

Training-Free Visual Potential Field Calibration for Balanced Hallucination Mitigation in MLLMs

Guangzong Si^{1,2,*}, Hao Yin^{1,*}, Xianfei Li^{3,4}, Qing Ding¹, Tao He³,
Wenlong Liao³, Guodong Guo², Pai Peng^{3,†}, Wentao Zhu^{2,†}

¹University of Science and Technology of China, ²Eastern Institute of Technology, Ningbo,
³COWAROBOT, ⁴Shanghai Jiao Tong University

{guangzongsi,yinhnavi}@mail.ustc.edu.cn pengpai@cowarobot.com wtzhu@eitech.edu.cn

* Equal contribution

† Corresponding authors

 Project

Abstract

Object hallucination in Multimodal Large Language Models (MLLMs) often appears in two forms: omitting objects that are present in the image and generating objects that are absent. Existing training-free mitigation methods usually apply a uniform intervention to both cases. We find that this strategy leads to an imbalanced effect: it can reduce some omission errors, but often introduces new fabrications. To understand this behavior, we analyze spatial attention patterns and layer-wise attention dynamics during generation. Our results suggest that omissions often occur when the model attends to the relevant visual regions but gradually loses visual confidence during decoding, whereas fabrications are more closely related to deep-layer semantic priors overriding visual evidence. Based on this observation, we propose **Visual Potential Field Calibration (VPFC)**, a training-free hidden-state calibration method. VPFC locates an intervention region using the spatial centroid of high-confidence visual tokens and applies a lightweight hidden-state update to adjust the model’s generative confidence. Experiments on multiple hallucination benchmarks show that VPFC consistently improves over existing training-free methods while maintaining inference efficiency close to standard greedy decoding.

1 Introduction

Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Touvron et al., 2023; Bai et al., 2025b; Dai et al., 2023; Lin et al., 2024; Wang et al., 2024b; Tong et al., 2024) have made substantial progress in visual-language tasks. However, object hallucination remains a major obstacle to their reliability. In this work, we focus on two common forms of hallucination: **omission hallucinations (omissions)**, where the model fails to mention objects that are present in the image, and **fabrication hallucinations (fabrications)**, where the model

generates objects that are absent. Although these two errors have opposite output forms, existing training-free mitigation methods often treat them as a single failure mode and apply a uniform intervention to both.

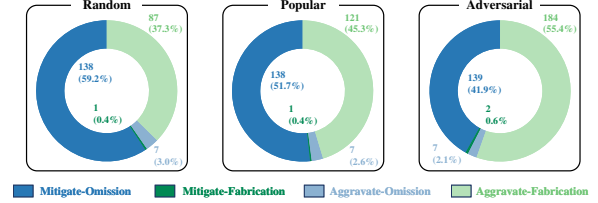


Figure 1: **Effects of VCD on omission and fabrication errors.** VCD reduces some omissions but also introduces new fabrications, especially in adversarial cases.

This unified treatment has clear limitations in practice. As shown in Fig. 1, a representative visual contrastive decoding method (Leng et al., 2024) can reduce some omission errors, but it also introduces new fabrications. This effect is especially clear when the queried object is semantically related to the image context but visually absent. The result suggests that **omissions** and **fabrications** do not always respond to the same intervention in the same way. A uniform enhancement or suppression strategy may therefore correct one type of error while amplifying the other.

To understand this imbalance, we analyze spatial attention patterns and layer-wise attention dynamics during generation. Our analysis shows that **omissions** do not necessarily arise because the visual encoder fails to perceive the target object. In many cases, the model already attends to the relevant visual region, but gradually loses generative confidence in this visual evidence during the visual-to-language mapping process, eventually producing a negative answer. In contrast, **fabrications** more often emerge in deeper representation layers, where semantic priors or co-occurrence biases override visual evidence and map contextually related

but absent objects to high-confidence outputs. In short, omissions are closer to the decay of visual evidence during generation, whereas fabrications are closer to semantic priors overriding visual grounding.

Based on these observations, we describe the model’s visual evidence aggregation with the **Visual-Semantic Attention Potential Field (VS-APF)**. In this view, High-Credibility Visual Regions (HCVRs) correspond to stable visual evidence that supports affirmative predictions, while Low-Credibility Visual Regions (LCVRs) correspond to ambiguous regions that tend to increase uncertainty and favor negative predictions. This formulation provides a simple way to explain why uniform visual interventions may reduce omissions but worsen fabrications: *they do not distinguish reliable visual anchors from ambiguous or spurious ones*.

We further propose **Visual Potential Field Calibration (VPFC)**, a training-free hidden-state calibration method. VPFC first identifies HCVRs and computes their spatial centroid. It then constructs a local intervention region around this centroid and applies a lightweight additive update to the hidden states. When the target object is present, the HCVRs are usually concentrated, so the centroid region tends to cover the object and strengthens valid visual evidence. When the target object is absent, the attention distribution is more dispersed, and the same centroid-based region covers more ambiguous areas, encouraging the model to re-evaluate visually unsupported semantic assumptions. In this way, VPFC reduces **omissions** without introducing additional **fabrications**.

Experiments on discriminative and generative hallucination benchmarks show that VPFC consistently improves over existing training-free methods. Moreover, unlike contrastive decoding methods that require additional forward passes during generation, VPFC only performs a lightweight hidden-state update and therefore maintains inference efficiency close to standard greedy decoding.

In summary, our contributions are threefold:

- We characterize the asymmetric responses of **omissions** and **fabrications** to uniform training-free interventions, showing that omission reduction can be accompanied by newly introduced fabrications.
- We analyze the two error types from spatial and layer-wise attention perspectives, show-

ing that **omissions** are mainly associated with visual confidence decay, while **fabrications** are mainly associated with deep-layer semantic priors overriding visual evidence.

- We propose VPFC, a centroid-based hidden-state calibration method that mitigates both **omissions** and **fabrications** without training, while preserving near-baseline inference efficiency.

2 Related Work

2.1 MLLMs

Recent MLLMs (Li et al., 2025; Dong et al., 2024; Sun et al., 2024; Li et al., 2024; McKinzie et al., 2024) have achieved rapid progress in visual understanding, multimodal reasoning, and instruction following (OpenAI et al., 2024; Comanici et al., 2025; Bai et al., 2025a,b). Architecturally, models such as BLIP-2 (Li et al., 2023a) and MiniGPT-4 (Zhu et al., 2024) employ lightweight cross-modal modules to bridge visual and textual representations, while the LLaVA and Qwen series adopt projection-based interfaces for efficient multimodal alignment. Beyond general perception, recent studies have further explored complex multimodal reasoning, visual grounding (Yang et al., 2023), and the internal flow of visual information in MLLMs. Despite these advances, generated responses can still deviate from the visual evidence, making hallucination a persistent challenge for reliable multimodal generation.

2.2 Hallucination Mitigation Methods

Contrastive decoding has become a representative training-free paradigm for hallucination mitigation, with different methods constructing alternative visual, textual, or intermediate representations to calibrate the output distribution (Leng et al., 2024; Wang et al., 2024c; Huo et al., 2025; Chen et al., 2024b; Wang et al., 2024a). Representative approaches include VCD, ICD, HALC, SID, and VaLiD, which differ mainly in how the contrasting signal is constructed. Despite these differences, they typically apply a unified correction during decoding without explicitly modeling omissions and fabrications as distinct failure modes, potentially resulting in imbalanced effects. Recent work (Zou et al., 2025) has also examined hallucination from the perspective of visual information utilization. Our prior work, VAF (Yin et al., 2025a) enhances

visual signals during modality fusion, while Mirage (Yin et al., 2025b) diagnoses contrastive decoding at the output level and attributes its apparent gains largely to systematic distribution shifts. Beyond these output-level analysis, VPFC characterizes omission–fabrication asymmetry through spatial and layer-wise visual dynamics and develops region-aware hidden-state calibration accordingly.

3 Motivating Observation

Existing training-free methods often apply a uniform intervention to both **omissions** and **fabrications**. A representative family of such methods is contrastive decoding, which compares output distributions under different visual conditions to reduce the model’s reliance on language priors. We take VCD (Leng et al., 2024), one of the most representative contrastive decoding methods, as a case study. VCD contrasts the output distribution under the original image with that under a distorted image and can mitigate hallucinations in some cases. However, this strategy does not explicitly distinguish between missing visual evidence and visually unsupported semantic priors. We therefore examine whether such a uniform intervention improves the two error types in a balanced way.

Experimental Setup. We use LLaVA-v1.5-7B as the backbone MLLM and apply greedy decoding for deterministic comparison. We evaluate VCD on the COCO split of POPE (Li et al., 2023c), a discriminative benchmark that asks whether a queried object is present in the image. For each prediction changed by VCD, we count whether it mitigates or aggravates an **omission** or a **fabrication** error.

Results and Analysis. Fig. 1 shows that VCD reduces some **omissions** but also introduces many new **fabrications**. This effect is most pronounced on the Adversarial subset, where the queried objects are semantically related to the image context but visually absent. The result suggests that the two hallucination types do not respond to the same intervention in the same way. A method that strengthens visual influence globally may rescue missed objects, but it can also reinforce spurious semantic associations and produce new false positives.

4 Analysis: How Omissions and Fabrications Diverge during Generation

We next analyze why the uniform intervention in Section 3 affects **omissions** and **fabrications** differently. Our analysis follows three steps. We first

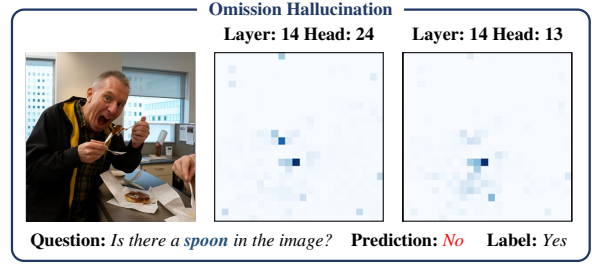


Figure 2: **Spatial attention divergence for omissions hallucination.**

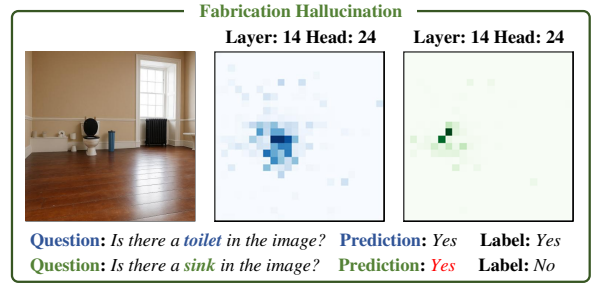


Figure 3: **Spatial attention divergence for fabrications hallucination.**

inspect spatial attention maps to examine whether the relevant visual evidence is present. We then track how this evidence changes across layers during decoding. Finally, we intervene on different visual regions to characterize their effects on output confidence. These observations motivate the VSAPF, which serves as the basis for our method in Section 5.

4.1 Spatial Evidence: The Model Often Looks at the Right Place

A common explanation for object hallucination is that the model fails to perceive the relevant visual content. However, spatial attention maps suggest a more nuanced picture. Following prior work on visual grounding heads (Kang et al., 2025), we inspect attention heads that consistently localize queried objects during generation.

Fig. 2 shows an **omission** case where the model answers that there is no spoon in the image. Although the object is small, the localization head still attends to the correct region. This suggests that the error does not necessarily come from a failure of visual encoding. Instead, the visual evidence is present in the model’s internal representation but is not converted into an affirmative prediction.

Fig. 3 shows a different pattern. When queried about a “sink” in an image containing a “toilet,” the model produces an affirmative answer. The

attention map focuses on visually and semantically related regions, but these regions do not support the queried object. This behavior is consistent with the influence of co-occurrence priors: objects that often appear in similar contexts may be incorrectly mapped to one another during generation.

Together, these cases suggest that object hallucination is not always a simple perception failure. The model may attend to relevant or semantically related visual regions, but assign them the wrong role when forming the final response.

4.2 Layer-wise Dynamics: Visual Evidence Decays or Gets Overridden

Spatial maps show where the model attends, but they do not explain how visual evidence evolves during decoding. We therefore track layer-wise attention allocation between localized visual evidence and contextual semantic priors on LLaVA-1.5 and Qwen-VL.

As shown in Fig. 4, **omissions** are associated with a gradual weakening of grounded visual evidence. The model attends to the relevant visual region in early layers, but this signal decreases as the forward pass moves into mid-to-late layers. This pattern explains why a visually grounded object can still be missed in the final answer: the model does not fully preserve its confidence in the visual evidence during the visual-to-language mapping process.

In contrast, **fabrications** show a sharper transition. Visual evidence is initially maintained, but in deeper layers the attention shifts toward contextual or semantic priors. We refer to this pattern as semantic hijacking, where visually related but unsupported concepts dominate the final prediction. This explains why fabrications are frequent in adversarial settings, where the queried objects are semantically plausible but visually absent.

4.3 Visual-Semantic Attention Potential Field

The above observations suggest that different visual regions contribute differently to generation confidence. To study this effect more directly, we divide visual tokens according to their attention salience into **High-Credibility Visual Regions (HCVRs)** and **Low-Credibility Visual Regions (LCVRs)**. We then ask a simple question: if these regions are involved in generation, how does enhancing each group change the model’s output distribution?

Fig. 5 shows that enhancing HCVRs and LCVRs leads to different prediction tendencies. Enhancing

HCVRs shifts the model toward object presence, while enhancing LCVRs increases negative responses. Fig. 6 shows the same pattern at the token-distribution level: HCVR intervention raises the probability of affirmative tokens, whereas LCVR intervention moves probability mass toward negative or uncertain outputs.

We summarize this behavior as the **Visual-Semantic Attention Potential Field (VSAPF)**. In this view, HCVRs form low-uncertainty regions that stabilize affirmative predictions, while LCVRs form high-uncertainty regions where the model tends to avoid unsupported assertions. The potential-field analogy is not intended as a physical model, but as a compact description of how visual evidence and uncertainty shape the output distribution.

Fig. 7 illustrates this confidence landscape. To further quantify it, Fig. 8 bins visual tokens by attention salience and measures the generation entropy associated with each bin. Entropy increases as attention moves from high-salience regions to low-salience regions, supporting the view that LCVRs correspond to more uncertain regions of the visual-semantic space.

4.4 Why Uniform Mitigation Is Imbalanced

The VSAPF view provides a simple explanation for the imbalance observed in Section 3. Existing contrastive or attention-based methods often change the overall influence of visual evidence without distinguishing reliable visual anchors from ambiguous or spurious ones. This can help when the model underuses valid evidence, thus reducing **omissions**. However, when the queried object is absent, the same intervention may strengthen semantically related distractors and introduce new **fabrications**.

A representative example is SID (Huo et al., 2025), whose contrastive term can be interpreted as increasing the relative contribution of high-credibility visual evidence over low-credibility regions. We provide a formal comparison in Appendix E. The key limitation is that such uniform reweighting does not decide whether the strengthened evidence corresponds to a real object or a semantically related distractor. This motivates a region-specific calibration strategy, which adapts the intervention to the spatial structure of the attended regions. We implement this idea with VPFC in Section 5.

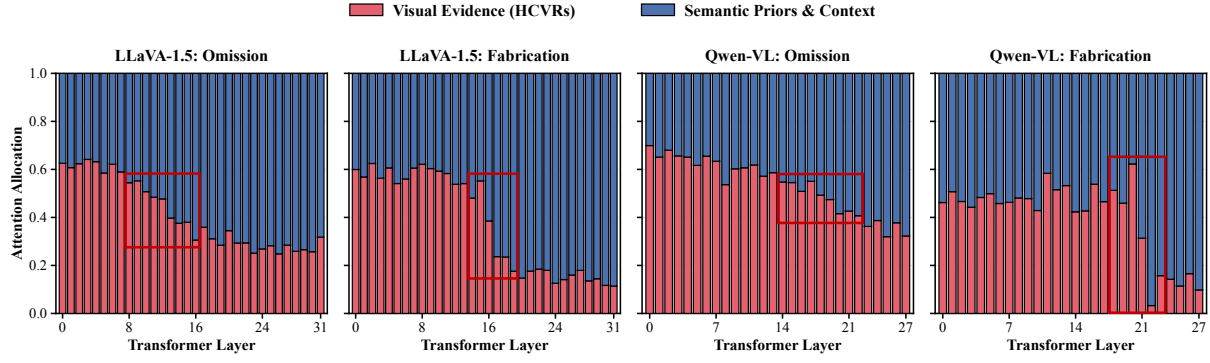


Figure 4: **Layer-wise attention allocation dynamics.** We decouple attention weights into visual evidence and semantic priors. Omissions show gradual visual signal decay, whereas fabrications show a sharper transition where semantic priors dominate the later layers.

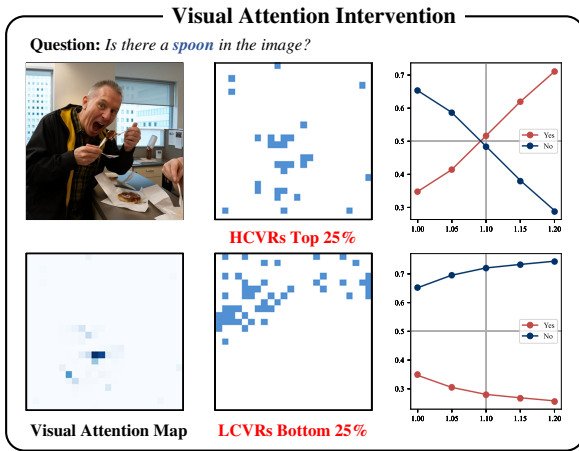


Figure 5: **Outcomes of visual interventions.** Targeted attention scaling outcomes on prediction accuracy.

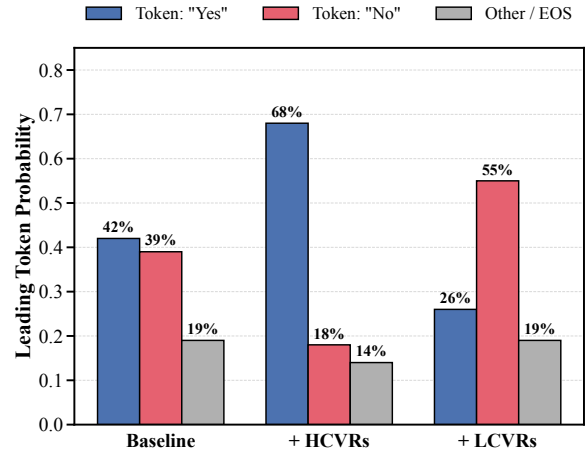


Figure 6: **Leading token probability shift.** Shift in the leading output token probabilities under visual interventions.

5 Visual Potential Field Calibration

The analysis in Section 4 suggests that hallucination mitigation should not globally amplify visual influence. Instead, the intervention should depend on the spatial structure of the attended visual regions. We therefore propose **Visual Potential Field Calibration (VPFC)**, a training-free method that extracts a centroid-based visual region and calibrates the hidden states along the corresponding confidence direction.

5.1 Centroid-based Region Extraction

We first examine the spatial distribution of HCVRs, defined as the top 25% visual tokens ranked by attention weights. On 5,000 samples from POPE and MME, we compute the spatial variance of these tokens. As shown in Fig. 9, HCVRs are usually concentrated when the queried object is present, but become more dispersed when the object is absent. This provides a simple cue for adaptive calibration.

VPFC uses this cue as follows. It computes the attention-weighted spatial centroid of HCVRs and constructs a square intervention region centered at this point. The region area is set to match the total area of the selected HCVRs. When the object is present, the centroid region tends to cover the localized visual evidence. When the object is absent, the same rule is more likely to include ambiguous low-credibility regions. Fig. 10 illustrates this behavior.

5.2 Hidden-State Calibration

Instead of directly scaling attention weights during generation, VPFC uses attention perturbation only to estimate a steering direction. Specifically, we mildly enhance the centroid region and compute the difference between the perturbed and original

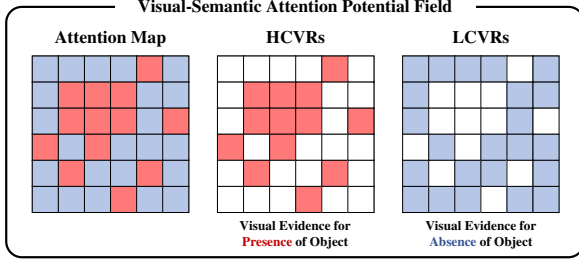


Figure 7: **Visual potential field landscape.** Conceptual 3D topology of the VSAPF.

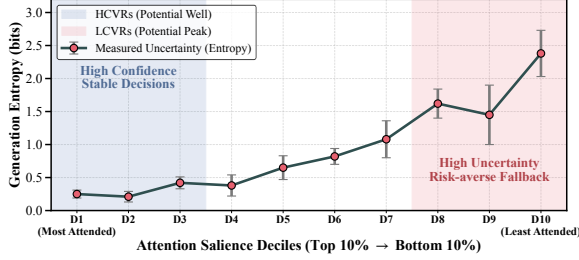


Figure 8: **Generation entropy vs. salience.** Generation entropy increases as attention moves from HCVRs to LCVRs.

hidden states:

$$\Delta_{l,h}(x) = h_{l,h}^+(x) - h_{l,h}^-(x), \quad (1)$$

where $h_{l,h}^+(x)$ and $h_{l,h}^-(x)$ denote the hidden states of the h -th attention head in the l -th layer under the perturbed and original conditions, respectively. VPFC then applies an additive update:

$$\tilde{h}_{l,h}(x) = h_{l,h}(x) + \alpha \Delta_{l,h}(x), \quad (2)$$

where α controls the intervention strength. This update calibrates the model’s confidence toward the selected visual region while avoiding the instability of strong attention scaling.

5.3 Attention Head Selection

Since not all attention heads contribute equally to visual-language alignment, VPFC applies the intervention only to important heads. Following prior head-importance estimation methods (Michel et al., 2019; Li et al., 2023b), we score each head by

$$I_{h,l} = \left\| A_{l,h} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{l,h}} \right\|_1, \quad (3)$$

where $A_{l,h}$ is the attention map and $\mathcal{L}(x)$ is the auto-regressive loss. We rank heads by $I_{h,l}$ and apply the hidden-state update only to the top $\gamma\%$ heads.

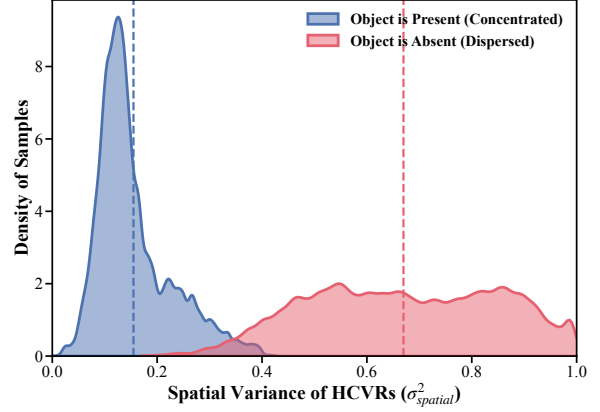


Figure 9: **Spatial variance ($\sigma_{spatial}^2$) distribution of HCVRs on POPE.** HCVRs are more concentrated when the queried object is present and more dispersed when it is absent.

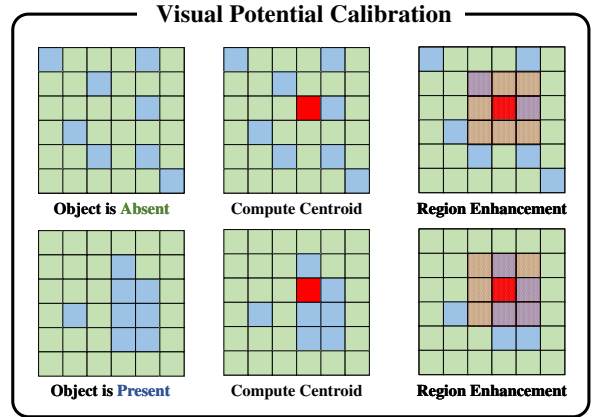


Figure 10: **Illustration of VPFC.** VPFC constructs a fixed-size region around the centroid of HCVRs and applies hidden-state calibration.

6 Experiments

6.1 Experimental Setup

Evaluation Datasets. We evaluate our method across a range of benchmarks, including discriminative tasks (POPE (Li et al., 2023c), MME-Hallucination (Fu et al., 2025)) and generative tasks (CHAIR (Rohrbach et al., 2018) and LLaVA-Bench-in-the-wild (Liu et al., 2023)). Furthermore, to assess broader reasoning generalizability, we incorporate comprehensive multimodal evaluations, including MMStar (Chen et al., 2024a), MMBench (Liu et al., 2024b), MMBench-CN (Liu et al., 2024b), and MMvet (Yu et al., 2024).

Baselines. We compare VPFC against standard greedy decoding (Regular) and three representative training-free mitigation approaches (Leng et al., 2024; Huo et al., 2025; Zou et al., 2025): VCD, SID, and Memory-Space Visual Retracing

Model	Method	Random		Popular		Adversarial	
		Acc.	F1	Acc.	F1	Acc.	F1
LLaVA-1.5	Regular	87.10 $\uparrow$ 0.00	85.53 $\uparrow$ 0.00	84.83 $\uparrow$ 0.00	83.33 $\uparrow$ 0.00	83.60 $\uparrow$ 0.00	82.29 $\uparrow$ 0.00
	VCD	88.60 $\uparrow$ 1.50	88.17 $\uparrow$ 2.64	85.20 $\uparrow$ 0.37	85.10 $\uparrow$ 1.77	81.93 $\downarrow$ 1.67	82.47 $\uparrow$ 0.18
	SID	87.53 $\uparrow$ 0.43	86.45 $\uparrow$ 0.92	85.21 $\uparrow$ 0.38	85.50 $\uparrow$ 2.17	80.88 $\downarrow$ 2.72	80.69 $\downarrow$ 1.60
	MemVR	88.50 $\uparrow$ 1.40	87.34 $\uparrow$ 1.81	86.10 $\uparrow$ 1.27	85.01 $\uparrow$ 1.68	79.20 $\downarrow$ 4.40	79.28 $\downarrow$ 3.01
	VPFC	89.80 $\uparrow$ 2.70	88.90 $\uparrow$ 3.37	87.60 $\uparrow$ 2.77	87.02 $\uparrow$ 3.69	85.80 $\uparrow$ 2.20	84.60 $\uparrow$ 2.31
Qwen-VL	Regular	87.43 $\uparrow$ 0.00	86.48 $\uparrow$ 0.00	84.70 $\uparrow$ 0.00	83.96 $\uparrow$ 0.00	82.50 $\uparrow$ 0.00	81.70 $\uparrow$ 0.00
	VCD	88.80 $\uparrow$ 1.37	88.11 $\uparrow$ 1.63	85.13 $\uparrow$ 0.43	84.69 $\uparrow$ 0.73	79.83 $\downarrow$ 2.67	79.23 $\downarrow$ 2.47
	SID	87.83 $\uparrow$ 0.40	87.17 $\uparrow$ 0.69	84.57 $\downarrow$ 0.13	84.67 $\uparrow$ 0.71	81.50 $\downarrow$ 1.00	80.90 $\downarrow$ 0.80
	MemVR	88.47 $\uparrow$ 1.04	87.62 $\uparrow$ 1.14	85.27 $\uparrow$ 0.57	84.73 $\uparrow$ 0.77	80.90 $\downarrow$ 1.60	79.80 $\downarrow$ 1.90
	VPFC	89.73 $\uparrow$ 2.30	89.07 $\uparrow$ 2.59	87.90 $\uparrow$ 3.20	87.00 $\uparrow$ 3.04	84.50 $\uparrow$ 2.00	83.40 $\uparrow$ 1.70
Qwen2.5-VL	Regular	88.52 $\uparrow$ 0.00	87.59 $\uparrow$ 0.00	87.83 $\uparrow$ 0.00	86.59 $\uparrow$ 0.00	86.73 $\uparrow$ 0.00	85.55 $\uparrow$ 0.00
	VCD	89.61 $\uparrow$ 1.09	88.40 $\uparrow$ 0.81	88.41 $\uparrow$ 0.58	87.32 $\uparrow$ 0.73	84.18 $\downarrow$ 2.55	83.08 $\downarrow$ 2.47
	SID	90.33 $\uparrow$ 1.81	89.17 $\uparrow$ 1.58	88.67 $\uparrow$ 0.84	87.30 $\uparrow$ 0.71	85.69 $\downarrow$ 1.04	84.75 $\downarrow$ 0.80
	MemVR	89.72 $\uparrow$ 1.20	88.73 $\uparrow$ 1.14	88.29 $\uparrow$ 0.46	87.36 $\uparrow$ 0.77	84.92 $\downarrow$ 1.81	83.65 $\downarrow$ 1.90
	VPFC	91.26 $\uparrow$ 2.74	90.18 $\uparrow$ 2.59	90.54 $\uparrow$ 2.71	89.63 $\uparrow$ 3.04	88.32 $\uparrow$ 1.59	87.25 $\uparrow$ 1.70

Table 1: **Performance on POPE.** We report accuracy and F1 on Random, Popular, and Adversarial splits. Numbers denote absolute changes over the regular decoding baseline.

(MemVR).

Implementation Details. The primary MLLM backbones are LLaVA-v1.5-7B (Liu et al., 2024a), Qwen-VL-7B (Bai et al., 2023), and Qwen2.5-VL-7B (Bai et al., 2025b). The steering coefficient α is set to 4, and the proportion of selected attention heads γ is set to 25%. All evaluations employ greedy search to ensure deterministic comparisons of the underlying logit distributions.

6.2 Main Results

Results on POPE. Tab. 1 presents the performance on POPE. Existing methods, including VCD, SID, and MemVR, improve performance on the Random and Popular subsets, where omission errors are more common. However, they show clear drops on the Adversarial subset, where queried objects are semantically related to the image but visually absent. This pattern is consistent with our analysis in Section 4: a uniform increase in visual influence can help recover missed objects, but may also reinforce spurious visual-semantic associations. In contrast, VPFC improves performance across all subsets, including the Adversarial subset, suggesting that centroid-based calibration can reduce omissions without introducing the same level of fabrication errors.

Results on MME and MME-Hallucination.

Tab. 2 reports the results on MME-Hallucination and MME. Existing contrastive methods improve some categories but degrade others, showing an unstable trade-off across different hallucination-related settings. VPFC maintains or improves performance on most subsets. This indicates that region-guided hidden-state calibration can mitigate object-level hallucinations while better preserving the model’s general multimodal reasoning ability.

Results on Generative Tasks. Tab. 3 reports performance on open-ended generation benchmarks. VPFC achieves the lowest hallucination rates on CHAIR and the best overall scores on LLaVA-Bench. These results suggest that the proposed calibration improves factual grounding in free-form generation, rather than only changing the output distribution in discriminative yes/no tasks.

6.3 Inference Efficiency

A primary limitation of existing contrastive methods is their substantial computational overhead, as they require dual probability calculations at each auto-regressive decoding step. To evaluate practical efficiency, we compare per-example inference latency on POPE and LLaVA-Bench tasks. As illustrated in Fig. 11, VCD and SID consistently incur a $> 2\times$ latency penalty that compounds linearly with the generated sequence length. Conversely, VPFC

Model	Method	MM-Hall Total	Object-Level		Attribute-Level	
			Existence	Count	Position	Color
LLaVA-1.5	Regular	620.00 $\uparrow$ 0.00	185.00 $\uparrow$ 0.00	146.67 $\uparrow$ 0.00	128.33 $\uparrow$ 0.00	160.00 $\uparrow$ 0.00
	VCD	598.36 $\downarrow$ 21.64	190.00 $\uparrow$ 5.00	128.33 $\downarrow$ 18.34	133.33 $\uparrow$ 5.00	146.70 $\downarrow$ 13.30
	SID	598.33 $\downarrow$ 21.67	185.00 $\uparrow$ 0.00	130.00 $\downarrow$ 16.67	128.33 $\uparrow$ 0.00	155.00 $\downarrow$ 5.00
	MemVR	610.00 $\downarrow$ 10.00	190.00 $\uparrow$ 5.00	130.00 $\downarrow$ 16.67	130.00 $\uparrow$ 1.67	160.00 $\uparrow$ 0.00
	VPFC	635.00 $\uparrow$ 15.00	190.00 $\uparrow$ 5.00	146.67 $\uparrow$ 0.00	133.33 $\uparrow$ 5.00	165.00 $\uparrow$ 5.00
Qwen-VL	Regular	618.33 $\uparrow$ 0.00	175.00 $\uparrow$ 0.00	140.00 $\uparrow$ 0.00	128.33 $\uparrow$ 0.00	175.00 $\uparrow$ 0.00
	VCD	603.33 $\downarrow$ 15.00	170.00 $\downarrow$ 5.00	130.00 $\downarrow$ 10.00	123.33 $\downarrow$ 5.00	180.00 $\uparrow$ 5.00
	SID	616.66 $\downarrow$ 1.67	175.00 $\uparrow$ 0.00	138.33 $\downarrow$ 1.67	128.33 $\uparrow$ 0.00	175.00 $\uparrow$ 0.00
	MemVR	608.33 $\downarrow$ 10.00	170.00 $\downarrow$ 5.00	135.00 $\downarrow$ 5.00	133.33 $\uparrow$ 5.00	170.00 $\downarrow$ 5.00
	VPFC	645.00 $\uparrow$ 26.67	185.00 $\uparrow$ 10.00	145.00 $\uparrow$ 5.00	135.00 $\uparrow$ 6.67	180.00 $\uparrow$ 5.00
Qwen2.5-VL	Regular	700.00 $\uparrow$ 0.00	195.00 $\uparrow$ 0.00	160.00 $\uparrow$ 0.00	155.00 $\uparrow$ 0.00	190.00 $\uparrow$ 0.00
	VCD	668.29 $\downarrow$ 21.71	195.00 $\uparrow$ 0.00	148.66 $\downarrow$ 11.34	160.33 $\uparrow$ 5.33	164.30 $\downarrow$ 25.70
	SID	686.83 $\downarrow$ 13.17	190.00 $\downarrow$ 5.00	153.50 $\downarrow$ 6.50	168.33 $\uparrow$ 13.33	175.00 $\downarrow$ 15.00
	VPFC	735.00 $\uparrow$ 35.00	195.00 $\uparrow$ 0.00	176.67 $\uparrow$ 16.67	173.33 $\uparrow$ 18.33	190.00 $\uparrow$ 0.00

Table 2: **Performance on MM-Hallucination.** We report the total MM-Hallucination score and four fine-grained scores. Numbers denote absolute changes over the regular decoding baseline.

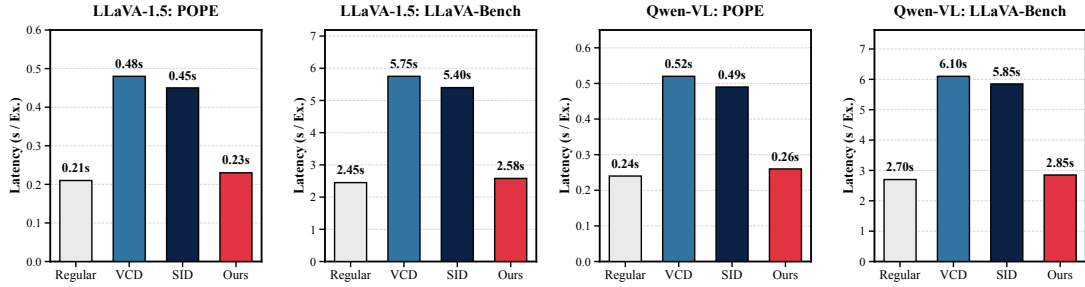


Figure 11: Inference latency comparison. Existing contrastive decoding methods severely degrade inference speed by requiring dual probability calculations. VPFC maintains an inference efficiency nearly identical to the standard greedy decoding baseline.

introduces negligible overhead. Because the centroid extraction and steering direction $\Delta_{l,h}(x)$ are computed efficiently, and hidden states are modified via a simple additive operation, VPFC achieves SOTA hallucination mitigation while preserving near-baseline inference speeds.

6.4 Ablation Studies

We conduct ablation experiments using LLaVA-v1.5-7B on the POPE benchmark to validate the core components of VPFC.

Hyperparameter Robustness. Fig. 12 and Fig. 13 analyze the head selection ratio γ and steering coefficient α . Applying intervention across too many attention heads $\gamma > 50\%$ injects excessive noise, causing a sharp decline in prediction accuracy. Constraining the selection ratio to $\gamma \leq 25\%$ yields

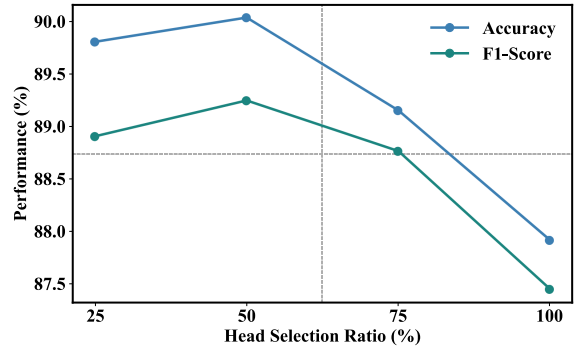


Figure 12: **Ablation on the head selection ratio.** Impact of the attention head selection ratio in VPFC.

Method	CHAIR_S ↓	CHAIR_I ↓	Avg. ↓
Regular	50.2 ↑ 0.00	15.6 ↑ 0.00	32.9 ↑ 0.00
VCD	54.8 ↑ 4.60	16.5 ↑ 0.90	35.6 ↑ 2.70
SID	49.2 ↓ 1.00	15.1 ↓ 0.50	32.1 ↓ 0.80
MemVR	51.2 ↑ 1.00	15.9 ↑ 0.30	33.5 ↑ 0.60
VPFC	46.8 ↓ 3.40	13.8 ↓ 1.80	30.3 ↓ 2.60

(a) Performance on CHAIR

Method	Conv.	Desc.	Reason.
Regular	59.6 ↑ 0.00	53.4 ↑ 0.00	75.6 ↑ 0.00
VCD	57.4 ↓ 2.20	50.9 ↓ 2.50	76.9 ↑ 1.30
SID	59.2 ↓ 0.40	51.3 ↓ 2.10	76.1 ↑ 0.50
MemVR	58.1 ↓ 1.50	51.2 ↓ 2.20	77.4 ↑ 1.80
VPFC	62.1 ↑ 2.50	53.8 ↑ 0.40	77.9 ↑ 2.30

(b) Performance on LLaVA-Bench

Table 3: Performance comparison on CHAIR and LLaVA-Bench.

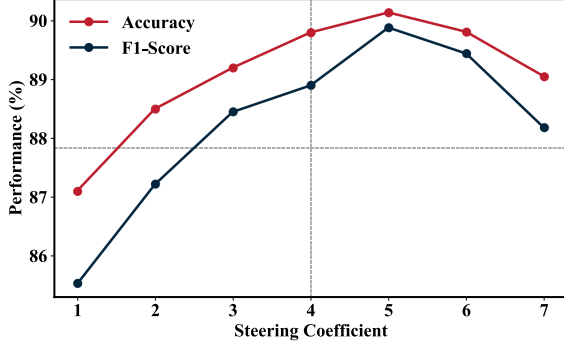


Figure 13: Ablation on the steering coefficient. Influence of the steering coefficient in VPFC.

strong results. Additionally, VPFC demonstrates stability across a broad range of the steering coefficient ($3 < \alpha < 6$), confirming that the method does not require exhaustive hyperparameter tuning for practical deployment.

7 Conclusion

We study **omissions** and **fabrications** hallucinations in MLLMs from the perspective of visual evidence usage. We find that uniform training-free interventions affect the two errors differently: **omissions** often reflect decayed valid evidence, whereas **fabrications** are tied to unsupported semantic priors. We therefore propose VPFC, a training-free hidden-state calibration method that adapts intervention to the spatial structure of attended visual regions. Experiments show that VPFC reduces hallucinations more consistently than existing training-free methods with near-baseline efficiency.

Limitations

VPFC relies on the initial visual grounding ability of the base MLLMs. If an object is entirely unrecognizable due to severe occlusion, low resolution, or failures in the visual encoder, VPFC cannot recover visual evidence that is absent from the model’s internal representation. In addition, the

current centroid-based intervention assumes that the relevant visual evidence is relatively compact. For queries involving multiple spatially dispersed instances or several disconnected visual anchors, a single center of mass may be insufficient and may not align well with all relevant regions. Finally, although VPFC improves hallucination mitigation in the evaluated settings, it should be viewed as a calibration mechanism rather than a complete safeguard for high-stakes deployment.

Ethical Considerations and Potential Risks

This work studies object hallucination mitigation in MLLMs for research purposes and does not collect new human-subject data, private user data, or personally identifiable information. All experiments use existing public benchmarks and model artifacts in accordance with their stated usage and access conditions, and any released code, configurations, or evaluation results are intended solely to support reproducible academic research. While VPFC reduces certain object-level hallucinations, it does not eliminate incorrect visual understanding and remains dependent on the visual grounding ability of the base model. Improved hallucination mitigation may therefore create an over-impression of model reliability. VPFC should not be used as the sole reliability mechanism in high-stakes settings, including medical, legal, safety-critical, or surveillance-related applications, nor should the released materials be used to make misleading claims of model safety or bypass task-specific validation.

Acknowledgments

We thank the High-Performance Computing Center at Eastern Institute of Technology and Ningbo Institute of Digital Twin for providing the computing resources that supported this work. This work was supported by Project 2024zd01, “Research and Application of Next-Generation General-Purpose Intelligent Robot Brain (Robo-GPT),” and Project

202423dl2050005, “Research and Application of Key Technologies for L4 End-to-End Autonomous Driving Based on Multimodal Large Language Models.” We also thank the anonymous reviewers for their constructive comments and valuable suggestions.

References

- 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*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. [Are we on the right way for evaluating large vision-language models?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 27056–27087. Curran Associates, Inc.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024b. [HALC: object hallucination reduction via adaptive focal-contrast decoding](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 7824–7846. PMLR / OpenReview.net.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsen, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xianguyu Zhang, Kaisheng Ma, and Li Yi. 2024. [Dream-llm: Synergistic multimodal comprehension and creation](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. [Self-introspective decoding: Alleviating hallucinations for large vision-language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. [Your large vision-language model only needs A few attention heads for visual grounding](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9339–9350. Computer Vision Foundation / IEEE.
- 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882. IEEE.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [Llava-onevision: Easy visual task transfer](#). *Trans. Mach. Learn. Res.*, 2025.

- 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.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. [Monkey: Image resolution and text label are important things for large multi-modal models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26753–26763. IEEE.
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. [VILA: on pre-training for visual language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, and 10 others. 2024. [MM1: methods, analysis and insights from multimodal LLM pre-training](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*, volume 15087 of *Lecture Notes in Computer Science*, pages 304–323. Springer.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). Preprint, arXiv:2303.08774.
- 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. [Towards VQA models that can read](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.
- Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yuezeng Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. [Generative multimodal models are in-context learners](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14398–14409. IEEE.
- Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. [Cambrian-1: A fully open, vision-centric exploration of multimodal llms](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). Preprint, arXiv:2307.09288.
- Jiaqi Wang, Yifei Gao, and Jitao Sang. 2024a. Valid: Mitigating the hallucination of large vision language

- models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024b. [Cogvlm: Visual expert for pretrained language models](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- 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, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 15840–15853. Association for Computational Linguistics.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv:2310.11441*.
- Hao Yin, Guangzong Si, and Zilei Wang. 2025a. [Clear-sight: Visual signal enhancement for object hallucination mitigation in multimodal large language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 14625–14634. Computer Vision Foundation / IEEE.
- Hao Yin, Guangzong Si, and Zilei Wang. 2025b. [The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms?](#) In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Weihaoyu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. [MM-vet: Evaluating large multimodal models for integrated capabilities](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754. PMLR.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. [Minigpt-4: Enhancing vision-language understanding with advanced large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.
- Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Ken-ling Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. 2025. [Look twice before you answer: Memory-space visual re-tracing for hallucination mitigation in multimodal large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80873–80899. PMLR.

A Overview

This supplementary material provides additional technical details and empirical evidence to support the main text.

We first present the step-by-step algorithm of VPFC in Sec. B. We then provide expanded motivation and analysis in Sec. C, including cross-architecture validation on Qwen2.5-VL, the spatio-temporal evolution of the VSAPF topology, and a token-level anatomy of semantic hijacking. Next, we present hallucination change analysis in Sec. D and provide a formal comparison with contrastive decoding in Sec. E. The remaining sections report additional qualitative results in Sec. H, benchmark configurations in Sec. I, evaluated model specifications in Sec. J, AI usage declarations in Sec. K.

B Algorithm of Visual Potential Field Calibration

Fig. 14 illustrates the complete procedure of VPFC. For each backbone, we randomly sample 50 image-question pairs from the TextVQA (Singh et al., 2019) training split to construct a calibration set. For each sample, the model’s greedy next-token prediction is used as a pseudo-target to compute the autoregressive loss. We then evaluate each attention head using the attention-gradient, average its importance score across the calibration samples, and retain the top $\gamma\%$ heads in each intervened layer. The selected head configuration is fixed for the corresponding backbone. During generation, VPFC aggregates the text-to-visual attention of these heads to identify HCVRs, constructs a local intervention region around their spatial centroid, and derives the steering direction for hidden-state calibration.

C More Motivation & Analysis

C.1 Cross-Architecture Validation on Qwen2.5-VL

To verify whether the observed hallucinatory mechanisms persist in advanced architectures, we extend the layer-wise attention analysis to Qwen2.5-VL-7B. As shown in Fig. 15, the model exhibits consistent structural behaviors despite its enhanced language backbone. During an omission error, visual evidence initially captures substantial attention but undergoes a fluctuating decay across the intermediate layers, ultimately yielding to semantic priors. Conversely, during a fabrication error, vi-

Algorithm 1: Visual Potential Field Calibration (VPFC)

Input: MLLM $\mathcal{M}$; unlabeled calibration set $\mathcal{C} = \{(I_n, q_n)\}_{n=1}^{50}$; test input (x, q) ; α, γ, τ

Output: Calibrated hidden states $\tilde{h}$

Head Configuration

```

1  for  $(I_n, q_n) \in \mathcal{C}$  do
2     $\hat{y}_n \leftarrow \arg \max_y p_\theta(y \mid I_n, q_n)$ 
3     $\mathcal{L}_n \leftarrow -\log p_\theta(\hat{y}_n \mid I_n, q_n)$ 
4     $I_{h,l}^{(n)} \leftarrow \left\| A_{l,h}^{(n)} \odot \frac{\partial \mathcal{L}_n}{\partial A_{l,h}^{(n)}} \right\|_1$ 
5  end for
6   $\bar{I}_{h,l} \leftarrow \frac{1}{|\mathcal{C}|} \sum_n I_{h,l}^{(n)}$ 
7   $\mathcal{H}_{\text{sel}}^l \leftarrow \text{Top}_\gamma(\{\bar{I}_{h,l}\}_h)$  for each intervened layer  $l$ 

```

Visual Potential Field Calibration

```

8   $(A, h^-) \leftarrow \mathcal{M}(x, q)$ 
9   $s_l \leftarrow \text{Aggregate}_{h \in \mathcal{H}_{\text{sel}}^l} A_{l,h}^{\text{text} \rightarrow \text{vis}}$ 
10  $\mathcal{V}_H \leftarrow \text{Top}_\tau(s_l)$ 
11  $(x_c, y_c) \leftarrow \text{Centroid}(\mathcal{V}_H)$ 
12  $\mathcal{R} \leftarrow \text{Region}(x_c, y_c), |\mathcal{R}| = |\mathcal{V}_H|$ 
13  $h^+ \leftarrow \text{Enhance}(\mathcal{M}, x, q, \mathcal{R})$ 
14  $\Delta \leftarrow h^+ - h^-$ 
15  $\tilde{h} \leftarrow h^- + \alpha \Delta$ 
16 return  $\tilde{h}$ 

```

Figure 14: Visual Potential Field Calibration (VPFC).

sual attention maintains a volatile presence through the middle layers before experiencing an abrupt collapse in the deep layers, where semantic hijacking overtakes the generative process. The minor attention spikes in the terminal layers primarily reflect residual gradient leakage rather than meaningful visual grounding. These observations confirm that signal decay and semantic hijacking are inherent structural challenges in cross-modal alignment, demonstrating the broad applicability of the proposed spatial intervention.

C.2 Spatio-Temporal Evolution of the VSAPF Topology

To investigate the formation of the VSAPF, we map the generation entropy across network depths and attention salience deciles. As depicted in Fig. 16, the generative confidence landscape exhibits a distinct structural bifurcation during the forward pass. Early layers display an undifferentiated high-entropy state, indicating that the model has not yet committed to a specific semantic output. Across the intermediate layers, the landscape undergoes structural turbulence reflecting cross-modal feature reorganization. In the deep layers, this spatial divergence solidifies: HCVRs collapse into stable potential wells, while LCVRs escalate

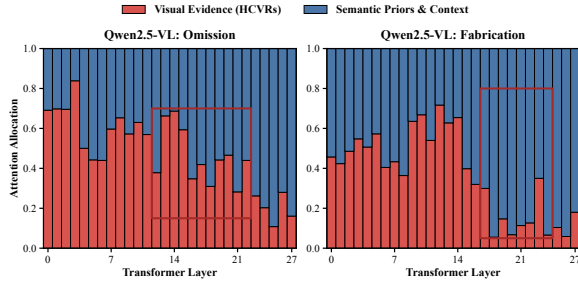


Figure 15: Layer-wise attention allocation dynamics on Qwen2.5-VL-7B. (a) Omission error, exhibiting a fluctuating decay of visual attention across intermediate layers. (b) Fabrication error, showing an abrupt collapse of visual attention in the deep layers as semantic priors dominate the generation.

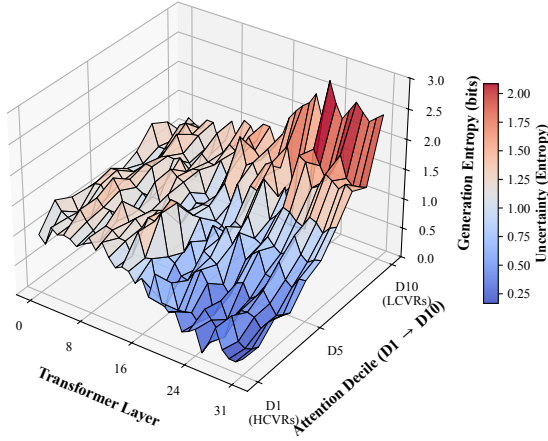


Figure 16: Spatio-temporal evolution of the VSAPF.

into unstable potential peaks.

C.3 Token-Level Anatomy of Semantic Hijacking

To explicitly demonstrate the semantic hijacking mechanism, we extract the token-wise attention weights during a fabrication hallucination. As illustrated in Fig. 17, the early decoding stage exhibits visually grounded generation, where the attention is correctly concentrated on the image representations. However, as the forward pass enters the deep layers, an attention shift is observed. The attention on visual tokens drops to a very low level, while the induced textual token receives a highly concentrated attention weight. This token-level evidence confirms that deep-layer computations decouple from visual inputs and align with textual co-occurrence biases, thereby validating the necessity of our latent calibration.

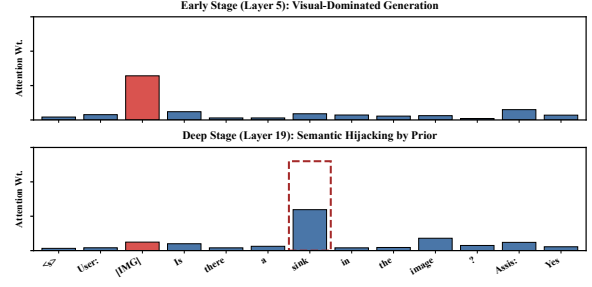


Figure 17: Token-wise attention distribution during a fabricated generation. In the early layer, attention is correctly grounded on the visual tokens. In the deep layer, the visual attention drops significantly, while the textual prior receives a highly concentrated attention weight, indicating a semantic hijacking.

D Hallucination Change Analysis

Tab. 4 further analyzes how different training-free methods change the two hallucination types on POPE. For each method, we categorize the changed predictions into four cases: mitigating omissions, mitigating fabrications, aggravating omissions, and aggravating fabrications. The results show that existing methods often improve one error type at the cost of the other. For example, VCD and SID mitigate a large number of omissions on LLaVA-1.5, but they also introduce 392 and 332 new fabrication errors, respectively. Similar trade-offs can also be observed on Qwen-VL and Qwen2.5-VL, where baseline methods frequently increase either omission or fabrication errors after intervention. In contrast, VPFC shows a more balanced change pattern. It reduces both omissions and fabrications while introducing substantially fewer new errors, especially fabrication errors. This supports our claim that region-aware calibration can mitigate hallucinations more consistently than uniform contrastive decoding.

E Formal Comparison with contrastive decoding

This section provides a formal comparison between contrastive decoding and direct visual-region enhancement under the VSAPF view. We use SID (Huo et al., 2025) as a representative example, since it constructs a contrastive distribution by comparing the original visual input with low-importance visual tokens.

Model	Method	Mitigate-Omission	Mitigate-Fabrication	Aggravate-Omission	Aggravate-Fabrication
LLaVA-1.5	VCD	415	4	21	392
	SID	274	0	0	332
	MemVR	141	0	0	194
	VPFC	216	39	0	27
Qwen-VL	VCD	70	2	18	82
	SID	114	0	0	136
	MemVR	42	13	40	15
	VPFC	61	166	4	0
Qwen2.5-VL	VCD	18	59	40	64
	SID	14	109	31	46
	MemVR	35	36	22	55
	VPFC	68	152	11	0

Table 4: **Hallucination changes on POPE.**

SID defines the contrastive distribution as:

$$\begin{aligned}
p_{\text{sid}}(y_i) &= \text{softmax} \left[\text{logit}_{\theta}(y_i | v, x) \right. \\
&\quad \left. + \alpha \cdot \Delta_{\text{sid}}(y_i) \right], \\
\Delta_{\text{sid}}(y_i) &= \text{logit}_{\theta}(y_i | v, x) \\
&\quad - \text{logit}_{\theta}(y_i | v_{\text{low}}, x).
\end{aligned} \tag{4}$$

Under the VSAPF view, v_{low} mainly corresponds to low-credibility or ambiguous visual regions. Therefore, the contrastive term in Eq. 4 can be interpreted as increasing the relative contribution of the remaining high-credibility visual evidence. For comparison, a direct enhancement of high-credibility visual regions v_{high} can be written as:

$$\begin{aligned}
p_{\text{enh}}(y_i) &= \text{softmax} \left[\text{logit}_{\theta}(y_i | v, x) \right. \\
&\quad \left. + \beta \cdot \Delta_{\text{enh}}(y_i) \right], \\
\Delta_{\text{enh}}(y_i) &= \text{logit}_{\theta}(y_i | v_{\text{high}}, x) \\
&\quad - \text{logit}_{\theta}(y_i | v, x).
\end{aligned} \tag{5}$$

We now show why these two operations can produce similar probability shifts. Let the final hidden state before the language modeling head be h . Using a local linear approximation of the cross-modal mapping, we decompose h into a linguistic prior component and two visual components:

$$\begin{aligned}
h &\approx h_{\text{text}} + W_{\text{high}} v_{\text{high}} \\
&\quad + W_{\text{low}} v_{\text{low}},
\end{aligned} \tag{6}$$

where v_{high} and v_{low} denote high- and low-credibility visual regions, respectively. Since the language modeling head is linear, the pre-softmax

logit for token y_i can be approximated as:

$$\begin{aligned}
\text{logit}_{\theta}(y_i | v, x) &\approx L_{\text{text}}(y_i) \\
&\quad + L_{\text{high}}(y_i) + L_{\text{low}}(y_i),
\end{aligned} \tag{7}$$

where $L_{\text{high}}(y_i) = EW_{\text{high}} v_{\text{high}}$ and $L_{\text{low}}(y_i) = EW_{\text{low}} v_{\text{low}}$ denote the logit contributions of the corresponding visual regions.

When the model is evaluated using only v_{low} , the high-credibility contribution is largely removed. Thus,

$$\text{logit}_{\theta}(y_i | v_{\text{low}}, x) \approx L_{\text{text}}(y_i) + L_{\text{low}}(y_i). \tag{8}$$

Substituting this approximation into Eq. 4, the SID correction term becomes:

$$\Delta_{\text{sid}}(y_i) \approx L_{\text{high}}(y_i). \tag{9}$$

Therefore, SID approximately modifies the original logit as:

$$\text{logit}_{\text{sid}}(y_i) \approx \text{logit}_{\theta}(y_i | v, x) + \alpha L_{\text{high}}(y_i). \tag{10}$$

Conversely, when the model is evaluated using only v_{high} , the low-credibility contribution is largely removed:

$$\text{logit}_{\theta}(y_i | v_{\text{high}}, x) \approx L_{\text{text}}(y_i) + L_{\text{high}}(y_i). \tag{11}$$

The direct enhancement correction term in Eq. 5 becomes:

$$\Delta_{\text{enh}}(y_i) \approx -L_{\text{low}}(y_i). \tag{12}$$

Thus, direct high-region enhancement approximately gives:

$$\text{logit}_{\text{enh}}(y_i) \approx \text{logit}_{\theta}(y_i | v, x) - \beta L_{\text{low}}(y_i). \tag{13}$$

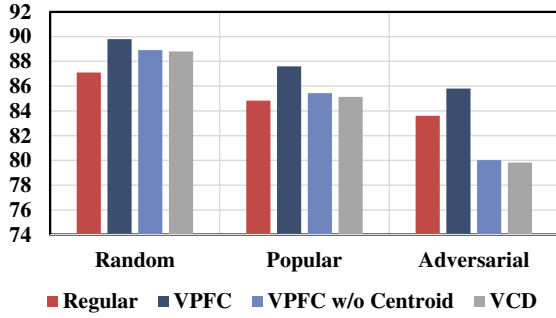


Figure 18: **Ablation on the centroid-focused strategy.** Effect of removing the centroid-based region extraction in VPFC.

Equations 10 and 13 reveal a structural duality: suppressing the effect of low-credibility regions and enhancing the effect of high-credibility regions both increase the relative dominance of salient visual evidence. Since the softmax function is invariant to global logit shifts, these two operations can induce similar probability changes when their scaling coefficients are matched to the attention distribution norms.

This explains why contrastive decoding can rescue missed objects when valid visual evidence is underused. However, it also reveals its limitation. The operation changes the relative strength of visual evidence but does not verify whether the strengthened evidence corresponds to a real object or to a semantically related distractor. As a result, uniform decode may reduce omissions while introducing new fabrications, especially when low-credibility regions contain semantically plausible but visually unsupported cues.

F More Ablation

Centroid-Focused Strategy. A central design of VPFC is anchoring the intervention around the spatial centroid of HCVRs. We ablate this by computing the steering direction directly from the top 25% visual tokens without spatial concentration constraints. As shown in Fig. 18, removing the centroid computation drastically degrades performance, pulling the Adversarial subset accuracy down to the baseline level. Unconstrained attention enhancement fails to differentiate between true visual presence and ambiguous features, re-triggering the fabrication dilemma. The centroid constraint is necessary to correctly navigate the potential field topology.

G Generalizability to Advanced Architectures

To verify the plug-and-play versatility of VPFC, we integrate it into the recent InternVL (Chen et al., 2025; Zhu et al., 2025) series (InternVL2.5-8B and InternVL3-8B). As illustrated in Tab. 5, VPFC yields consistent performance gains across six distinct benchmarks, including comprehensive multimodal evaluations like MMStar and MM-Bench. The omission-fabrication imbalance remains a structural characteristic even in newer architectures. The proposed potential field calibration provides a reliable solution that scales effectively with model advancements.

H More Qualitative Results

Fig. 19 presents qualitative comparisons on LLaVA-Bench. Across both examples, VPFC produces responses that are better grounded in the visual content and avoids hallucinated details introduced by the baseline and contrastive decoding method.

We provide additional qualitative examples from LLaVA-Bench to evaluate the generation quality in open-ended scenarios. As illustrated in Fig. 20, the generated responses demonstrate accurate comprehension of complex scenes without fabricating ungrounded contents.

I Benchmark Details

We provide additional details of the benchmarks used in our experiments.

Polling-based Object Probing Evaluation. POPE (Li et al., 2023c) evaluates object hallucination through binary object-presence questions. Given an image, the model is asked whether a particular object appears in the visual content and responds with a yes-or-no answer. The official evaluation protocol reports Accuracy, Precision, Recall, F1 score, and the ratio of affirmative responses. In our experiments, we report Accuracy and F1 on the Random, Popular, and Adversarial subsets, following the standard POPE evaluation setting. The official POPE repository is released under the MIT License.

Multimodal Model Evaluation. MME (Fu et al., 2025) evaluates multimodal models over 14 sub-tasks, consisting of ten perception tasks and four cognition tasks. The perception tasks cover existence, count, position, color, poster, celebrity, scene, landmark, artwork, and OCR, while the cognition tasks cover commonsense reasoning, numeri-

Model	Method	MME	MMStar	POPE	MMBench-CN	MMBench	MMVet	Avg.
InternVL3-8B	Baseline	2415 $\uparrow$ 0.00	68.2 $\uparrow$ 0.00	91.1 $\uparrow$ 0.00	82.2 $\uparrow$ 0.00	83.4 $\uparrow$ 0.00	81.3 $\uparrow$ 0.00	100% $\uparrow$ 0.00
	Ours	2563 $\uparrow$ 148	68.4 $\uparrow$ 0.2	91.8 $\uparrow$ 0.7	83.1 $\uparrow$ 0.9	84.2 $\uparrow$ 0.8	82.0 $\uparrow$ 0.7	101.68% $\uparrow$ 1.68
InternVL2.5-8B	Baseline	2344 $\uparrow$ 0.00	62.8 $\uparrow$ 0.00	90.6 $\uparrow$ 0.00	82.6 $\uparrow$ 0.00	84.6 $\uparrow$ 0.00	62.8 $\uparrow$ 0.00	100% $\uparrow$ 0.00
	Ours	2471 $\uparrow$ 127	62.5 $\downarrow$ 0.3	91.1 $\uparrow$ 0.5	82.8 $\uparrow$ 0.2	85.3 $\uparrow$ 0.7	63.6 $\uparrow$ 0.8	101.31% $\uparrow$ 1.31

Table 5: **Performance comparison on six benchmarks.** Our method consistently improves InternVL2.5-8B and InternVL3-8B.



Q: Is there any strawberry-flavored yogurt in the fridge?
 LLaVA-v1.5: *Yes, there is strawberry-flavored yogurt.*
 SID: *Yes, strawberry yogurt is visible.*
 Ours: *No, there is no strawberry-flavored yogurt.*



Q: What are the animals doing in the painting?
 LLaVA-v1.5: *The bear, cat, and rabbit are all drinking tea.*
 SID: *The bear and cat are drinking, while the rabbit is pouring tea.*
 Ours: *The bear eats a cookie, the cat holds a mug, and the rabbit drinks.*

Figure 19: **Qualitative comparison of object hallucination mitigation on LLaVA-Bench.**

cal calculation, text translation, and code reasoning. For our hallucination-oriented evaluation, we use four perception subtasks—existence, count, position, and color—to measure object- and attribute-level errors.

Caption Hallucination Assessment with Image Relevance. CHAIR (Rohrbach et al., 2018) measures object hallucination in generated image captions with respect to objects supported by MSCOCO annotations. It provides two complementary metrics: CHAIR_i is the fraction of hallucinated object mentions among all object mentions in generated captions, whereas CHAIR_s is the fraction of generated captions containing at least one hallucinated object. The metric considers the 80 object categories used in the MSCOCO segmentation task. Following our evaluation setting, we randomly sample 500 images from the MSCOCO val2014 split and generate captions using the prompt “Please describe this image in detail.” The official CHAIR implementation is released under the BSD 2-Clause License.

J Model Card

We provide a detailed description of MLLMs here. **LLaVA-1.5.** LLaVA-1.5 (Liu et al., 2024a) is built upon the Vicuna-7B language model. For visual encoding, it employs a pre-trained CLIP ViT-L/14 model with an input resolution of 336×336 pixels. The visual features are mapped into the language model’s embedding space using a two-layer MLP

cross-modal projector. LLAMA 2 Community License.

Qwen-VL. Qwen-VL (Bai et al., 2023) integrates the Qwen-7B language backbone with a ViT-bigG visual encoder initialized from OpenCLIP. It processes images at a standard resolution of 448×448 pixels. A defining architectural feature is its position-aware vision-language adapter, which uses a single-layer cross-attention module to compress dense visual features into a fixed-length visual token sequence before feeding them into the language model. Tongyi Qianwen LICENSE AGREEMENT.

Qwen2.5-VL. Qwen2.5-VL (Bai et al., 2025b) is built on the Qwen2.5 language backbone. Architecturally, it departs from fixed-resolution visual processing by adopting a native dynamic-resolution ViT, allowing images of different sizes and aspect ratios to be processed more flexibly. It also incorporates multimodal rotary positional embeddings to preserve spatial and temporal positional information, together with visual token merging/window-attention designs for efficient high-resolution perception. Apache license 2.0.

K AI Use

The authors used AI Tools solely for the purpose of refining the grammatical structure and improving the linguistic clarity of the manuscript. During this process, the authors reviewed and edited the content as needed and take full responsibility for



Q: Describe this photo in detail.

A: The image depicts a group of mangosteens, a tropical fruit known for its sweet and tangy flavor, arranged on a wooden surface. There are four mangosteens in total, three of which are intact with a deep purple, almost burgundy, rind. The fourth mangosteen is cut in half, revealing the white, segmented flesh inside, which is reminiscent of a small, white pumpkin. Each segment is juicy and appears succulent, with a translucent quality that suggests ripeness and freshness.

No Hallucination.



Q: What are the possible reasons of the formation of this sight?

A: The image showcases a volcanic cone, a geological feature typically formed as a result of volcanic eruptions. The formation of such a cone can be attributed to several factors, with the most direct cause being the eruption of a volcano. When magma rises from beneath the Earth's crust to the surface and erupts, it can accumulate around the vent, creating a conical shape.

No Hallucination.



Q: Which iconic movie scene is being parodied in the meme?

A: The meme is parodying the iconic scene from the movie "The Lion King," where Rafiki, the mandrill, holds up Simba, the young lion cub, on Pride Rock for all the animals of the Pride Lands to see. This moment is a significant and emotional one in the film, symbolizing the introduction of the new king to his kingdom. The meme humorously replaces Simba with a Shiba Inu dog, a breed that has become popular in internet culture, often associated with the "Doge" meme. The character in the image is depicted in a similar pose to Rafiki, lifting the dog up high with one hand.

No Hallucination.



Q: Describe this painting in detail.

A: The painting is a portrait of a woman set against a distant, imaginary landscape. It features a figure with an enigmatic smile and direct gaze that has captivated viewers for centuries. The woman has long, dark hair parted in the middle and draped over her shoulders. Her complexion is fair, and she has a serene and gentle expression.

No Hallucination.



Q: What is the name of the character in the image?

A: The character in the image is named Conan Edogawa. He is the main protagonist of the long-running manga and anime series "Detective Conan" (also known as "Case Closed" in some regions), created by Gosho Aoyama. Conan is known for his intelligence and detective skills, as well as his signature bowtie and glasses.

No Hallucination.

Figure 20: **More qualitative results.** Using LLaVA-bench, we have demonstrated the effectiveness of our method on OpenQA, where the model's hallucination-free responses fully showcase the validity of our approach.

the final version of the paper.