

# ADAPT: Attention Dynamics Alignment with Preference Tuning for Faithful MLLMs

Zhiyuan Yao<sup>1,3</sup>, Zheren Fu<sup>1</sup>, Zhixiao Zheng<sup>1</sup>,  
Jiajun Li<sup>2</sup>, Yi Tu<sup>2</sup>, and Zhendong Mao<sup>1,3\*</sup>

<sup>1</sup> University of Science and Technology of China, Hefei, China

<sup>2</sup> Huawei Technologies Ltd., China

<sup>3</sup> State Key Laboratory of Communication Content Cognition, People’s Daily Online, China

{yaozhiyuan,zhixiao.zheng}@mail.ustc.edu.cn, {fzr,zdmao}@ustc.edu.cn  
{jiajun.work,xsg.tuyi}@huawei.com

**Abstract.** Multimodal Large Language Models (MLLMs) are critically hampered by hallucination—generating content inconsistent with the provided image. In this paper, we identify an internal signature of hallucination: progressive degradation of text-to-image cross-attention during generation, leading to specific failure patterns like unfocused or biased attention. Existing mitigation strategies are largely outcome-driven and do not explicitly target this failure mode. To address this problem, we propose ADAPT (Attention Dynamics Alignment with Preference Tuning), an attention-based framework that intervenes directly on text-to-image cross-attention dynamics. We propose ADAPT with three key contributions: a cross-attention visual anchor refined from early decoding to provide stable spatial grounding, an attention-supervised inference mechanism that detects and corrects attention drift online, and a Visual Attention Guidance DPO that aligns preferences toward visually grounded responses. Experiments show that each component of ADAPT contributes to hallucination reduction, and the full framework achieves new best results across multiple hallucination benchmarks, reducing hallucination rates by 40–60% across mainstream backbones while preserving general multimodal capabilities. Our work provides an attention-based perspective on mitigating hallucinations by exploring the model’s internal text-to-image cross-attention behaviors. Code is available at: <https://github.com/yao-ustc/ADAPT>.

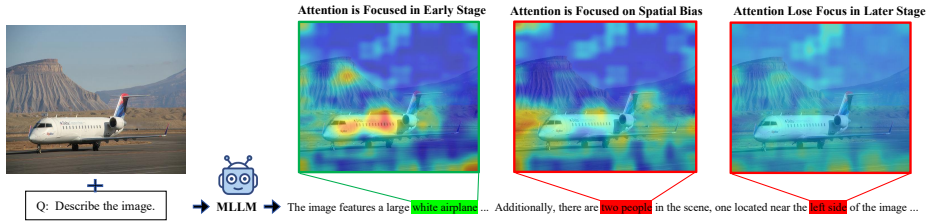
**Keywords:** MLLMs · Hallucination Mitigation · Cross-Attention

## 1 Introduction

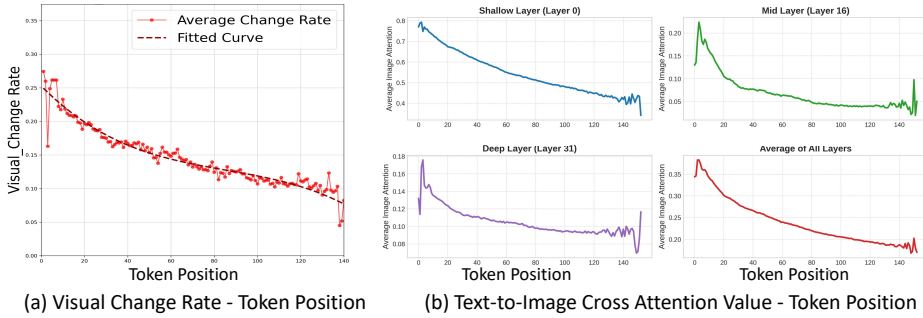
Multimodal Large Language Models (MLLMs) often produce hallucinations, which means generating descriptions or answers that are not grounded in the input image [2, 14]. Hallucinations can take the form of asserting the presence of

---

\* Corresponding author.



**Fig. 1: Cross-attention degradation is associated with hallucinations.** Text-to-image cross-attention is first aligned with query-relevant evidence (the airplane), then collapsing onto spatially biased regions (the left side of the airplane) and later becoming diffuse (across the scene), where hallucinated tokens appear.



**Fig. 2: Cross-attention degradation during generation.** (a) The cross-attention change rate to newly generated text decreases as generation proceeds, dropping from about 25% to below 10%. (b) The average attention value on visual tokens decays across layers as decoding proceeds, with more than a 40% reduction across layers, reflecting reduced reliance on visual evidence.

objects that do not exist, attributing wrong properties such as color or count, or inventing relationships that are not visually supported, which severely undermines the trustworthiness of MLLMs. This limitation persists even in commonly used models such as GPT-5 [25], Qwen-VL [1], and LLaVA [16].

Current efforts to mitigate MLLM hallucination mainly concentrate on two directions: inference-time interventions and training-time alignment. Inference-time methods apply plug-and-play control signals during generation, including contrastive decoding [11], ensemble decoding [3], and penalty-based strategies [7]. These methods primarily adjust the model outputs by manipulating token-level probabilities during decoding, aiming to steer generation toward more faithful responses. Training-time alignment methods improve visual faithfulness by updating model parameters through fine-tuning objectives, ranging from RLHF [4] to more efficient alternatives such as DPO [21], with various vision-aware variants [31, 35, 39, 42]. These methods strengthen preference alignment by shaping the model’s output distribution toward human preferences for faithful descriptions. Overall, inference-time methods adjust token-level logits and training-time alignment supervises token predictions, both of which are outcome-driven.



In contrast, internal state dynamics such as text-to-image cross-attention are closely tied to hallucinations, yet this connection is less studied.

Therefore, we analyze the model’s text-to-image cross-attention as a key internal signal for tracing hallucinations. We find that cross-attention first anchors on salient, text-relevant visual evidence and then gradually degrades during generation, shifting the model toward language priors and raising the probability of hallucinated outputs. This degradation manifests in three aspects. (1) Cross-attention drifts away from text-relevant visual evidence and becomes unfocused or spatially biased (Fig. 1). (2) Cross-attention changes less in response to newly generated text, with a decreasing change rate as generation proceeds (Fig. 2a). (3) Cross-attention value on visual tokens drops across layers, indicating reduced reliance on visual evidence (Fig. 2b). Together, these three signals show that cross-attention becomes less aligned with visual evidence, less responsive to new tokens, and less weighted on image tokens as generation proceeds, making hallucinations increasingly probable in the later part of the output (Fig. 3).

To address this issue, we first compute a reliable, semantics-aware cross-attention anchor from early grounded stage and then use it as a supervision signal to monitor and steer cross-attention in later prior-dominated stage, maintaining semantic consistency between the generation and the image evidence. Building on this design, we propose **ADAPT (Attention Dynamics Alignment with Preference Tuning)**, an attention-anchor-based framework that mitigates attention degradation during long-form generation. Our framework consists of three synergistic stages. We first perform Visual Enhance by extracting early, reliable text-to-image cross-attention, refining it into an anchor  $A_{\text{anchor}}$ . We then apply Attention-Supervised Inference by monitoring cross-attention against the anchor and using sparse corrective steering when attention degrades and becomes unfocused or spatially biased. Finally, we perform Visual Attention Guidance DPO by tuning preferences under different visual evidence conditions, using anchor-enhanced images as chosen inputs and noise images as rejected inputs to encourage visually grounded outputs. Therefore, ADAPT refines multi-layer text-to-image cross-attention into anchors and couples anchor-guided attention supervision with preference tuning.

Our contributions are fourfold: (1) We systematically reveal the two-stage attention degradation process behind MLLM hallucinations, identifying key failure patterns. (2) We propose a cross-attention refinement method that uses three complementary criteria with positional bias correction to produce reliable cross-attention anchors for downstream supervision and alignment. (3) We design an attention supervision based framework ADAPT, which performs online attention drift correction during inference and reinforces grounding through visual cross attention guidance preference alignment. (4) Experiments show that ADAPT consistently outperforms strong baselines across benchmarks and backbones.

## 2 Related Work

### 2.1 Hallucination of MLLMs

Hallucination occurs when MLLM outputs are inconsistent with the input image [2, 14]. Existing mitigation methods mainly follow two directions: training-time alignment and inference-time interventions.

**Training-time alignment.** Preference optimization adapts RLHF/DPO-style objectives to align model outputs with visual facts, including DPO [21] and vision-aware variants such as mDPO [31], CLIP-DPO [19], SIMA [33], and RLHF-V [38]. HDPO [6] further constructs hallucination-targeted preference pairs to better cover diverse failure patterns.

#### Inference-time interventions.

Inference-time methods reduce hallucinations in a plug-and-play manner by modifying decoding [12, 18]. Contrastive decoding improves specificity by comparing alternative distributions [11]. IBD [43] amplifies image-biased attention to counter language-prior over-reliance, while SID [8] introduces token-level disturbances for efficient suppression. DeCo [30] leverages intermediate signals for dynamic correction during generation, and Layer Contrastive Decoding [29] contrasts layer-wise predictions to suppress hallucinated outputs.

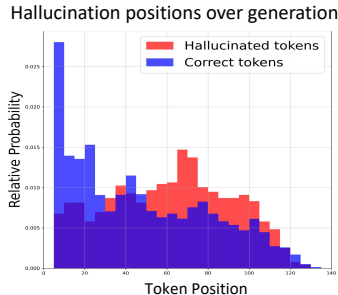
## 2.2 Cross-attention in MLLMs

Diagnostic studies connect anomalous attention behaviors, such as attention sinks, with failures of visual grounding [9, 10]. Accordingly, attention-based interventions manipulate attention heads or calibrate attention to encourage visual focus. For example, SPIN [24] suppresses heads with low cumulative attention to image tokens, and AVISC [34] explores attentional calibration for improved faithfulness. Vision-guided attention methods further steer the model toward informative regions during generation [41].

ViT-based MLLMs also exhibit attention decay to visual tokens and positional biases over long sequences. CCA [36] reduces effective visual-text distance via token rearrangement and causal masking, FarSight [28] improves masking and positional handling for long-range reasoning, and MemVR [44] re-injects visual tokens as key-value memory under uncertainty. Beyond architecture-level changes, data and signal augmentation can further strengthen grounding, such as phrase-level alignment with augmented data [23].

## 3 Cross-Attention and Hallucinations

We first conducted an in-depth exploration of the text-to-image cross-attention behaviors by which MLLMs generate hallucinations. Through meticulous obser-



**Fig. 3: Hallucination positions over generation.** We compute the relative probability of hallucinatory and correct tokens across token positions and find that hallucination probability increases markedly in the later stage of generation.

vation of the model’s attention behavior and output content, we identified two distinct stages in the generation process and three recurring degradation signals that occur when hallucinations arise.

### 3.1 Cross-Attention Degradation

We observe a two-stage shift in visual dependency during generation, from an early Visually Grounded stage to a later Prior-Dominated stage, reflected by progressive degradation of text-to-image cross-attention. This degradation manifests in three aspects:

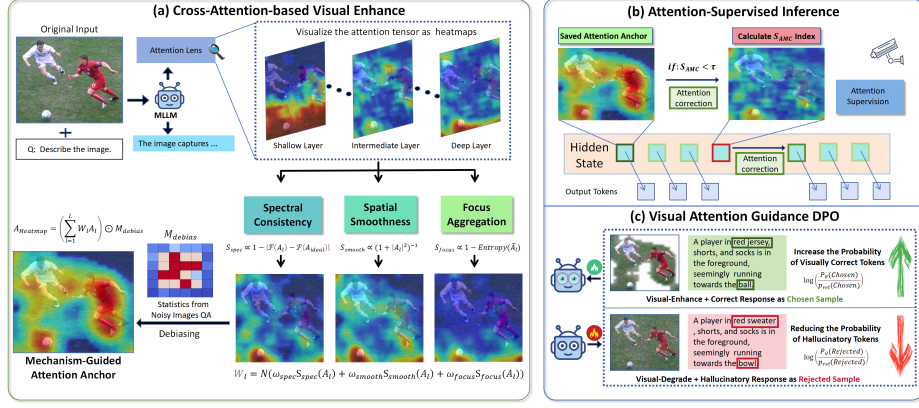
- (1) Cross-attention progressively drifts away from text-relevant visual evidence and becomes unfocused or spatially biased, meaning that attention is either spread broadly across many visual tokens or concentrated on irrelevant positions. In both cases, next-token prediction is empirically no longer constrained by query-relevant regions, so visually unsupported tokens can be selected even when the image provides contrary evidence (Fig. 1).
- (2) Cross-attention changes less in response to newly generated text. This indicates that attention updates are no longer driven by new context, and the model fails to re-locate when later tokens require different image regions (Fig. 2a).
- (3) The overall attention value on visual tokens drops rapidly across layers and token positions. This redistribution weakens the visual constraint on decoding and makes language priors more influential (Fig. 2b).

Together, these signals indicate cross-attention degradation, mark the shift to the prior-dominated stage, and explain why hallucinations become more frequent in the later portion of the output (Fig. 3).

### 3.2 Limitations of Native Cross-Attention

Beyond the two-stage degradation and its failure patterns, we identify a structural limitation of native text-to-image cross-attention in Transformer-based MLLMs: attention sinks [9, 10]. In AMBER, we observe that even when the visual input is uninformative (e.g., pure black/white images or heavily corrupted noise), cross-attention can still become sharply concentrated. This concentration is therefore not necessarily evidence-driven; instead, it often collapses onto a few recurring tokens/positions, revealing a content-independent positional bias. Such sink-induced bias can falsely signal high visual confidence and consequently misguide autoregressive decoding, making native cross-attention unreliable as a grounding signal.

Motivated by this limitation, we explicitly *debias* cross-attention before using it for supervision or steering. Specifically, we estimate a model-intrinsic cross-attention bias map from noise images (Eq. 7), derive a debiasing mask to suppress sink-prone positions (Eq. 8), and apply it to obtain a refined anchor (Eq. 9). This correction reduces spurious concentration on sink tokens, yielding anchors that better reflect query-relevant visual evidence and enabling more reliable supervision and generation.



**Fig. 4: Overview of ADAPT.** (a) **Cross-Attention-based Visual Enhance.** We extract cross-attention from the first  $K$  tokens, refine it with three complementary modules, and apply debiasing to obtain a stable attention anchor. (b) **Attention-Supervised Inference.** We compute  $S_{AMC}$  at each step; when  $S_{AMC} < \tau$ , we trigger sparse attention correction. (c) **Visual Attention Guidance DPO (VAG DPO).** We construct preference pairs under different visual signal strengths, using anchor-enhanced inputs as chosen samples and noise inputs as rejected samples, encouraging visually grounded generation.

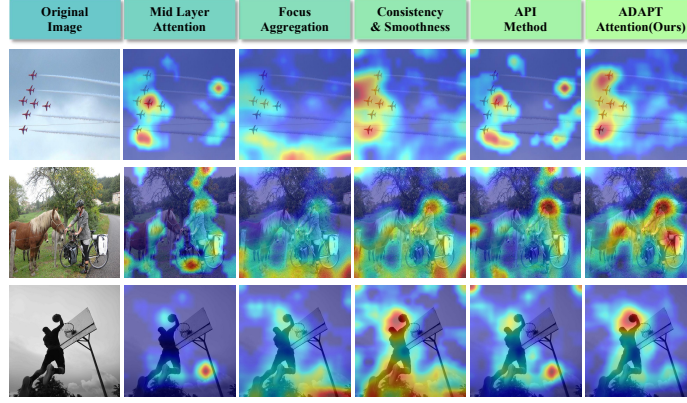
## 4 Methodology

Building on the Cross-Attention Visual Anchor, we propose **ADAPT (Attention Dynamics Alignment with Preference Tuning)**, an attention driven framework that uses a shared visual anchor across coordinated components, as shown in Fig. 4. ADAPT intervenes at three stages: a lightweight pre-inference visual enhancement step, an inference-time attention supervision module, and a training-time preference optimization module. Among them, the core interventions are Attention-Supervised Inference and Visual Attention Guidance DPO, which address hallucination during decoding and alignment.

### 4.1 Cross-Attention-based Visual Enhance

Motivated by Section 3, we treat text-to-image cross-attention as a controllable internal signal for hallucination mitigation. Since raw cross-attention is noisy and position-biased, and a reliable reference must be query-dependent, we refine early multi-layer cross-attention into a Cross-Attention Visual Anchor via complementary criteria and explicit debiasing, yielding a stable, query-relevant anchor for supervision and preference tuning.

To obtain a stable, factually grounded reference, we construct a cross-attention visual anchor from early-generation cross-attention. Let  $\mathcal{A} = \{A_1, \dots, A_L\}$  be the 2D cross-attention maps from  $L$  Transformer layers, where  $A_l \in \mathbb{R}^{H \times W}$  is conditioned on the query  $Q$  and the initial prefix. We score each  $A_l$  with



**Fig. 5: Qualitative comparison of cross-attention anchors** generated by ADAPT against various baselines, including strategies in Section 4.2 and the API method [37]. All heatmaps are obtained by fusing multi-layer query-to-image cross-attention.

three criteria targeting common failures—noise, unfocused attention, and query-mismatched concentration  $S(A_l, Q) \rightarrow [0, 1]^3$ , yielding  $\mathbf{s}_l = [S_{\text{spec}}, S_{\text{smooth}}, S_{\text{focus}}]^\top$ , and fuse the maps accordingly.

**Dynamic Fusion Weight Calculation.** Unreliable attention maps often exhibit excessive high-frequency components, which correspond to visually noisy and unstable grounding. We quantify this property using the high-frequency energy ratio:

$$\Psi(A_l) = \frac{\sum_{(u,v) \in \Omega_H} |\mathcal{F}(A_l)_{u,v}|^2}{\sum_{u,v} |\mathcal{F}(A_l)_{u,v}|^2 + \epsilon}, \quad (1)$$

where  $\mathcal{F}(\cdot)$  is the 2D FFT and  $\Omega_H$  denotes the high-frequency band. We then score deviation from a layer-wise reference  $\mu_l^{\text{spec}}$ :

$$S_{\text{spec}}(A_l) = \exp(-\gamma_1 |\Psi(A_l) - \mu_l^{\text{spec}}|). \quad (2)$$

To suppress fragmented or scattered activations (the unfocused pattern) and favor contiguous object regions, we compute an inverse TV regularizer on the normalized map  $\hat{A}_l$ :

$$S_{\text{smooth}}(A_l) = \left( 1 + \lambda_{\text{TV}} \frac{1}{HW} \sum_{i,j} \|\nabla \hat{A}_l(i,j)\|_2 \right)^{-1}. \quad (3)$$

Over-diffuse attention may ignore key evidence, while over-concentrated attention can miss required context; importantly, the desired focus depends on the query. Let  $\bar{H}(\hat{A}_l)$  denote the normalized Shannon entropy of  $\hat{A}_l$ . We define a query-dependent target entropy

$$\bar{H}^*(Q) = (1 - g(Q))\bar{H}_{\text{obj}} + g(Q)\bar{H}_{\text{global}}, \quad (4)$$

and score by deviation:

$$S_{\text{focus}}(A_l, Q) = 1 - \left| \bar{H}(\hat{A}_l) - \bar{H}^*(Q) \right|. \quad (5)$$

After obtaining  $\mathbf{s}_l$ , we compute layer-wise fusion weights via a linear projection followed by Softmax normalization:

$$W_l = \frac{\exp(\boldsymbol{\omega}^\top \mathbf{s}_l)}{\sum_{k=1}^L \exp(\boldsymbol{\omega}^\top \mathbf{s}_k)}, \quad (6)$$

where  $\boldsymbol{\omega} = [\omega_{\text{spec}}, \omega_{\text{smooth}}, \omega_{\text{focus}}]^\top$  controls the relative importance of the criteria. This emphasizes layers whose attention better matches the above grounding properties. The fused attention map is  $\sum_{l=1}^L W_l \cdot A_l$ .

**Spatial Bias Correction.** Transformer-based MLLMs can exhibit content-independent positional priors, where certain spatial tokens are systematically over-attended regardless of image content (the biased pattern). To estimate this content-free bias, we feed noise images while keeping the same questions and average the resulting fused attention maps:

$$M_{\text{bias}} = \mathbb{E}_{(I, Q) \sim \mathcal{D}_{\text{SVA}}} [\text{Normalize}(A_{\text{fused}}(I_{\text{noise}}, Q))], \quad (7)$$

where  $\text{Normalize}(\cdot)$  rescales the map to sum to 1. We then construct a debiasing mask by applying an element-wise inverse (with smoothing) followed by renormalization:

$$M_{\text{debias}} = \text{Normalize}((M_{\text{bias}} + \epsilon)^{-1}). \quad (8)$$

Finally, we apply the debiasing mask  $M_{\text{debias}}$  to the fused multi-layer cross-attention map via element-wise multiplication, removing layer-aggregated positional bias and yielding the debiased visual anchor:

$$A_{\text{anchor}} = \left( \sum_{l=1}^L W_l \cdot A_l \right) \odot M_{\text{debias}}. \quad (9)$$

The resulting anchor is conditioned on the query semantics and shifts to different image regions as the question changes. It is also layer-selective and corrected for positional bias, and is used for both inference-time attention supervision and preference-pair construction.

**Hyperparameter Selection for Fusion Weights.** We select  $\boldsymbol{\omega}$  via a constrained grid search on AMBER with  $\omega_k \geq 0$  and  $\sum_k \omega_k = 1$  using the Chair metric, and use the same default across backbones:  $\omega_{\text{spec}} = 0.4$ ,  $\omega_{\text{smooth}} = 0.3$ ,  $\omega_{\text{focus}} = 0.3$ . As visualized in Fig. 7(a), the 3D surface over the weight simplex exhibits a broad low-score region, which allows the hyperparameter choice to be inspected directly and indicates low sensitivity around the selected setting. Moreover, the optimum does not collapse to a single criterion weight, suggesting that the three refinement criteria provide complementary benefits in forming a reliable cross-attention anchor. Full results are provided in the Appendix.

**Anchor-Guided Visual Enhancement.** As a lightweight pre-inference step, we use the visual anchor to create a visually grounded input that reduces the risk of later attention drift. Specifically, we first perform a short preliminary inference run (the first  $K$  tokens) during the reliable Strong Visual Anchoring Phase to obtain the anchor  $A_{\text{anchor}}$ . We apply the attention heatmap and a mild binary mask to the original image, producing an anchor-guided image that remains visually interpretable for manual inspection of anchor quality. During extended QA processes, critical visual anchors are generated first; these anchors are subsequently fed back into the model as Visual Enhance to guide the rest of the response generation. Furthermore, these masked images are stored for later training, providing the model with explicit visual cues about which regions should be treated as important.

## 4.2 Attention-Supervised Inference

To maintain visual grounding as decoding transitions to the Prior-Dominated stage, we introduce Attention-Supervised Inference. While hallucinations may also stem from other factors, we treat attention degradation as a practical, model-internal signal for grounding. Rather than forcing cross-attention to strictly follow a fixed anchor, our module monitors attention fidelity online and intervenes only when substantial degradation is detected, applying sparse corrective steering to recover from unfocused or spatially biased drift.

At each decoding step  $t$ , we evaluate the alignment between the current cross-attention distribution  $A_t \in \mathbb{R}^N$  and the previously synthesized visual anchor  $A_{\text{anchor}}$ . Importantly, we use the anchor as a reference for detection, not as a hard target: the model is free to adapt attention as the textual context evolves, and we intervene only when attention drifts toward unfocused or spatially biased attention. We define the Anchor-Modulated Concentration (AMC) index  $S_{\text{AMC}}$  as an online attention-fidelity score for triggering such selective corrections:

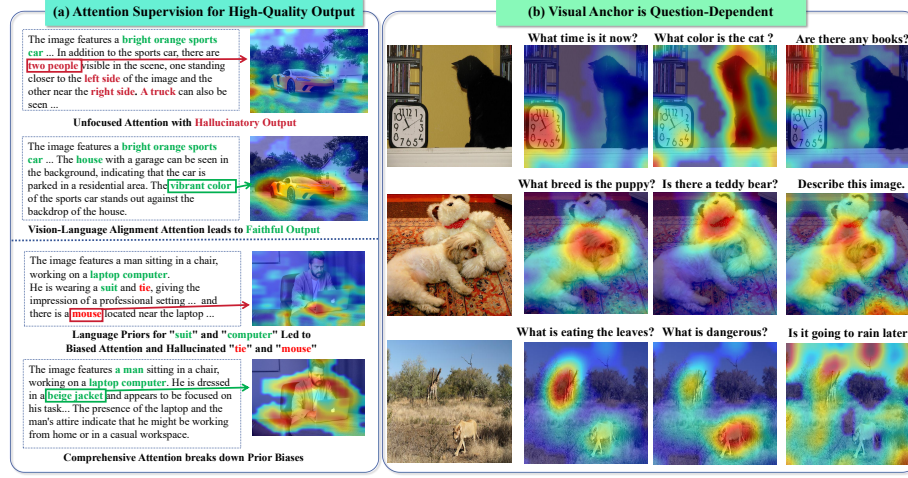
$$S_{\text{AMC}}(A_t \parallel A_{\text{anchor}}) = \frac{\sum_{i=1}^N (a_{i,t}^2 \cdot w_i^{\text{anchor}})}{\left(\sum_{i=1}^N a_{i,t}\right)^2 + \epsilon} \quad (10)$$

where  $a_{i,t}$  is the attention weight of the  $i$ -th image token at step  $t$ , and  $w_i^{\text{anchor}}$  is its anchor weight. AMC combines concentration ( $a_{i,t}^2$ ) with anchor consistency (down-weighting regions with near-zero  $w_i^{\text{anchor}}$ ), enabling online drift detection during autoregressive decoding.

We monitor attention fidelity during decoding. When  $S_{\text{AMC}} < \tau$ , we apply a steering operator  $\mathcal{O}$  to the attention logits:

$$\hat{a}_{i,t} = \begin{cases} a_{i,t} + \alpha \log(w_i^{\text{anchor}} + \epsilon), & \text{if } S_{\text{AMC}} < \tau \\ a_{i,t}, & \text{otherwise} \end{cases} \quad (11)$$

We set  $\tau = 0.6$  using a validation sweep to select the Pareto-optimal trade-off between hallucination reduction and runtime overhead (Appendix).



**Fig. 6: (a) Qualitative examples of attention-supervised inference.** We visualize how the refined cross-attention anchor enables attention correction and removes hallucination. **(b) Query-dependent cross-attention anchors.** The refined cross-attention anchor shifts to different image regions depending on the question semantics.

### 4.3 Visual Attention Guidance DPO

While Attention-Supervised Inference mitigates hallucination at inference time, we further improve the model’s intrinsic preference for visually grounded responses through VAG-DPO. Unlike standard DPO, which contrasts responses primarily at the textual level, VAG-DPO contrasts responses under different visual evidence conditions, thereby explicitly decoupling visual grounding from linguistic priors.

We construct preference pairs as follows:

- **Chosen sample** ( $I_{VE}, y_c$ ): The factual response  $y_c$  is conditioned on the visually enhanced image  $I_{VE}$ , which is derived from the Cross-Attention Visual Anchor. This strengthens the association between correct tokens and salient visual evidence.
- **Rejected sample** ( $I_{noise}, y_r$ ): The hallucinatory response  $y_r$  is paired with a noise-injected image  $I_{noise}$ , creating a “blind” condition in which the model should not assign high confidence to visually unsupported content.

Formally, we optimize the following objective:

$$\mathcal{L}_{VAG-DPO} = -\mathbb{E}_{\mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{P_{\theta}(y_c | I_{VE}, Q)}{P_{ref}(y_c | I_{VE}, Q)} - \beta \log \frac{P_{\theta}(y_r | I_{noise}, Q)}{P_{ref}(y_r | I_{noise}, Q)} \right) \right] \quad (12)$$

By explicitly distinguishing the visual evidence conditions ( $I_{VE}$  versus  $I_{noise}$ ), VAG-DPO penalizes the model’s tendency to generate high-confidence hallucinations from language priors alone, and encourages generation probabilities to remain contingent on valid visual support.



**Table 1: Main results.** Performance comparison on multiple hallucination benchmarks. ADAPT-TF denotes our training-free variant with Visual Enhance and Attention-Supervised Inference, while full ADAPT adds VAG-DPO. Full ADAPT achieves the best overall performance among compared methods across backbones.

| Methods             | Obj-Hal [22]         |                      | AMBER [32] |             |            |             | MM-Hal [27] |             | POPE-Adv [13] |
|---------------------|----------------------|----------------------|------------|-------------|------------|-------------|-------------|-------------|---------------|
|                     | Chair <sub>s</sub> ↓ | Chair <sub>i</sub> ↓ | Chair↓     | Hal↓        | Cog↓       | Acc.↑       | Score↑      | Hal↓        | Precision↑    |
| LLaVA-v1.5-7B [15]  | 54.4                 | 15.8                 | 7.8        | 36.4        | 4.2        | 71.7        | 1.86        | 0.64        | 76.1          |
| +DPO [21]           | 49.0                 | 13.0                 | 6.5        | 34.5        | 2.3        | 71.3        | 2.14        | 0.65        | 77.8          |
| +mDPO [31]          | 35.7                 | 9.8                  | 4.4        | 24.5        | 2.4        | 71.1        | 2.39        | 0.54        | 77.5          |
| +SIMA [33]          | 40.9                 | 10.4                 | -          | -           | -          | -           | 2.30        | -           | -             |
| +VCD [11]           | 51.2                 | 15.0                 | 7.4        | 34.7        | 4.1        | 71.8        | 2.12        | 0.54        | 73.4          |
| +OPERA [7]          | 50.8                 | 14.3                 | 6.9        | 30.3        | 3.2        | 75.2        | 2.15        | 0.54        | 82.2          |
| +AVISC [34]         | -                    | -                    | 6.25       | 25.6        | 2.0        | -           | -           | -           | 79.8          |
| <b>+ADAPT-TF</b>    | 35.6                 | 11.8                 | 4.0        | 16.7        | <b>1.1</b> | 73.5        | 2.55        | 0.53        | 90.3          |
| <b>+ADAPT</b>       | <b>30.7</b>          | <b>8.5</b>           | <b>3.8</b> | <b>15.2</b> | 1.2        | <b>79.6</b> | <b>2.62</b> | <b>0.53</b> | <b>91.9</b>   |
| LLaVA-v1.5-13B [15] | 49.8                 | 14.6                 | 7.0        | 29.0        | 3.2        | 71.4        | 2.38        | 0.53        | 84.5          |
| +DPO [21]           | 43.9                 | 12.4                 | 6.1        | 26.3        | 2.7        | -           | 2.47        | 0.51        | 82.7          |
| +RLHF-V [38]        | -                    | -                    | -          | -           | -          | 72.6        | 2.50        | 0.52        | -             |
| +VCD [11]           | 53.6                 | 15.3                 | 7.2        | 28.7        | 3.3        | 73.5        | 2.31        | 0.54        | 82.4          |
| +OPERA [7]          | 42.6                 | 13.2                 | 6.8        | 28.5        | 3.1        | -           | 2.46        | 0.52        | 84.7          |
| <b>+ADAPT-TF</b>    | 35.4                 | 11.6                 | <b>4.0</b> | 18.0        | 1.5        | <b>82.4</b> | 2.55        | 0.51        | 90.6          |
| <b>+ADAPT</b>       | <b>34.8</b>          | <b>10.6</b>          | 4.2        | <b>17.7</b> | <b>1.4</b> | <b>82.4</b> | <b>2.64</b> | <b>0.49</b> | <b>90.6</b>   |
| Qwen2.5-VL-3B [1]   | 43.2                 | 12.2                 | 8.0        | 44.4        | 4.3        | 86.0        | 2.62        | 0.49        | 93.2          |
| +VCD [11]           | 41.4                 | 12.3                 | 7.7        | 40.8        | 4.2        | 85.3        | 3.14        | 0.42        | 92.9          |
| +DPO [21]           | 30.6                 | 10.7                 | 5.7        | 22.6        | 1.4        | 82.8        | 2.53        | 0.52        | 93.0          |
| <b>+ADAPT-TF</b>    | 19.8                 | 8.1                  | 6.2        | 23.7        | 1.1        | <b>88.9</b> | 3.11        | 0.39        | 93.5          |
| <b>+ADAPT</b>       | <b>15.2</b>          | <b>5.3</b>           | <b>3.1</b> | <b>13.5</b> | <b>0.5</b> | 88.3        | <b>3.16</b> | <b>0.37</b> | <b>94.0</b>   |
| Qwen2.5-VL-7B [1]   | 32.3                 | 9.6                  | 4.7        | 21.9        | 1.2        | 87.6        | 3.55        | 0.4         | 93.5          |
| +VCD [11]           | 35.1                 | 10.0                 | 4.8        | 23.9        | 1.1        | 74.2        | 3.42        | 0.4         | 93.7          |
| +DPO [21]           | 29.6                 | 11.7                 | 3.9        | 17.3        | <b>0.6</b> | 85.8        | 3.59        | 0.39        | 92.4          |
| +SPIN [24]          | 28.6                 | 7.0                  | -          | -           | -          | -           | -           | -           | -             |
| <b>+ADAPT-TF</b>    | 27.8                 | 8.2                  | 5.3        | 19.5        | 0.8        | 86.3        | <b>3.83</b> | 0.31        | <b>93.9</b>   |
| <b>+ADAPT</b>       | <b>16.8</b>          | <b>5.6</b>           | <b>3.5</b> | <b>13.1</b> | 0.7        | <b>88.1</b> | 3.75        | <b>0.30</b> | 93.5          |

## 5 Experiments

### 5.1 Experimental Setup

**Benchmark:** We evaluate our method on four widely-recognized hallucination benchmarks: MM-Hal [27] for general QA; Obj-Hal [22] (using the *Chair* metric) for object-level errors; AMBER [32], providing fine-grained evaluation across generative and discriminative tasks; and POPE [13] for probing object existence.

**Compared Methods:** We compare our approach against a suite of baselines and state-of-the-art methods. These include the basic LLaVA-v1.5 [15] and Qwen2.5-VL [1] model, various inference-time decoding strategies such as VCD [11], OPERA [7], ED [3], AVISC [34], SPIN [24] and GF-SCD [40], as well as preference optimization techniques like RLHF-V [38], mDPO [31] and SIMA [33].

**Implementation Details:** We ensure a fair comparison for open-source methods by utilizing the exact same datasets and training parameters. Furthermore, for all GPT evaluations, we employed the consistent gpt-4-turbo version, with

**Table 2: Ablation and efficiency analysis.**

(a) Ablation study of ADAPT modules. Each ADAPT module improves hallucination metrics, and the full ADAPT performs best.

| Method                       | Chair↓     | Hal↓        | Cog↓       |
|------------------------------|------------|-------------|------------|
| LLaVA-v1.5-7B                | 7.8        | 36.4        | 4.2        |
| + Visual Enhance (4.1)       | 4.7        | 18.7        | 1.8        |
| + Attention-Supervised (4.2) | 6.3        | 27.8        | 2.9        |
| + VAG-DPO (4.3)              | 6.0        | 25.0        | 2.1        |
| <b>+ ADAPT</b>               | <b>3.8</b> | <b>15.2</b> | <b>1.2</b> |

(b) Inference efficiency and performance. ADAPT achieves the strongest hallucination reduction with only  $1.42\times$  runtime overhead.

| Method         | Avg. time (s)↓ | Time Ratio↓ | Chair <sub>s</sub> ↓ |
|----------------|----------------|-------------|----------------------|
| LLaVA-v1.5-7B  | 2.71           | 1.00        | 53.6                 |
| + VCD [11]     | 5.50           | 2.03        | 48.8                 |
| + OPERA [7]    | 19.51          | 7.20        | 45.1                 |
| + GF-SCD [40]  | 10.75          | 3.96        | 48.8                 |
| <b>+ ADAPT</b> | <b>3.84</b>    | <b>1.42</b> | <b>26.3</b>          |

the final results reported as the average of three evaluation runs. Training parameters are detailed in the Appendix.

**ADAPT Parameters:** We set parameters based on our analysis and empirical studies. According to the sensitivity study in Fig. 7a, we choose the fusion weights  $\omega_{\text{spec}} = 0.4$ ,  $\omega_{\text{smooth}} = 0.3$ , and  $\omega_{\text{focus}} = 0.3$  for cross-attention refinement. Based on the ablation study in the appendix and the analysis in Section 4.2, we generate the visual anchor from the first  $K = 5$  tokens and set the AMC intervention threshold to  $\tau = 0.6$ .

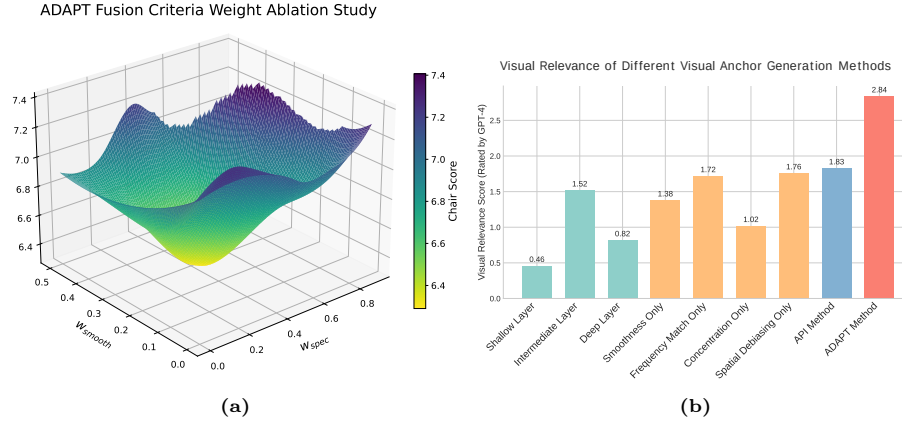
**Preference Dataset Construction:** For the VAG-DPO Training stage, we constructed our preference dataset based on the subset of the RLAIIF-V [39] and RLHF-V [38] dataset. While we directly adopted the response preference pairs from the sampled dataset, the chosen image is the visual anchor image generated by our approach and the rejected image is a noise image. This process transforms the original text-preference dataset into a vision-conditioned preference dataset tailored to our framework.

## 5.2 Main Results

As shown in Table 1, our ADAPT method achieves State-of-the-Art (SOTA) performance across all mainstream hallucination benchmarks (AMBER, POPE, MM-Hal, Obj-Hal). Our core motivation was to mitigate hallucination in long-form generation, successfully achieving SOTA on long captioning tasks. Crucially, ADAPT demonstrates architectural generalizability across diverse MLLM families, validated by strong results on both LLaVA-v1.5-7B/13B and Qwen2.5-VL-3B/7B. Specifically, ADAPT establishes superior control over object errors ( $Chair_s, Chair_i$ ) and achieves over 90% precision on POPE Adversarial. Furthermore, our lightweight, training-free variant, ADAPT-TF, surpasses most advanced methods, proving the practical potential and overall effectiveness of our multi-stage intervention framework in mitigating model hallucinations.

## 5.3 Ablation Study

**Hallucination Mitigation:** We evaluated the contribution of each high-level module on the AMBER dataset. As shown in Table 2a, we started with the



**Fig. 7: (a) Fusion Weight Sensitivity of ADAPT.** 3D surface of AMBER Chair score as a function of fusion weights  $w_{spec}$  and  $w_{smooth}$ . For Chair Score, Lower is better. **(b) Evaluation of Visual Anchor Semantic Relevance.** We compare our ADAPT anchor against ablated variants and API baseline; higher scores indicate better highlighting of query-relevant evidence, and ADAPT performs best.

LLaVA-v1.5-7B baseline and progressively added our components. The results show that each module brings a notable improvement. The full framework achieves the best performance across all metrics, highlighting a strong synergistic effect and demonstrating that all three components are crucial to the result.

**Visual Anchor Quality:** To quantitatively assess the visual relevance of our generated visual anchors, we compare our approach against several baselines: single-layer attention variants, ablated fusion criteria, and the SOTA API method [37]. Figure 7b shows our full approach significantly outperforms all baselines.

We conducted more ablation studies on both hyperparameter selection and construction of preference pairs (detailed results in Appendix).

#### 5.4 Further Analysis

**Inference Efficiency:** We evaluate the inference efficiency of our method to demonstrate its practicality. As detailed in Table 2b, our ADAPT introduces a modest time overhead, requiring only 1.42 times the inference time of the regular baseline. This is significantly more efficient than others.

**Visual Grounding Evaluation:** To provide a rigorous, objective validation of our Cross-Attention Visual Anchor quality, we assess its efficacy as a visual grounding feature. We measure selection Accuracy (Acc) and mean Intersection over Union (mIoU) on RefCOCOg Test and RefCOCO Val datasets [26] by selecting the bounding box that maximizes the contained Visual Anchor’s average attention value. As shown in Table 3a, ADAPT achieves SOTA Acc and mIoU, outperforming baselines LLaVA-v1.5-7B [15], Vanilla CLIP [20], and the SOTA

**Table 3: Further analysis of grounding and general capability.**

(a) **Evaluation of using attention for the grounding task** on RefCOCO [26]. Using our visual anchor attention for localization yields significant improvements over the LLaVA-v1.5-7B [15] and outperforms the Vanilla CLIP [20].

| Method        | RefCOCOg    |             | RefCOCO     |             |
|---------------|-------------|-------------|-------------|-------------|
|               | Acc↑        | mIoU↑       | Acc↑        | mIoU↑       |
| CLIP [20]     | 45.9        | 51.3        | 39.3        | 45.3        |
| 7B Base [15]  | 32.6        | 41.8        | 30.2        | 37.3        |
| +API [37]     | 39.8        | 56.0        | 32.5        | 39.1        |
| +Focus        | 27.9        | 37.0        | 25.9        | 34.9        |
| +Spectral     | 42.3        | 49.8        | 37.1        | 45.6        |
| +Spatial      | 37.0        | 45.1        | 32.6        | 40.5        |
| <b>+ADAPT</b> | <b>50.3</b> | <b>57.0</b> | <b>43.0</b> | <b>50.4</b> |

(b) **General capability on MME [5] and MMBench [17]**. While both DPO variants slightly reduce general performance, our VAG-DPO causes a smaller degradation than vanilla DPO, indicating better capability retention.

| Method          | MME [5]        |               | MMBench [17] |              |
|-----------------|----------------|---------------|--------------|--------------|
|                 | Score↑         | $\Delta$ ↑    | Score↑       | $\Delta$ ↑   |
| 7B Base [15]    | 1511.47        | –             | 64.60        | –            |
| +DPO            | 1443.20        | -68.27        | 63.14        | -1.46        |
| <b>+VAG-DPO</b> | <b>1459.95</b> | <b>-51.52</b> | <b>63.23</b> | <b>-1.37</b> |
| 13B Base [15]   | 1528.02        | –             | 68.47        | –            |
| +DPO            | 1465.29        | -62.73        | 66.31        | -2.16        |
| <b>+VAG-DPO</b> | <b>1473.74</b> | <b>-54.28</b> | <b>67.62</b> | <b>-0.85</b> |

API method. This confirms ADAPT successfully imparts robust grounding capabilities to a non-localizing base VLM.

**General Capability Evaluation:** Since preference optimization can inadvertently reduce a model’s general-purpose multimodal capability, we further evaluate the impact of our training stage on two widely-used general benchmarks, MME [5] and MMBench [17]. As shown in Table 3b, both vanilla DPO and our VAG-DPO cause a mild drop relative to the base model. This suggests that contrasting responses under different visual evidence conditions strengthens visual alignment while better preserving general capabilities.

**Qualitative Case Studies:** We further provide qualitative evidence to illustrate the practicality of our approach. Figure 5 visualizes the cross-attention visual anchors produced by our multi-criteria refinement, showing that the anchors consistently concentrate on query-relevant objects and suppress spurious attention on backgrounds or border regions across diverse scenes. Figure 6 further compares model outputs and text-to-image cross-attention before and after intervention: when cross-attention becomes unfocused or spatially biased, the baseline tends to introduce unsupported details, whereas Attention-Supervised Inference uses the anchor to re-steer attention toward relevant evidence during later decoding, resulting in more faithful, question-dependent generation. In the car and laptop examples, this correction prevents hallucinated entities/attributes by redirecting attention back to the visually supported regions.

## 6 Conclusion

In this paper, we study hallucinations in MLLMs through the lens of text-to-image cross-attention and identify a consistent signature: cross-attention progressively degrades during generation, shifting decoding toward language priors. Based on this analysis, we propose ADAPT, which refines multi-layer cross-attention into a query-relevant anchor and leverages it for anchor-guided atten-

tion supervision and preference tuning. Extensive experiments across architectures and benchmarks show that ADAPT effectively reduces hallucinations with a favorable efficiency–performance trade-off, highlighting cross-attention interventions as a practical path toward more faithful MLLMs.

## Acknowledgements

This work was supported by the Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200) and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM103).

## References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
3. Cho, Y., Kim, K., Hwang, T., Cho, S.: Do you keep an eye on what i ask? mitigating multimodal hallucination via attention-guided ensemble decoding. arXiv preprint arXiv:2505.17529 (2025)
4. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. *Advances in neural information processing systems* **30** (2017)
5. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: MME: A comprehensive evaluation benchmark for multimodal large language models. In: Belgrave, D., Zhang, C., Montoya, L.N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I.V.M., Loaiza-Bonilla, A. (eds.) *Advances in Neural Information Processing Systems* (2025)
6. Fu, Y., Xie, R., Sun, X., Kang, Z., Li, X.: Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In: *Findings of the Association for Computational Linguistics: ACL 2025*. pp. 16563–16577 (2025)
7. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: 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*. pp. 13418–13427 (2024)
8. Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., Zhao, P.: Self-introspective decoding: Alleviating hallucinations for large vision-language models. arXiv preprint arXiv:2408.02032 (2024)
9. Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., Yang, X.: Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 25004–25014 (2025)

10. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. arXiv preprint arXiv:2503.03321 (2025)
11. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: 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. pp. 13872–13882 (2024)
12. Li, J., Zhang, J., Jie, Z., Ma, L., Li, G.: Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. arXiv preprint arXiv:2501.01926 (2025)
13. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
14. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
15. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
17. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: ECCV (6). Lecture Notes in Computer Science, vol. 15064, pp. 216–233. Springer (2024)
18. Lyu, X., Chen, B., Gao, L., Shen, H., Song, J.: Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems* **37**, 122811–122832 (2024)
19. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in vlms. In: European Conference on Computer Vision. pp. 395–413. Springer (2024)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* **36**, 53728–53741 (2023)
22. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018)
23. Sarkar, P., Ebrahimi, S., Etemad, A., Beirami, A., Arik, S.Ö., Pfister, T.: Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025 (2025)
24. Sarkar, S., Che, Y., Gavin, A., Beerel, P.A., Kundu, S.: Mitigating hallucinations in vision-language models through image-guided head suppression. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 12481–12500 (2025)
25. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)

26. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5198–5215 (2022)
27. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 13088–13110 (2024)
28. Tang, F., Liu, C., Xu, Z., Hu, M., Huang, Z., Xue, H., Chen, Z., Peng, Z., Yang, Z., Zhou, S., et al.: Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26147–26159 (2025)
29. Tong, B., Xia, J., Zhou, K.: Mitigating hallucination in multimodal llms with layer contrastive decoding. arXiv preprint arXiv:2509.25177 (2025)
30. Wang, C., Chen, X., Zhang, N., Tian, B., Xu, H., Deng, S., Chen, H.: MLLM can see? dynamic correction decoding for hallucination mitigation. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025. OpenReview.net (2025)
31. Wang, F., Zhou, W., Huang, J.Y., Xu, N., Zhang, S., Poon, H., Chen, M.: mdpo: Conditional preference optimization for multimodal large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 8078–8088 (2024)
32. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: AMBER: an llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397 (2023)
33. Wang, X., Chen, J., Wang, Z., Zhou, Y., Zhou, Y., Yao, H., Zhou, T., Goldstein, T., Bhatia, P., Kass-Hout, T., et al.: Enhancing visual-language modality alignment in large vision language models via self-improvement. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 268–282 (2025)
34. Woo, S., Kim, D., Jang, J., Choi, Y., Kim, C.: Don’t miss the forest for the trees: Attentional vision calibration for large vision language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 1927–1951 (2025)
35. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 13258–13273 (2024)
36. Xing, Y., Li, Y., Laptev, I., Lu, S.: Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems* **37**, 92012–92035 (2024)
37. Yu, R., Yu, W., Wang, X.: Attention prompting on image for large vision-language models. In: European Conference on Computer Vision. pp. 251–268. Springer (2024)
38. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
39. Yu, T., Zhang, H., Yao, Y., Dang, Y., Chen, D., Lu, X., Cui, G., He, T., Liu, Z., Chua, T.S., et al.: Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. arXiv preprint arXiv:2405.17220 (2024)

40. Zhang, C., Wan, Z., Kan, Z., Ma, M.Q., Stepputtis, S., Ramanan, D., Salakhutdinov, R., Morency, L.P., Sycara, K., Xie, Y.: Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. arXiv preprint arXiv:2502.06130 (2025)
41. Zhao, J., Zhang, F., Sun, X., Feng, C., Tan, Z.: Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. arXiv preprint arXiv:2511.20032 (2025)
42. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023)
43. Zhu, L., Ji, D., Chen, T., Xu, P., Ye, J., Liu, J.: Ibid: Alleviating hallucinations in large vision-language models via image-biased decoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1624–1633 (2025)
44. Zou, X., Wang, Y., Yan, Y., Lyu, Y., Zheng, K., Huang, S., Chen, J., Jiang, P., Liu, J., Tang, C., et al.: Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. arXiv preprint arXiv:2410.03577 (2024)