIR benchmark across different layer and strength combinations.

Layer	Strength	$C_S\downarrow$	$C_I\downarrow$	Rec↑	Len
18	0.0	57.8	17.10	78.63	105.6
18	0.3	61.2	17.52	77.35	106.1
18	0.5	58.6	17.44	77.29	107.0
18	0.7	57.8	16.82	77.42	106.8
19	0.5	55.4	16.83	78.06	104.7
19	0.6	56.4	17.12	77.67	104.2
20	**0.6**	**55.0**	**16.60**	76.33	105.4
20	0.7	55.4	16.43	76.52	105.0
21	0.5	57.0	16.98	77.42	106.3
22	0.5	58.4	17.05	77.93	106.7

Baseline (Greedy): C_S=57.8, C_I=17.1, Rec=78.6, Len=105.6

Key Findings: For mPLUG-Owl2, the optimal

configuration is layer 20 with α=0.6, achieving C_S=55.0 (-2.8 vs baseline) and C_I=16.60 (-0.5). The improvement is more modest compared to LLaVA-1.5, consistent with our observation that mPLUG-Owl2's modality collaboration mechanism partially addresses hallucination. Notably, layer 18 (optimal for POPE) shows minimal improvement on CHAIR, confirming task-dependent optimal layers.

C Hyperparameter Transfer Across Models

We investigate whether optimal hyperparameters transfer across different VLM architectures.

Table 20: Optimal hyperparameters across different VLMs.

Model	Layer	α	Range
LLaVA-1.5-7B	18	0.6	16–20
mPLUG-Owl2-7B	18	0.5–0.7	14–20
InstructBLIP-7B	14	0.5	14–17

Key Finding: For LLaVA-style models (LLaVA-1.5, mPLUG-Owl2), layer 18 is consistently optimal with strength in the 0.5–0.7 range. InstructBLIP, which uses a different Q-Former architecture, shows optimal performance at an earlier layer (14) with lower strength (0.5). This suggests that the critical layers for textual bias are architecturally determined, with Q-Former-based models showing different layer distributions.

D Limitations and Future Work

Task-Specific Tuning. While FADE achieves strong results with a single hyperparameter setting for discriminative tasks (POPE) and caption generation (CHAIR), the MME benchmark requires significantly smaller attenuation strength (0.01 vs 0.6). This suggests that different task types may require task-specific tuning, which we leave for future work on adaptive strength selection.

Larger Models. Our experiments focus on 7B-scale models. Preliminary experiments on larger models (e.g., InternVL3-14B) suggest that the optimal layer may shift proportionally with model depth, but comprehensive evaluation is needed.

Training-Time Integration. FADE operates at inference time without model modification. Future work could explore training-time regularization that explicitly minimizes directional drift during instruction tuning.