Spotlight and Shadow: Attention-Guided Dual-Anchor Introspective Decoding for MLLM Hallucination Mitigation

Yebo Wu[1†], **Han Jin**[1†], **Zhijiang Guo**[2,3*], **Li Li**[1*]
[1]State Key Laboratory of IOTSC, University of Macau
[2]HKUST, [3]HKUST (Guangzhou)
{yc37926, mc56727, llili}@um.edu.mo
zhijiangguo@hkust-gz.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable reasoning capabilities yet continue to suffer from hallucination, where generated text contradicts visual content. In this paper, we introduce Dual-Anchor Introspective Decoding (DAID), a novel contrastive decoding framework that dynamically calibrates each token generation by mining the model's internal perceptual discrepancies. Specifically, DAID identifies a Spotlight layer to amplify visual factual signals and a Shadow layer to suppress textual inertia. By leveraging visual attention distributions to guide this dual-anchor selection process, our method ensures precise, token-specific adaptation. Experimental results across multiple benchmarks and MLLMs demonstrate that DAID significantly mitigates hallucination while enhancing general reasoning capabilities.

1 Introduction

Multimodal Large Language Models (MLLMs) (Li et al., 2024b) have demonstrated exceptional versatility, excelling in tasks ranging from image captioning (Bucciarelli et al., 2024; Sarto et al., 2025) to complex reasoning (Yang et al., 2025; Xu et al., 2026b,d, 2025). By synergizing the perceptual acuity of vision encoders with the cognitive reasoning capabilities of LLMs, MLLMs achieve unprecedented proficiency in comprehending and generating content grounded in multimodal inputs, fundamentally reshaping the landscape of artificial intelligence (Wu et al., 2026c, 2025b).

Despite these significant strides, the practical deployment of MLLMs remains severely constrained by the persistent issue of hallucination (Zhang et al., 2025b; Zhu et al., 2025; Wu et al., 2025a), where models generate textual descriptions that are unfaithful to the provided visual content. This phenomenon critically undermines the trustworthiness

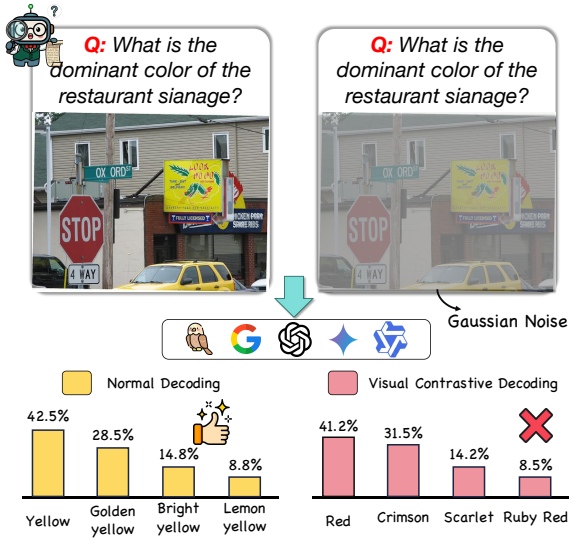


Figure 1: A failure case of Visual Contrastive Decoding (VCD) caused by external perturbations.

of MLLMs in high-stakes scenarios (Xu et al., 2024, 2026c), such as financial analysis (Gan et al., 2024) and medical diagnosis (Qiu et al., 2024). For instance, in medical imaging diagnosis (Al-Saad et al., 2024), a model might hallucinate a non-existent tumor driven by spurious textual correlations, leading to dangerous consequences.

To address this challenge, training-free Contrastive Decoding (CD) strategies have emerged as a promising direction (Wang et al., 2024c). The fundamental premise involves penalizing hallucinated tokens by contrasting the model's logits against a constructed negative distribution (Suo et al., 2025). However, existing CD methods, such as Visual CD (VCD; Leng et al. 2024) and Instruction CD (ICD; Wang et al. 2024c), are constrained by two critical limitations. First, these methods necessitate an additional forward pass for the negative sample at each decoding step, imposing substantial computational burden and latency; for instance, VCD increases inference latency by $1.83\times$ compared to standard decoding on NVIDIA V100. Second, existing approaches rely on heuristic external per-

*Corresponding authors.
†Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2026, pages 13219–13233

turbations (e.g., visual masking) to synthesize the negative distribution, which can introduce stochastic noise. As illustrated in Figure 1, such perturbations lead to erroneous semantic shifts: while standard decoding correctly identifies "yellow" attributes, VCD inadvertently prioritizes incorrect "red" vocabulary due to perturbation-induced noise.

In this paper, we resolve these dilemmas by shifting the paradigm from external intervention to internal introspection. Our approach is grounded in the observation that MLLMs exhibit distinct layer-wise behaviors: shallow layers show severe hallucination tendencies, while intermediate layers possess superior visual perception capabilities. We posit that these internal perceptual discrepancies serve as valuable, intrinsic contrastive sources that can calibrate the generation process. Leveraging this insight, we introduce Dual-Anchor Introspective Decoding (DAID), a novel CD method that mines contrastive signals directly from the model's intermediate states. Specifically, DAID dynamically identifies two anchors guided by visual attention distributions: a Spotlight layer to amplify factual visual signals and a Shadow layer to suppress linguistic priors. This design enables effective hallucination mitigation within a single forward pass while eliminating the uncertainty noise inherent in traditional methods. Our main contributions are summarized as follows:

- We conduct a granular layer-wise diagnosis of MLLM decoding dynamics, revealing that shallow layers exhibit severe hallucination tendencies, whereas intermediate layers possess superior visual perception capabilities.
- We propose DAID, a novel CD method that simultaneously amplifies factual visual signals and suppresses language priors to mitigate hallucinations within a single forward pass.
- Our extensive experiments across multiple benchmarks demonstrate that DAID significantly outperforms existing methods in hallucination mitigation while enhancing general multimodal reasoning capabilities.

2 Motivation and Observation

In this section, we investigate whether MLLMs possess inherent internal signals necessary to self-correct hallucinations. To answer this, we conduct a layer-wise analysis of the decoding trajectory by probing hidden states and attention distributions across layers. This analysis reveals how visual perception emerges and how hallucination tendencies evolve during generation, thereby identifying optimal internal contrastive anchors for decoding.

2.1 Evolution of Hallucination Rate

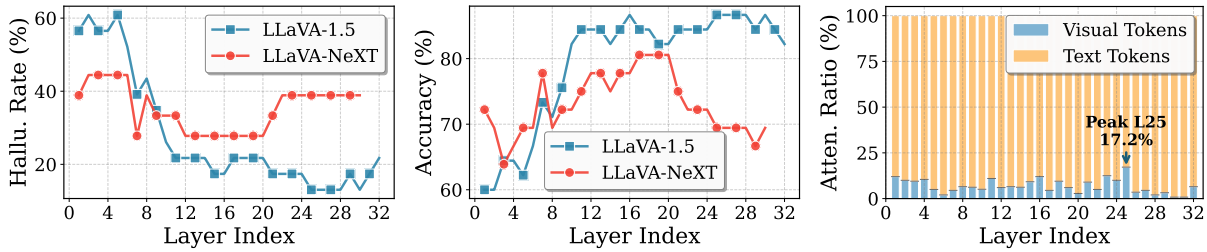
We first trace the genesis of hallucination by evaluating the factual consistency of layer-wise decoded sequences on the POPE (Li et al., 2023c) benchmark. As shown in Figure 2(a), models exhibit divergent trajectories: LLaVA-1.5 (Liu et al., 2023) shows a generally monotonic decline in hallucination rates, while LLaVA-NeXT (Li et al., 2024a) follows a non-monotonic pattern with rates rebounding in deeper layers. Despite these variations, a critical commonality emerges: shallow layers exhibit the most severe hallucination tendencies on both models (peaking at 60.87% for LLaVA-1.5 and 44.44% for LLaVA-NeXT), generating fluent but factually detached descriptions.

We attribute this phenomenon to visual agnosia. In initial layers, visual encoder representations are not yet semantically aligned with the LLM's latent space due to the modality gap. While these layers may encode low-level features (e.g., edges, textures), they suffer from a semantic void, lacking the reasoning capability to ground these features into text. This misalignment forces the model to rely on linguistic patterns learned during pre-training, rather than actual visual evidence. Consequently, generation at this stage is driven almost exclusively by unimodal language priors.

TAKEAWAY 1: Shallow layers, characterized by linguistic noise prevailing over visual reasoning, naturally constitute an ideal "Shadow" (negative anchor). This enables us to effectively isolate and subtract the language inertia driving hallucination, leaving clearer visual signals.

2.2 Evolution of Visual Perception Capability

To pinpoint layers where visual comprehension is maximized, we probe object recognition accuracy across the model's depth. As illustrated in Figure 2(b), recognition accuracy does not follow a simple linear progression. Instead, we identify a distinct seeing-then-forgetting phenomenon, particularly pronounced in LLaVA-NeXT. Specifically, accuracy rises rapidly to a peak at intermediate layers (e.g., layer 25 for LLaVA-1.5 and layer 17 for LLaVA-NeXT), indicating that the model attains optimal visual fidelity within the network's intermediate stages. Crucially, this fidelity is not



(a) Hallucination Rate. (b) Object Recognition Accuracy. (c) Attention Distribution of LLaVA-1.5.

Figure 2: Layer-wise diagnosis of MLLM decoding dynamics on the POPE benchmark, revealing the evolution of hallucination rates, visual perception capability, and attention distribution across layers.

sustained; it degrades significantly in deeper layers, dropping by 4.45% for LLaVA-1.5 and a striking 11.12% for LLaVA-NeXT. We attribute this decline to textual inertia, where dominant pre-trained language priors progressively override visual signals to prioritize linguistic fluency and coherence.

TAKEAWAY 2: Intermediate layers, characterized by peak visual fidelity, constitute an ideal "Spotlight" (positive anchor). This enables us to amplify authentic perceptual signals, ensuring visually faithful generation.

2.3 Visual Attention Bridging the Gap

To elucidate the mechanistic underpinnings of the above phenomena and identify these anchors without expensive probes, we employ attention to quantify model's reliance on visual cues. In Figure 2(c), we visualize attention weights allocated to visual tokens for LLaVA-1.5. A striking correlation emerges: visual attention reaches its zenith at layer 25, aligning precisely with the layer where object recognition accuracy peaks. This synchronization indicates that high visual attention is a definitive signature of the model's peak perceptual state.

Conversely, in shallower layers (e.g., around layer 6), attention reaches a local minimum, corresponding directly to regions with high hallucination rates. This validates that visual attention distribution serves as a robust, training-free proxy for the model's cognitive state. Critically, attention weights reveal the intrinsic balance between visual and linguistic processing; this makes them ideal for pinpointing layers with optimal visual grounding, ensuring our dual-anchor selection is driven by the model's actual visual reliance.

TAKEAWAY 3: Visual attention serves as a faithful, training-free proxy for internal model states. Leveraging this indicator enables token-specific anchoring, precisely identifying the Spotlight and Shadow layers for each decoding step.

3 Methodology: DAID

Drawing on the empirical insights from Section 2, we introduce DAID, a training-free framework designed to mitigate hallucinations. DAID dynamically calibrates generation by pinpointing token-specific internal states. As shown in Figure 3, DAID employs a Spotlight layer to amplify authentic visual signals and a Shadow layer to suppress linguistic noise.

3.1 Attention-Guided Dynamic Anchoring

Our preliminary analysis establishes that the distribution of attention weights serves as a robust proxy for the model's real-time reliance on visual cues. To leverage this, we propose the Visual Attention Score (VAS), which dynamically gauges visual reliance at each decoding step.

Visual Attention Score. Consider the generation of the t-th token, where the model processes a context sequence containing N_v visual tokens. Let $A_{t,k}^{l,h} \in [0,1]$ denote the attention weight in layer l and head h, representing the affinity between the current query token t and the key token k. We define the VAS for layer l at step t as the cumulative attention mass allocated to visual tokens, averaged across all H heads:

$$\text{VAS}_t(l) = \frac{1}{H} \sum_{h=1}^{H} \sum_{k \in \text{V}} A_{t,k}^{l,h}, \tag{1}$$

where $\text{V} = \{1, \ldots, N_v\}$ represents the set of visual token indices. Intuitively, a high $\text{VAS}_t(l)$ indicates that layer l actively grounds the current generation in visual features, whereas a low score implies a predominance of linguistic priors.

Spotlight Anchor: Peak Perception. To counteract the seeing-then-forgetting phenomenon, where visual signals are attenuated in deeper layers,

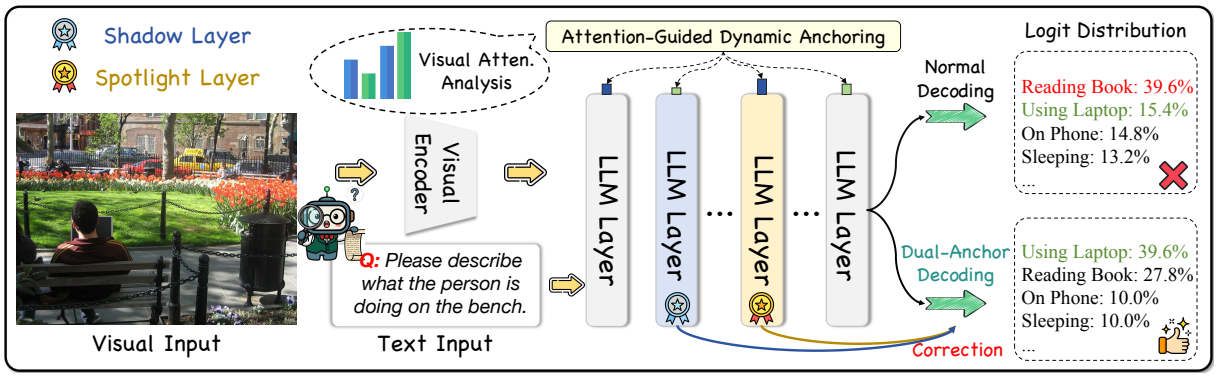


Figure 3: Framework of DAID. DAID dynamically identifies dual-anchor layers guided by the visual attention and calibrates the final output by leveraging both the Spotlight and Shadow anchors to ensure visual grounding.

we identify the positive anchor (Spotlight layer) as the layer exhibiting maximum visual grounding:

$$L^t_{spot.} = \underset{l \in \{1,\dots,L\}}{\text{argmax}} \, \text{VAS}_t(l). \tag{2}$$

By anchoring to $L^t_{spot.}$, we retrieve authentic visual details that are otherwise diluted in the final layers.

Shadow Anchor: Visual Agnosia. Conversely, the negative anchor (Shadow layer) is selected to capture a pre-perception state where visual signals are minimal and linguistic noise dominates. Crucially, we constrain the search space to layers preceding the Spotlight layer to isolate textual inertia before visual features are integrated:

$$L^t_{shad.} = \underset{l \in \{1,\dots,L^t_{spot.}-1\}}{\text{argmin}} \, \text{VAS}_t(l). \tag{3}$$

This constraint ($L^t_{shad.} < L^t_{spot.}$) ensures that the Shadow layer represents pure linguistic noise, providing an ideal state for contrastive suppression.

3.2 Dual-Anchor Introspective Decoding

DAID rectifies the final output via a dual-anchor contrastive mechanism. The calibrated logits $\mathcal{L}^t_{\text{DAID}}$ at step t are derived by modulating the final layer's distribution with signals from both anchors:

$$\begin{aligned}\mathcal{L}^t_{\text{DAID}} = &[\mathcal{L}^t_{L_{final}} + \alpha \cdot \mathcal{L}^t_{L_{spot.}}] \cdot (1+\beta) \\ &- \beta \cdot \mathcal{L}^t_{L_{shad.}},\end{aligned} \tag{4}$$

where α and β modulate the correction intensity, and $\mathcal{L}^t_{L_{spot.}}$ and $\mathcal{L}^t_{L_{shad.}}$ denote the logits of the Spotlight and Shadow anchors, respectively. This formulation achieves two simultaneous objectives:

- **Visual Perception Reinforcement** ($+\alpha \cdot \mathcal{L}^t_{L_{spot.}}$)**:** Re-injects the peak visual signals to strengthen the model's perceptual acuity.
- **Language Prior Suppression** ($-\beta \cdot \mathcal{L}^t_{L_{shad.}}$)**:** Penalizes tokens driven primarily by linguistic correlations to mitigate textual inertia.

Adaptive Plausibility Constraint. Since the Spotlight layer ($L_{spot.}$) typically resides in intermediate stages, directly injecting its logits may introduce syntactically inappropriate tokens despite high visual relevance. Similarly, contrastive subtraction on negligible-probability tokens causes instability. To address these, we restrict dual-anchor contrastive decoding to a candidate set $\mathcal{V}'$, dynamically determined by the final layer's distribution:

$$\begin{aligned}\mathcal{V}'(y_{<t}) = \{y_t \in \mathcal{V} \mid &\, p_\theta(y_t|\mathcal{L}_{L_{final}}, y_{<t}) \\ &\geq \gamma \cdot \max_w p_\theta(w|\mathcal{L}_{L_{final}}, y_{<t})\},\end{aligned} \tag{5}$$

where $\gamma \in [0,1]$ confines adjustment to linguistically valid candidates. The dual-anchor adjustment is applied exclusively to candidates within $\mathcal{V}'$:

$$p_{\text{DAID}}(y_t) = \begin{cases} p_{\text{DAID}}(y_t) & \text{if } y_t \in \mathcal{V}'(y_{<t}), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

This constraint ensures DAID sharpens factual details and eliminates inertia while maintaining the generated text's fluency and coherence.

4 Experiments

4.1 Experimental Settings

Baselines. We compare DAID with the following baselines. 1) DoLa (Chuang et al., 2023) contrasts early and late layers to isolate factual knowledge; 2) VCD (Leng et al., 2024) mitigates hallucination by contrasting original and distorted visual inputs; 3) OPERA (Huang et al., 2024) employs an attention penalty to prevent token over-trust; 4) SID (Huo et al., 2024) exploits internal state discrepancies

for intrinsic correction; and 5) EAZY (Che et al., 2025) eliminates hallucinations by zeroing out hallucinatory image tokens.

Evaluation Models. We validate DAID across five representative MLLMs: LLaVA-1.5 (Liu et al., 2023), LLaVA-NeXT (Li et al., 2024a), Qwen2-VL (Wang et al., 2024b), MiniGPT-4 (Zhu et al., 2023), and InstructBLIP (Dai et al., 2023). All models are evaluated at the 7B parameter scale.

Implementation Details. We conduct evaluations on NVIDIA RTX 5070 Ti GPUs. For the dual-anchor contrastive decoding, we set $\alpha = 0.8$ to govern the intensity of visual perception reinforcement and $\beta = 0.2$ to modulate the suppression of linguistic priors. Additionally, the confidence threshold for the adaptive plausibility constraint is set to $\gamma = 0.9$ for the POPE benchmark and 0.1 for others, ensuring that the dual-anchor adjustments are confined to a linguistically valid candidate set.

4.2 Research Questions and Benchmarks

We evaluate DAID on three visual hallucination and five general vision-language benchmarks to investigate the following research questions:

- **RQ1:** Can DAID mitigate hallucinations without sacrificing core reasoning capabilities?
- **RQ2:** What is the impact of α (visual reinforcement) and β (language suppression)?
- **RQ3:** Is DAID consistently effective across diverse MLLM architectures?
- **RQ4:** What are the individual contributions of the Spotlight and Shadow anchors?
- **RQ5:** How does DAID qualitatively rectify token distribution during decoding?

Visual Hallucination Tasks. These benchmarks assess the factual consistency of model responses.

- **POPE** (Li et al., 2023c): Assesses object hallucination by asking binary questions (e.g., "`Is there a <object> in the image?`"). We report the Accuracy and F1 score.
- **CHAIR** (Rohrbach et al., 2018): Evaluates object hallucination in image captioning by comparing generated descriptions against ground-truth annotations. We report the CHAIR_I (Eq. (7)) and CHAIR_S (Eq. (8)).
- **MME** (Fu et al., 2025): Measures perceptual and cognitive abilities. We focus on the object-level (Existence, Count) and attribute-level (Position, Color) subsets, reporting the corresponding perception score.

$$\text{CHAIR}_\text{I} = \frac{|\{\text{hallucinated objects}\}|}{\text{all mentioned objects}}. \quad (7)$$

$$\text{CHAIR}_\text{S} = \frac{|\{\text{captions with hall. objects}\}|}{\text{all captions}}. \quad (8)$$

General Vision-Language Tasks. These benchmarks evaluate reasoning capabilities of MLLMs.

- **GQA** (Hudson and Manning, 2019): Focuses on visual reasoning and compositional question answering through image scene graphs.
- **VQA$^{\text{v2}}$** (Goyal et al., 2017): A standard benchmark requiring open-ended visual question answering grounded in image content.
- **MMB** (Liu et al., 2024): Evaluates 20 ability dimensions using a circular strategy.
- **Seed^I** (Li et al., 2023a): Assesses generative comprehension and spatial relationship understanding via multiple-choice questions.
- **VizWiz** (Gurari et al., 2018): Tests the model's ability to resolve diverse visual queries originating from blind individuals.

4.3 Overall Performance

To answer **RQ1**, we evaluate DAID's effectiveness in mitigating visual hallucinations while preserving core multimodal reasoning capabilities.

Visual Hallucination Reduction. Table 1 compares DAID with existing methods on LLaVA-1.5 and LLaVA-NeXT across visual hallucination benchmarks. DAID consistently outperforms all baselines. On POPE for LLaVA-1.5, DAID achieves 85.08% accuracy and 85.92% F1-score, improving over baselines by up to 3.7% and 3.72%, respectively. On CHAIR, DAID achieves the lowest CHAIR_S (35.9%) and CHAIR_I (11.3%), significantly reducing object hallucinations. DAID also delivers optimal performance on MME. Similar gains are observed on LLaVA-NeXT. These results demonstrate DAID's effectiveness in mitigating hallucinations through reinforced visual grounding and suppressed linguistic priors.

Method	POPE		CHAIR		MME				
	Acc.↑	F1↑	$CHAIR_S$↓	$CHAIR_I$↓	Existence↑	Count↑	Position↑	Color↑	Total↑
LLaVA-1.5-7B									
Greedy	81.38	82.20	49.6	14.4	173.45	122.72	113.39	149.92	559.48
Beam Search	84.66	84.60	46.3	12.9	175.67	124.67	114.00	151.00	565.34
DoLa (Chuang et al., 2023)	84.06	84.62	47.1	13.8	180.10	127.40	119.30	154.60	581.40
VCD (Leng et al., 2024)	84.66	84.52	49.2	14.8	184.66	**137.33**	128.67	153.00	603.66
OPERA (Huang et al., 2024)	84.88	85.21	45.4	12.7	180.67	133.33	111.67	123.33	549.00
SID (Huo et al., 2024)	84.82	85.50	44.2	12.2	183.90	132.20	127.80	155.90	599.80
EAZY (Che et al., 2025)	84.97	85.78	38.8	11.4	182.29	131.73	126.93	155.21	596.16
DAID	**85.08**	**85.92**	**35.9**	**11.3**	**186.14**	130.70	**154.46**	**162.38**	**633.68**
LLaVA-NeXT									
Greedy	83.78	82.24	32.8	9.1	179.78	127.40	121.92	151.82	580.92
Beam Search	83.77	81.69	33.0	9.2	180.27	128.90	120.34	155.29	584.80
DoLa (Chuang et al., 2023)	84.53	84.77	31.3	9.0	184.45	133.07	126.72	160.38	604.62
VCD (Leng et al., 2024)	83.37	83.95	32.8	8.9	187.20	**140.16**	139.91	157.41	624.68
OPERA (Huang et al., 2024)	84.37	84.21	34.0	8.7	183.98	135.85	118.00	131.72	569.55
SID (Huo et al., 2024)	84.61	85.32	32.8	8.3	185.03	134.96	137.23	161.74	618.96
EAZY (Che et al., 2025)	84.91	85.40	26.8	8.3	184.55	134.18	133.52	158.89	611.14
DAID	**85.32**	**85.76**	**24.2**	**8.2**	**189.47**	133.12	**156.29**	**165.52**	**644.40**

Table 1: Performance comparison of DAID against state-of-the-art methods on POPE, CHAIR, and MME benchmarks. The best results are highlighted in **bold**, and the second-best results are underlined.

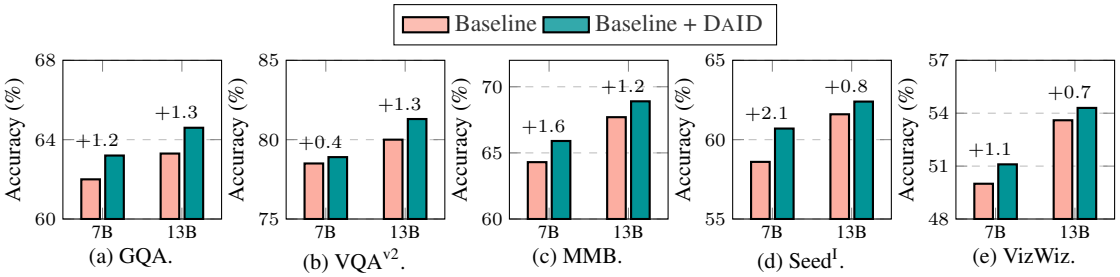


Figure 4: Performance evaluation of LLaVA-1.5-7B/13B across five general vision-language benchmarks.

Reasoning Capability Preservation. We further evaluate whether DAID preserves core multimodal reasoning capabilities across five general vision-language benchmarks. As shown in Figure 4, DAID not only maintains but consistently enhances performance across 7B and 13B scales of LLaVA-1.5. For instance, on Seed^{I}, DAID achieves a +2.1% improvement at the 7B scale. These results indicate that amplifying authentic visual signals while filtering linguistic noise refines the model's representational quality, facilitating robust multimodal understanding beyond error correction.

4.4 Hyperparameter Analysis

To answer **RQ2**, we analyze the impact of α and β on DAID's performance.

Understanding the Visual Perception Reinforcement (α). As shown in Figure 5, performance exhibits an inverted-U pattern across both models, peaking at $\alpha = 0.8$. When $\alpha = 0.4$, both models achieve relatively low performance (e.g., 83.44% accuracy for LLaVA-1.5), indicating insufficient reinforcement. As α increases to 0.6

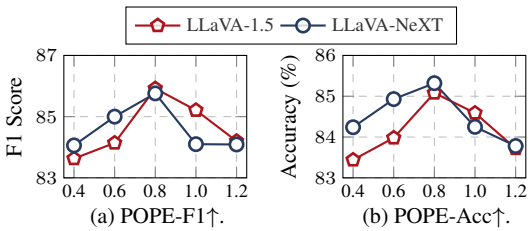


Figure 5: Impact of visual perception reinforcement intensity α on LLaVA-1.5 and LLaVA-NeXT.

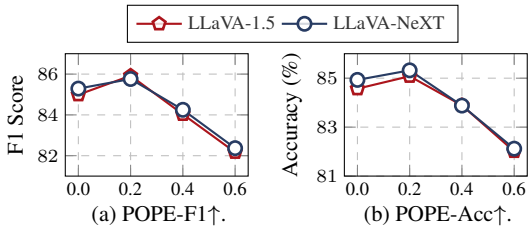
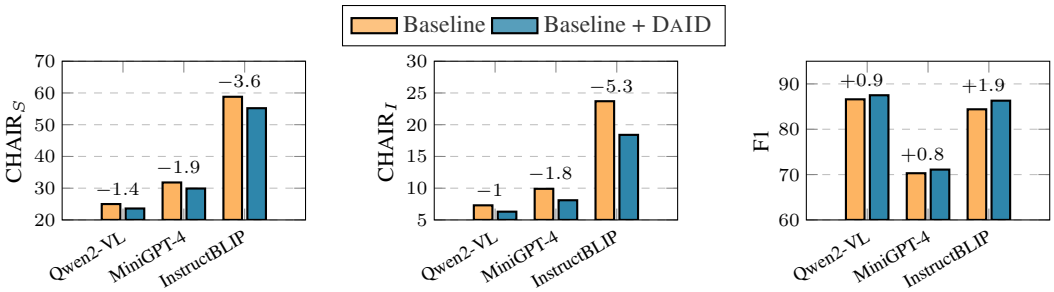


Figure 6: Impact of language prior suppression intensity β on LLaVA-1.5 and LLaVA-NeXT.



(a) CHAIR$_S$↓. (b) CHAIR$_I$↓. (c) POPE-F1↑.

Figure 7: Generalization analysis of DAID across diverse MLLM architectures on CHAIR and POPE benchmarks.

and then to 0.8, performance improves significantly (e.g., LLaVA-1.5 reaches 85.92% F1 and 85.08% accuracy), demonstrating that reintroducing visual signals effectively counteracts the seeing-then-forgetting phenomenon. However, performance degrades when α exceeds 0.8: at $\alpha = 1.0$, F1 scores drop by 0.71% and 1.66%, respectively, while accuracy drops by 0.49% and 1.70%, with the decline intensifying at $\alpha = 1.2$. This occurs because excessive visual signals from the Spotlight layer overpower the final linguistic modeling, disrupting syntactic coherence.

Understanding the Language Prior Suppression (β). Figure 6 shows that performance peaks at $\beta = 0.2$ across both models. When $\beta = 0.0$ (no suppression), both models achieve suboptimal performance (e.g., 84.99% F1 and 84.57% accuracy for LLaVA-1.5) due to unmitigated linguistic priors. As β increases to 0.2, performance improves significantly: LLaVA-1.5 reaches 85.92% F1 (+0.93%) and 85.08% accuracy (+0.51%), while LLaVA-NeXT achieves 85.76% F1 (+0.47%) and 85.32% accuracy (+0.39%), demonstrating that a mild penalty neutralizes visual agnosia without harming valid generation. However, performance declines when β exceeds 0.2: at $\beta = 0.4$, F1 scores drop by 1.51% and 1.89%, while accuracy falls by 1.19% and 1.44%, respectively. The decline intensifies at $\beta = 0.6$, indicating that excessive subtraction over-penalizes the vocabulary space, removing not only hallucinations driven by textual inertia but also valid linguistic connectors.

4.5 Generalization Analysis

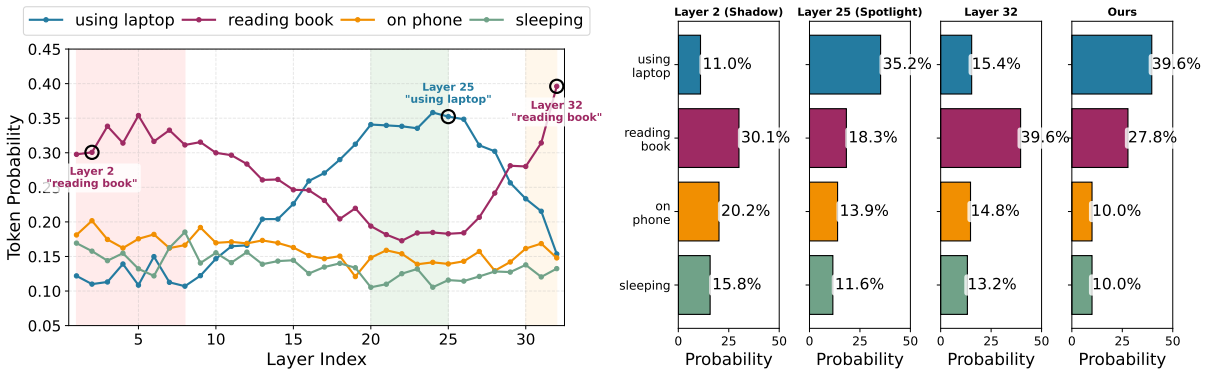
To answer **RQ3**, we evaluate DAID across diverse MLLMs, including Qwen2-VL, MiniGPT-4, and InstructBLIP. As shown in Figure 7, DAID consistently yields significant performance gains across all models and benchmarks. Specifically, DAID markedly suppresses erroneous object generation: CHAIR$_S$ decreases by 3.6% for InstructBLIP and 1.9% for MiniGPT-4 (Figure 7(a)), while CHAIR$_I$ drops by 5.3% for InstructBLIP (Figure 7(b)). Beyond error reduction, DAID consistently enhances grounding capabilities, achieving F1-score improvements of +1.9% for InstructBLIP and +0.9% for Qwen2-VL on POPE (Figure 7(c)). These results demonstrate that our attention-guided dual-anchor mechanism serves as a universal, training-free solution for enhancing MLLM reliability across diverse architectures.

Method	POPE		CHAIR	
	Acc.↑	F1↑	CHAIR$_S$↓	CHAIR$_I$↓
LLaVA-1.5				
+ DAID	**85.08**	**85.92**	**35.9**	**11.3**
w/o PA	84.57	84.99	38.1	12.0
w/o NA	82.79	82.85	48.9	14.3
LLaVA-NeXT				
+ DAID	**85.32**	**85.76**	**24.2**	**8.2**
w/o PA	84.93	85.29	26.7	8.4
w/o NA	83.11	83.52	33.6	9.1

Table 2: Ablation study of DAID. PA and NA denote the positive and negative anchors, respectively.

4.6 Ablation Study

To answer **RQ4**, we analyze the individual contributions of the Spotlight and Shadow anchors through ablation experiments. Table 2 shows that removing either anchor degrades performance. The Spotlight anchor (Positive Anchor) contributes to visual grounding: excluding it (w/o PA) reduces POPE accuracy from 85.08% to 84.57% on LLaVA-1.5, confirming that reinforcing visual perceptual signals is essential to counteract the seeing-then-forgetting phenomenon. The Shadow anchor (Negative Anchor) contributes to hallucination suppression: its absence (w/o NA) causes CHAIR$_S$ to jump from 35.9% to 48.9% on LLaVA-1.5, underscoring its necessity in isolating and penalizing linguistic priors. Similar gains are also observed on LLaVA-NeXT,



(a) Token probability distribution across layers.

(b) Token probabilities for specific layers.

Figure 8: Qualitative case study of DAID on a representative instance (input image and query detailed in Figure 3).

where both anchors contribute significantly to performance improvements. DAID consistently outperforms single-anchor variants across both models, demonstrating that simultaneously amplifying visual signals and suppressing linguistic noise is imperative for effective hallucination mitigation.

4.7 Qualitative Analysis

To answer **RQ5**, we analyze how DAID rectifies token distribution during decoding through a representative case where the model erroneously prioritizes "reading book" over the visually grounded "using laptop". Figure 8(a) shows that the correct token ("using laptop") peaks at layer 25 but is suppressed by "reading book" at the final layer, illustrating how deeper layers prioritize linguistic fluency over visual evidence. During decoding, DAID rectifies this distribution by adaptively designating layer 2 as the Shadow anchor to capture linguistic noise and layer 25 as the Spotlight anchor to amplify visual signals. As shown in Figure 8(b), DAID successfully inverts the erroneous distribution, boosting the correct token's probability from 15.4% to 39.6% while suppressing the hallucinated candidate, demonstrating effective token-level rectification during decoding.

5 Related Work

Hallucination in MLLMs. Hallucination in LLMs is typically defined as the generation of factually incorrect yet syntactically plausible text (Ji et al., 2023; Xu et al., 2026a), and this phenomenon is further compounded in MLLMs where the objective shifts from maintaining internal consistency to ensuring cross-modal fidelity (Bai et al., 2024; Zhao et al., 2025). Despite architectural advancements like InstructBLIP's Q-Former (Dai et al., 2023), MLLMs often exhibit a perceptual-cognitive paradox, prioritizing linguistic plausibility over visual evidence (Huo et al., 2024; Chuang et al., 2023). Recent studies characterize this as degenerative alignment, where visual signals diminish through deeper layers and eventually succumb to linguistic priors (Wang et al., 2024a), ultimately triggering hallucinations that favor probabilistic token sequences over visual grounding.

Contrastive Decoding for Hallucination Mitigation. To mitigate hallucinations without retraining, Contrastive Decoding (CD) (Li et al., 2023b) has emerged as a predominant paradigm. Existing methods fall into two categories: external methods like VCD (Leng et al., 2024) and ICD (Wang et al., 2024c) require manual construction of contrastive samples and additional forward passes, introducing extraneous noise and substantial computational overhead, while internal methods like SID (Huo et al., 2024) and DeCo (Wang et al., 2024a) rely on heuristics or fixed layer selection, failing to adapt to token-specific patterns. In contrast, DAID leverages visual attention as a dynamic proxy to adaptively identify dual anchors, enabling token-aware hallucination mitigation without manual intervention or fixed heuristics.

6 Conclusion

This paper introduces DAID, a novel contrastive decoding framework that mitigates MLLM hallucinations by adaptively identifying dual anchors—a Spotlight layer to amplify visual signals and a Shadow layer to suppress linguistic priors—ensuring visually grounded token generation. Experimental results demonstrate that DAID significantly reduces hallucination rates while enhancing general multimodal reasoning capabilities.

Limitations

Despite the significant improvements in hallucination mitigation and reasoning efficiency, our work has several limitations that warrant further exploration. First, computational hardware constraints limited our evaluation to models at the 7B parameter scale. While DAID consistently demonstrates superior performance, its scalability and internal layer dynamics in ultra-large-scale MLLMs (e.g., those exceeding 70B parameters) remain to be fully investigated. However, given that the seeing-then-forgetting phenomenon is a fundamental characteristic of Transformer-based cross-modal alignment, we expect our dual-anchor mechanism to generalize effectively to larger architectures. Second, our current investigation primarily focuses on image-text multimodal tasks. While DAID effectively mitigates hallucinations by mining internal perceptual discrepancies in image-based MLLMs, its applicability to other modalities, such as video or audio, has not yet been explored. Since temporal dynamics in video understanding introduce more complex attention shifts over time, extending the Dual-Anchor Introspective Decoding framework to handle sequential multimodal data would be a promising direction for future research.

Ethical Considerations

In this work, we present DAID to enhance the factual consistency of MLLMs. All experiments were conducted using publicly available, standard benchmarks (e.g., POPE, CHAIR, MME) that do not contain sensitive personal information or violate privacy. We strictly follow the licenses of all datasets and models used in this work, ensuring compliance with their terms of use and intended purposes. Our method is designed to improve the factual consistency of vision-language models in their intended application domains, and we do not advocate for any use that could cause harm, violate privacy, or be inconsistent with the intended use of the underlying models and datasets.

By reducing hallucinations where generated text contradicts visual content, our method promotes the development of more reliable and honest AI systems for high-stakes scenarios. As a training-free decoding strategy, DAID does not introduce new training data that could exacerbate existing societal or demographic biases inherent in pre-trained models. Furthermore, DAID operates within a single forward pass, significantly reducing computational overhead and carbon footprint compared to traditional contrastive decoding methods. We acknowledge potential risks: while DAID reduces visual-textual inconsistencies, it may not eliminate all forms of hallucinations or errors, and users should exercise appropriate caution in critical applications.

Acknowledgements

This work is supported in part by the Science and Technology Development Fund of Macau (0107/2024/RIA2, 0061/2025/RIB2), Joint Science and Technology Research Project with Hong Kong and Macau in Key Areas of Nansha District's Science and Technology Plan (EF2024-00180-IOTSC) and the Multi-Year Research Grant of University of Macau (MYRG-GRG2023-00211-IOTSC-UMDF, MYRG-GRG2024-00180-IOTSC).

References

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505.

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

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Personalizing multimodal large language models for image captioning: an experimental analysis. In *European Conference on Computer Vision*, pages 351–368. Springer.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. 2025. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21635–21644.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, et al. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

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

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

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

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

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Benji Peng, Qian Niu, Junyu Liu, Jinlang Wang, Sen Zhang, Xuanhe Pan, et al. 2024b. Surveying the mllm landscape: A meta-review of current surveys. *arXiv preprint arXiv:2409.18991*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Jianing Qiu, Wu Yuan, and Kyle Lam. 2024. The application of multimodal large language models in medicine. *The Lancet Regional Health–Western Pacific*, 45.

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

Sara Sarto, Marcella Cornia, and Rita Cucchiara. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29904–29914.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2024a. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 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. *arXiv preprint arXiv:2403.18715*.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025a. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8460–8468.

Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. 2025b. Elastic mixture of rank-wise experts for knowledge reuse in federated fine-tuning. *arXiv preprint arXiv:2512.00902*.

Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. 2026a. Developmental federated tuning: A cognitive-inspired paradigm for efficient llm adaptation. In *The Fourteenth International Conference on Learning Representations*.

Yebo Wu, Jingguang Li, Chunlin Tian, Zhijiang Guo, and Li Li. 2025c. Memory-efficient federated fine-tuning of large language models via layer pruning. *arXiv preprint arXiv:2508.17209*.

Yebo Wu, Jingguang Li, Chunlin Tian, Kahou Tam, Zhijiang Guo, and Li Li. 2026b. Beyond end-to-end: Dynamic chain optimization for private llm adaptation on the edge. *Preprint*, arXiv:2604.06819.

Yebo Wu, Li Li, and Cheng-zhong Xu. 2025d. Breaking the memory wall for heterogeneous federated learning via progressive training. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1623–1632.

Yebo Wu, Feng Liu, Ziwei Xie, Zhiyuan Liu, Changwang Zhang, Jun Wang, and Li Li. 2026c. Tsembed: Unlocking task scaling in universal multimodal embeddings. *arXiv preprint arXiv:2603.04772*.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. 2026a. When agents" misremember" collectively: Exploring the mandela effect in llm-based multi-agent systems. *arXiv preprint arXiv:2602.00428*.

Naen Xu, Changjiang Li, Tianyu Du, Minxi Li, Wenjie Luo, Jiacheng Liang, Yuyuan Li, Xuhong Zhang, Meng Han, Jianwei Yin, et al. 2024. Copyrightmeter: Revisiting copyright protection in text-to-image models. *arXiv preprint arXiv:2411.13144*.

Naen Xu, Jiayi Sheng, Changjiang Li, Chunyi Zhou, Yuyuan Li, Tianyu Du, Jun Wang, Zhihui Fu, Jinbao Li, and Shouling Ji. 2026b. "i see what you did there": Can large vision-language models understand multimodal puns? *Preprint*, arXiv:2604.05930.

Naen Xu, Jinghuai Zhang, Ping He, Chunyi Zhou, Jun Wang, Zhihui Fu, Tianyu Du, Zhaoxiang Wang, and Shouling Ji. 2026c. Fraudshield: Knowledge graph empowered defense for llms against fraud attacks. *arXiv preprint arXiv:2601.22485*.

Naen Xu, Jinghuai Zhang, Changjiang Li, Hengyu An, Chunyi Zhou, Jun Wang, Boyu Xu, Yuyuan Li, Tianyu Du, and Shouling Ji. 2026d. Bridging the copyright gap: Do large vision-language models recognize and respect copyrighted content? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35949–35957.

Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. 2025. Videoeraser: Concept erasure in text-to-video diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5965–5994.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2025. Look-back: Implicit visual re-focusing in mllm reasoning. *arXiv preprint arXiv:2507.03019*.

Xiangtao Zhang, Eleftherios Kofidis, Ruituo Wu, Ce Zhu, Le Zhang, and Yipeng Liu. 2026. Coupled tensor train decomposition in federated learning. *Pattern Recognition*, 170:112067.

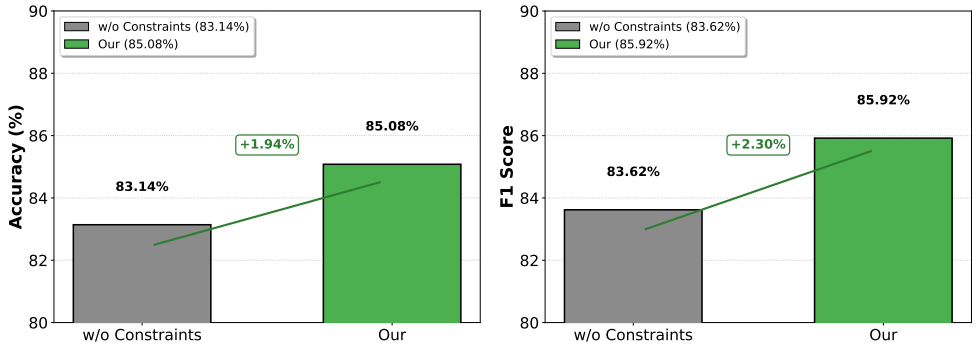
Xiangtao Zhang, Sheng Li, Ao Li, Yipeng Liu, Fan Zhang, Ce Zhu, and Le Zhang. 2025a. Subspace constraint and contribution estimation for heterogeneous federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20632–20642.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025b. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3512–3534.

Pengfei Zhao, Rongbo Luan, Wei Zhang, Peng Wu, and Sifeng He. 2025. Guiding cross-modal representations with mllm priors via preference alignment. *arXiv preprint arXiv:2506.06970*.

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

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. 2025. Mitigating object hallucinations in large vision-language models via attention calibration. *arXiv preprint arXiv:2502.01969*.



(a) POPE-Acc. (b) POPE-F1.

Figure 9: Analysis of the topological constraint ($L_{shad.} < L_{spot.}$) on LLaVA-1.5-7B. We compare the performance between the constrained and unconstrained settings.

A Discussion and Analysis

A.1 Averaging Attention for Anchor Selection

Our method computes the Visual Attention Score (VAS) by averaging attention patterns across all heads within each layer. While individual heads may exhibit diverse behaviors, we adopt this layer-level strategy based on three key justifications:

- **Alignment with Decoding Granularity.** As the decoding process operates on layer representations—which aggregate outputs from all heads—selecting anchors at the layer level ensures structural consistency with the model's natural information flow (Wu et al., 2026a, 2025c). The head-averaged VAS effectively captures the collective visual grounding, mirroring how multi-head information is integrated during inference.

- **Robustness via Consensus.** Averaging across heads distills a consensus visual attention signal while mitigating spurious patterns from individual heads. This robustness is critical for our dynamic, token-specific selection mechanism, ensuring stability across diverse generation contexts (Zhang et al., 2026, 2025a).

- **Empirical Effectiveness.** Despite head-level variations, our VAS successfully identifies layers with peak visual perception (Section 2) and delivers significant performance gains across all benchmarks, demonstrating that the consensus signal effectively captures the essential visual grounding information needed for anchor selection.

A.2 Constraints on Shadow Layer Selection

We enforce a topological constraint requiring the Shadow layer to precede the Spotlight layer ($L_{shad.} < L_{spot.}$). This design is grounded in both theoretical rationale and empirical evidence.

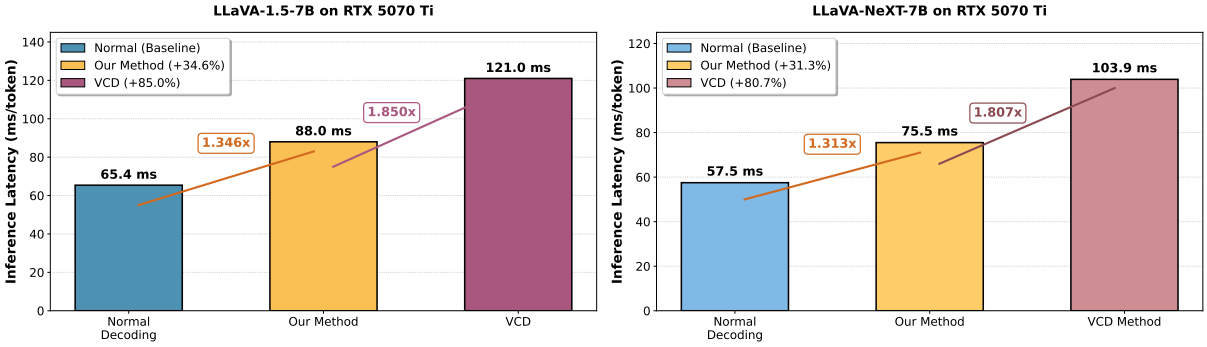
- **Conceptual Rationale.** This constraint guarantees that the Shadow layer captures a pre-perceptual state (Wu et al., 2026b, 2025d), where visual signals are minimal and linguistic priors dominate. This ordering aligns with the model's internal progression from visual agnosia (in shallow layers) to peak perception (in intermediate layers). By enforcing this precedence, we isolate pure linguistic noise rather than the semantic drift found in deeper layers, which is critical for maximizing the efficacy of contrastive decoding.

- **Empirical Validation.** Figure 9 demonstrates the necessity of this topology: relaxing the restriction ("w/o Constraints") noticeably degrades performance, whereas applying our constraint boosts POPE accuracy from 83.14% to 85.08% (+1.94%) and F1 from 83.62% to 85.92% (+2.30%). These results confirm that anchoring the negative reference to shallow, visually agnostic layers is essential for effective hallucination mitigation.

A.3 Analysis of Computational Efficiency

Our approach introduces minimal computational overhead compared to standard decoding while maintaining significantly higher efficiency than external perturbation-based methods (e.g., VCD).

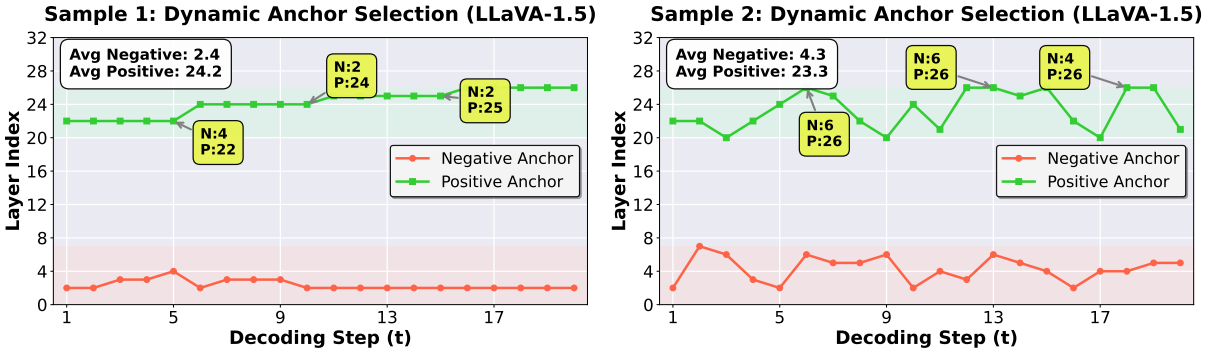
- **Theoretical Analysis.** Standard decoding computes attention weights and hidden states



(a) Inference Latency on LLaVA-1.5-7B.

(b) Inference Latency on LLaVA-NeXT-7B.

Figure 10: Inference latency comparison (ms/token) on NVIDIA RTX 5070 Ti. We evaluate the decoding speed of normal decoding (Baseline), our method, and VCD on LLaVA-1.5-7B and LLaVA-NeXT-7B.



(a) Sample 1 (Stable Progression).

(b) Sample 2 (High Variability).

Figure 11: Visualization of token-specific dynamic anchor selection on LLaVA-1.5. We track the layer indices selected for the Spotlight Anchor and Shadow Anchor across 20 decoding steps.

for all layers in a single forward pass. DAID incurs only two minor sources of overhead: 1) averaging attention weights for VAS (negligible as it reuses intermediate computations), and 2) projecting the hidden states of the selected Shadow and Spotlight layers to the vocabulary space. Crucially, this design circumvents the computational bottleneck of methods like VCD, which necessitate a second full forward pass on perturbed inputs. Consequently, DAID maintains a single-pass inference regime ($\approx 1\times$ cost), whereas VCD doubles the computational cost ($\approx 2\times$).

- **Practical Efficiency.** We evaluate inference latency on an NVIDIA RTX 5070 Ti (Figure 10). DAID incurs moderate overhead (approx. 1.31–$1.35\times$) compared to the baseline, primarily due to multi-layer logit projections. However, this is significantly more efficient than VCD, which increases latency by over $1.8\times$. On LLaVA-NeXT, DAID achieves **75.5 ms/token**, 37.6% faster than VCD (103.9 ms/token), demonstrating that mining internal states is effective and efficient.

A.4 Rationale for γ Selection

The adaptive plausibility constraint utilizes distinct γ values to align with the vocabulary distributions of different benchmarks. For discriminative benchmarks like POPE, the probability mass is heavily concentrated on a few tokens (e.g., Yes/No). A high threshold ($\gamma = 0.9$) ensures that the constraint remains tight, focusing only on the most confident predictions. In contrast, open-ended tasks (e.g., CHAIR) exhibit dispersed probability distributions across a wide vocabulary. A lower threshold ($\gamma = 0.1$) allows the method to operate on a wider plausible search space, balancing diversity with the suppression of implausible tokens.

A.5 Token-Specific Anchor Selection

The variability in anchor choices across tokens in our dynamic selection mechanism is beneficial and well-controlled through several mechanisms.

- **Context-Dependent Adaptation.** Layer se-

lection is highly responsive to context, reflecting varying perceptual requirements. As shown in Figure 11(a), tokens requiring consistent visual grounding exhibit natural smoothness. Conversely, Figure 11(b) illustrates that when processing complex visual semantics, the model demonstrates adaptive agility, shifting anchors (e.g., oscillating between layers 20 and 26) to calibrate specifically for individual tokens. This fluctuation reflects DAID's fine-grained responsiveness.

- **Operational Zone Stability.** Despite index-level fluctuations, the functional roles of anchors remain consistent. As visualized by the shaded regions in Figure 11, the Shadow Anchor is confined to the visual agnosia zone (shallow layers, indices 0–7), while the Spotlight Anchor operates in the peak perception zone (intermediate layers, indices 20–26). This ensures that the contrastive signal, suppressing linguistic noise and amplifying visual features, remains stable throughout generation, regardless of specific layer indices.

- **Linguistic Coherence Guarantee.** The adaptive plausibility constraint (Section 3) safeguards coherence by restricting decoding adjustments to the linguistically valid candidate set $\mathcal{V}'$. Consequently, even if internal layer selections vary dynamically, the final output tokens remain syntactically and semantically consistent with the preceding context.

A.6 Metric for Anchor Selection

We choose Visual Attention Score (VAS) as our metric for layer selection, motivated by several key considerations over alternative internal signals.

- **Direct Interpretability.** VAS serves as a direct, training-free proxy for visual reliance, explicitly measuring the attention weight assigned to image tokens. This avoids the semantic entanglement often present in hidden state representations.

- **Computational Efficiency.** VAS is computationally free, utilizing attention maps generated during standard forward passes. It avoids the latency overhead associated with computing complex vector similarities or gradients.

- **Empirical Validation.** As detailed in Section 2, VAS strongly correlates with visual perception accuracy. The alignment between peak VAS and optimal object recognition layers confirms its reliability.

A.7 Impact of Anchor Cardinality

Our method employs two anchors (Spotlight and Shadow), balancing effectiveness and efficiency through complementary mechanisms.

- **Synergistic Roles.** Ablation studies (Section 4.6) confirm that both anchors are indispensable: the Spotlight anchor amplifies visual evidence (+0.51% accuracy), while the Shadow anchor attenuates linguistic priors (-13% CHAIR_S). Removing either component degrades performance, validating that coupling positive visual reinforcement with negative linguistic suppression is essential for effective mitigation.

- **Mechanistic Rationale.** The dual-anchor design aligns with contrastive learning principles: amplifying positive signals (visual grounding) while suppressing negative signals (linguistic noise). This configuration avoids the insufficiency of single-anchor methods while circumventing the diminishing returns associated with multi-anchor complexity.

A.8 Robustness Analysis

While motivated by the seeing-then-forgetting phenomenon, DAID is not strictly bound to its intensity. The method operates on the relative differential of visual attention (ΔVAS) rather than absolute layer indices. Even in instances where the visual attenuation is subtle (a flat forgetting curve), our dynamic selection maximizes the available contrast by locating the optimal local peaks and valleys. This property, combined with the universality of attention mechanisms, ensures robustness across varied architectures and diverse input complexities without requiring heuristic tuning.

B The Use of Large Language Models

In the preparation of this manuscript, we utilized LLMs exclusively for linguistic refinement and editing to improve the clarity and readability of the text. We affirm that the conceptualization, methodology, experimental design, and analysis presented in this work are entirely original and were not generated by AI tools. All modifications suggested by the LLMs were critically reviewed and verified by the authors.

Method	GQA	VQA^{v2}	Seed^{I}	VizWiz
Qwen3VL-2B				
Baseline	60.9	79.5	60.5	47.2
DAID	**61.7**	**80.3**	**61.2**	**47.7**

Table 3: Performance comparison on Qwen3VL-2B.

C More Experimental Results

We also present experimental results on Qwen3VL-2B in Table 3. As shown in the table, DAID consistently outperforms the baseline across all four benchmarks. These results indicate that our method remains highly effective on recent models, further demonstrating its strong generalization ability across different MLLM architectures.