# Clearer Sight, Fewer Lies: Oriented Pickup Preference Optimization for Multimodal Hallucination Mitigation

**Xin Zou**[1,2], **Haolin Deng**[1,2], **Yibo Yan**[1,2], **Shuliang Liu**[1,2], **Zhiwei Jin**[3], **Chen Chen**[3], **Haonan Lu**[3], **Xuming Hu**[1,2*]

[1]HKUST (GZ) [2]HKUST [3]OPPO AI Center

## Abstract

Multimodal Large Language Models (MLLMs) are prone to hallucination as their generation preferences are insufficiently calibrated to visual evidence, causing them to fall back on linguistic priors, rather than faithful grounding. In this work, we start from an empirical observation: when query-relevant visual evidence is explicitly strengthened using the model's own attention, generation becomes more accurate, suggesting that many failures do not arise solely from missing perception, but from an insufficient tendency to trust the evidence the model has already attended to. Motivated by this finding, we propose Oriented Pickup Preference Optimization (OPPO), an evidence-aware alignment objective that learns preferences over the strength of visual evidence, rather than only response quality. Concretely, OPPO contrasts the same faithful response under stronger, anchored, weaker-evidence views, turning naive visual preference into ordered visual-evidence alignment. We further combine this objective with fine-grained span-level and token-level regularization to stabilize the training. Besides, we provide a theoretical analysis showing that ordered evidence margins induce a positive lower bound on local visual sensitivity. Extensive evaluations across hallucination and general-purpose benchmarks demonstrate that OPPO consistently outperforms baseline methods.

## 1 Introduction

The rapid evolution of MLLMs [2, 1, 29, 71, 67] has fundamentally advanced document intelligence [23, 40] and complex reasoning tasks [24, 68]. Yet, their failures still share a common pattern: when visual evidence is weak, atypical, or conflicts with semantic expectations, the model often trusts its internal linguistic prior more than visual evidence, leading to what is known as multimodal hallucination [26]. Fig. 1 illustrates this shared pattern across *generalized* and *text-scene* hallucination. This deficiency poses significant risks in safety-critical [32] and text-based VQA tasks [58, 27], where any wrong in object or character recognition can completely undermine the model's reliability [6].

The root of this misalignment may stem from a fundamental limitation in current post-training paradigms [18, 57, 19]. Previous preference alignment methods typically treat the visual input as a static and passive condition while focusing on optimizing response likelihood [25, 66], thus they may improve response ranking without explicitly enforcing evidence sensitivity. Consequently, when confronted with low-quality or counter-intuitive visual signals, the decoder




Figure 1: Examples of hallucination in MLLMs.

*Corresponding author. <dylan.zoux@gmail.com>

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

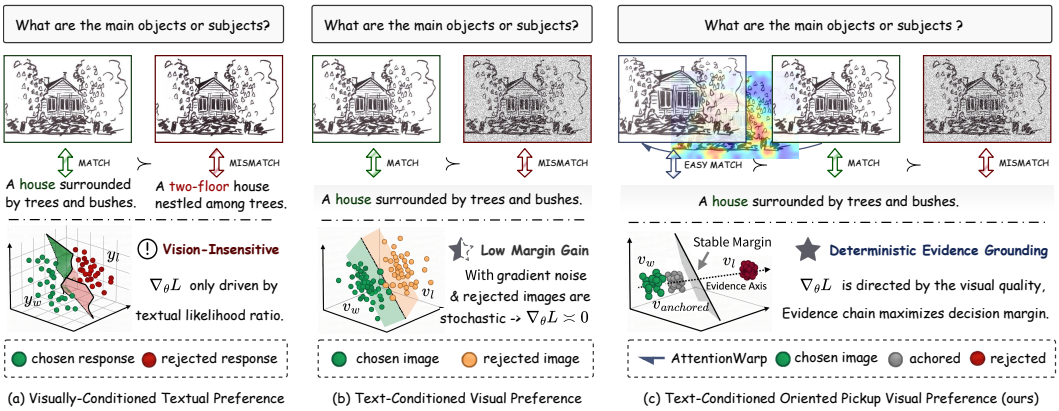



Figure 2: Comparison of preference optimization paradigms. (a) DPO focuses on linguistic alignment but is vision-insensitive. (b) Multimodal DPO utilizes stochastic visual augmentation but suffers from high gradient noise and low margin gain. (c) Our method constructs a deterministic visual preference chain to align $\nabla_\theta L$ with a query-aware evidence axis, thus maximizing the stable decision margin.

tends to "short-circuit" the evidence and retreat to its parameterized knowledge [21], *i.e.*, the model prefers to predict *"what should be there"* based presumptively on linguistic probability, rather than to decode *"what is actually there"* directly according to task-relevant visual groundedness.

This disconnect raises a pivotal question: *How can we explicitly align the model's generation preferences with the strength of its visual evidence, rather than merely ranking the final text outputs?*

Interestingly, our preliminary analysis (§ 2.1) reveals a promising paradox, *i.e.*, even when MLLMs generate hallucinatory content, their cross-modal attention maps often successfully localize the relevant visual regions. This observation suggests that the failure is not necessarily a lack of "sight", but rather a policy-level deficiency: the model does not know how to prioritize the evidence it has already attended to. To bridge this gap, we propose oriented pickup preference optimization (OPPO), an evidence-aware alignment paradigm that transforms preference learning from pure *answer ranking* into *evidence-aware ranking*. Specifically, we identify query-relevant regions through attention priors, synthesize stronger and weaker visual views, and train the model to prefer faithful responses more strongly when the supporting evidence is stronger. Compared with conventional DPO-style objectives, OPPO adds an orthogonal supervision signal, *i.e.*, it teaches the model that the confidence assigned to a faithful answer should scale with the strength of the evidence that supports it. To make this supervision more stable and localized, we further introduce span-level and token-level regularization, so that the alignment signal is concentrated on critical content and remains consistent throughout autoregressive generation. Ultimately, OPPO ensures that the model's responses are not just linguistically plausible, but are causally anchored in the visual evidence it perceives.

To further clarify the mechanics of this causal anchoring, we contrast our approach with existing preference optimization paradigms in Figure 2. Direct Preference Optimization (DPO) operates on the language manifold, but it often yields a decision boundary that is insensitive to fine-grained visual hallucinations. Existing multimodal variants of DPO attempt to bridge this gap by introducing visual perturbations; however, their reliance on random noise or generic cropping often misses the core visual evidence. This leads to a "noisy" latent space where the distinction between positive and negative samples remains ill-defined. In contrast, OPPO explicitly constructs a visual preference chain along a query-aware "evidence axis". By leveraging attention priors to warp and amplify task-relevant regions, we pull the faithful samples toward high-confidence regions of the latent space while pushing hallucinatory counterparts across the decision hyperplane. This mechanism forces the model to ground its preference scores on explicit visual features, thereby enlarging the optimization margin and effectively mitigating the gradient noise inherent in naive visual preference learning.

Overall, our contributions are summarized as follows:

- We propose to redefine MLLM hallucination as a policy-level distrust of internal visual signals rather than a mere deficiency in visual perception, an insight that allows us to pivot the alignment objective from stochastic linguistic preference to structurally-ordered evidence-aware calibration.
- We propose OPPO, which anchors the decision boundary to an ordered visual preference chain, reinforced by span-level regularization and a theoretical lower bound on local visual sensitivity.

- We demonstrate that OPPO consistently outperforms baseline methods, effectively mitigating hallucinations while preserving general capabilities across both generalized and text-scene benchmarks.

## 2 Preliminary and Motivation

**Direct Preference Optimization for MLLMs**. To align MLLMs with human intent, RLHF [3] and RLAIF [33] frameworks optimize a policy $\pi_\theta$ to maximize the KL-constrained expected reward:

$$\max_{\pi_\theta} \mathbb{E}_{(x,v)\sim\mathcal{D}, y\sim\pi_\theta(y|x,v)}[r(x,y,v)] - \beta\mathbb{D}_{\mathrm{KL}}[\pi_\theta(\cdot\,|x,v)||\pi_{\mathrm{ref}}(\cdot\,|x,v)], \tag{1}$$

where $\mathcal{D}$ denotes a preference dataset, $v$ is an image, and $\beta$ is the regularization parameter. DPO [52] bypasses explicit reward modeling by deriving a closed-form solution, the reward is reparameterized:

$$r(x,y,v) = \beta\log\frac{\pi_\theta(y|x,v)}{\pi_{\mathrm{ref}}(y|x,v)} + \beta\log Z(x,v). \tag{2}$$

Here, $Z(x,v)$ is a partition function dependent only on the inputs $(x,v)$. By integrating Bradley-Terry model [5] and preference pairs $(y_w, y_l)$, the partition function $Z$ cancels out, yielding DPO objective:

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x,v)}{\pi_{\mathrm{ref}}(y_w|x,v)} - \beta\log\frac{\pi_\theta(y_l|x,v)}{\pi_{\mathrm{ref}}(y_l|x,v)}\right)\right], \tag{3}$$

where $\sigma(\cdot)$ is `sigmoid` function, which enables direct policy optimization without explicit rewarding.

**Multimodal Direct Preference Optimization**. To align MLLMs with visual preferences, multimodal DPO extends the original framework by contrasting visual contexts $(v_w, v_l)$ instead of textual responses. Given the instruction $x$ and response $y$, the objective optimizes the policy $\pi_\theta$ to favor the preferred image $v_w$ over the dispreferred or hallucinatory image $v_l$. The loss function is defined as:

$$\mathcal{L}_{\mathrm{MDPO}} = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y|x,v_w)}{\pi_{\mathrm{ref}}(y|x,v_w)} - \beta\log\frac{\pi_\theta(y|x,v_l)}{\pi_{\mathrm{ref}}(y|x,v_l)}\right)\right]. \tag{4}$$

By maximizing the likelihood ratio of $v_w$ relative to $v_l$, which compels the model to ground its generations more accurately in the visual input, effectively suppressing the visual hallucination issue.

### 2.1 Causes of Multimodal Hallucination and Motivation

Multimodal hallucinations are often attributed to a variety of factors, including pre-training biases [87], visual uncertainty [34], attention sinks [85], and the "recency bias" in long autoregressive generation, where the model increasingly relies on previously generated text tokens over the initial visual context [11]. However, we argue that these factors culminate in a common failure mode: a misalignment between the model's internal perception (what it "sees") and its final decoding policy.

This misalignment is particularly acute in safety-sensitive or text-rich scenarios. Unlike open-ended storytelling, these tasks require visual faithfulness over semantic plausibility. For instance, when identifying a serial code or an object, the model must suppress its linguistic prior in favor of rarer but visually-grounded tokens. We categorize this failure into two types: (a) *Prior-driven overconfidence*, where the language prior dominates the output; and (b) *Evidence under-utilization*, where the model fails to translate valid visual cues into a sufficiently strong generation preference. While existing works often focus on the former by suppressing priors, our work targets the latter, *i.e.*, we want to empower the model to "***pick up***" and trust the visual evidence it has already captured in the encoding.

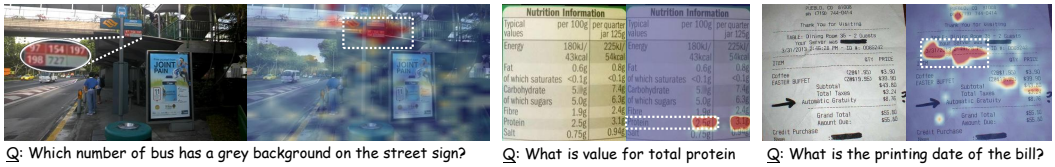



Figure 3: Cases of the visual grounding. The attention heatmaps show MLLMs maintain consistent textual-visual attention on query-relevant clues, regardless of the correctness of the final prediction.

To further validate that the bottleneck lies in *utilization* rather than *perception*, we conducted an empirical analysis on hallucinatory samples. As illustrated in Fig. 3, a striking "sight-text" paradox emerges, *i.e.*, MLLMs often still know where to look [83, 30], even when they output "the wrong answer. For example, when misidentifying a bus number 727' as a more common 720', the model's cross-modal attention heatmap often remains precisely concentrated on the correct numerical region.

Motivated by this observation, we argue that the MLLMs' hallucination is not necessarily "blindness" but a policy-level inability to prioritize attended evidence. Then, we conduct a pilot intervention study to verify whether manually amplifying these "attended but ignored" visual signals can rectify the model's decision-making. We employ two *visual-enhancement* strategies, cropping and warping, which zoom in on or spatially distort the image based on the model's own attention heatmaps to provide a "magnified" view of the evidence. As shown in Table 1, these interventions consistently improve performance, with particularly pronounced gains on hallucination, text-dense, general benchmarks. When visual evidence is stronger, the policy generation naturally shifts towards visual factuality, as [76, 8, 36] demonstrate.

Table 1: Results of visual-enhance strategies.

| Method | TextVQA | GQA | DocVQA | POPE | MMMU |
|---|---|---|---|---|---|
| LLaVA-1.5 | 49.3 | 60.5 | 18.1 | 85.3 | 36.9 |
| + API [10] | 49.9 | 60.6 | 17.4 | 85.9 | 36.9 |
| + AttCrop | 56.3 | 60.9 | 22.5 | 87.0 | 37.2 |
| + AttWarp | 58.1 | 63.7 | 25.5 | 87.5 | 40.4 |
| *vs.* baseline | (Δ8.8) | (Δ3.2) | (Δ7.4) | (Δ2.2) | (Δ3.5) |
| Qwen-VL | 81.0 | 62.4 | 77.3 | 86.1 | 47.3 |
| + API [10] | 81.6 | 61.1 | 68.4 | 85.8 | 47.4 |
| + AttCrop | 83.8 | 60.6 | 82.5 | 86.7 | 47.1 |
| + AttWarp | 84.7 | 64.0 | 84.1 | 87.4 | 50.4 |
| *vs.* baseline | (Δ3.7) | (Δ1.6) | (Δ6.8) | (Δ1.3) | (Δ3.1) |

However, relying on manual image manipulation during inference is computationally expensive and inflexible, which leads to our central hypothesis: *if "stronger evidence leads to better factuality", then the alignment objective itself should explicitly encode preferences over evidence strength, rather than merely comparing final text outputs*. By training the model to prefer responses supported by high-confidence visual cues inherently, we can bridge the gap between perception and generation without external intervention and extra overhead, which provides the direct impetus for our method.

## 3 Oriented Pickup Preference Optimization

OPPO transforms standard response-level alignment into evidence-aware preference optimization. While conventional DPO-style methods focus on which response is superior under a fixed image, we investigate how a model's confidence in a faithful response should scale with evidence strength. By explicitly modeling the relationship between visual clarity and policy preference, we provide a unified framework (Fig. 4) to mitigate hallucination where linguistic priors override visual grounding.

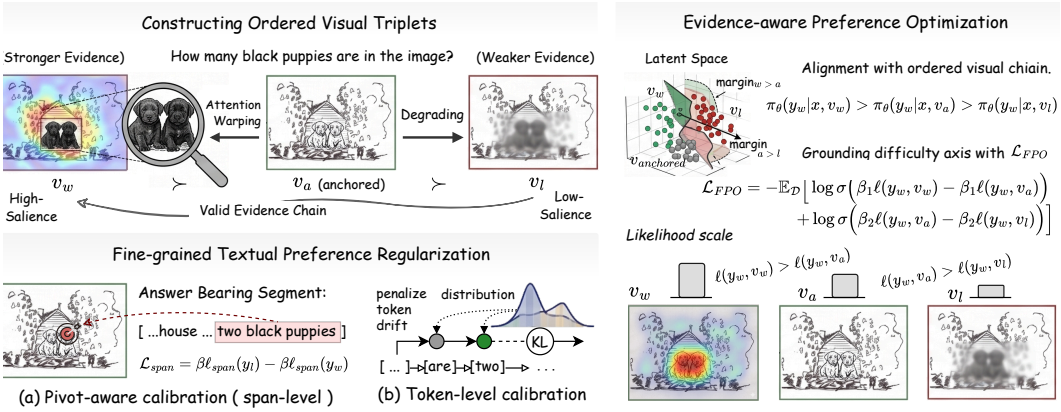



Figure 4: Overview of the OPPO framework. We establish an ordered visual triplet $v_w \succ v_a \succ v_l$ to enable evidence-aware preference alignment. By optimizing the FPO objective, the model learns to scale its confidence based on the salience of visual evidence. Further, the framework incorporates span-level and token-level calibration to penalize textual distribution drift via deterministic grounding.

### 3.1 Constructing Ordered Visual Triplets

To treat evidence strength as a controllable supervisory signal, we construct an ordered triplet of views for each query-answer pair: *stronger-evidence view* $v_w$, *anchored view* $v_a$, and *weaker-evidence view* $v_l$. These views maintain the same semantic target while systematically varying the difficulty of visual grounding. Specifically, we derive $v_w$ by amplifying query-relevant regions identified by the model's internal attention. Given the anchored image $v_a$ and query $x$, we apply a differentiable spatial transformation $\mathcal{T}$ [10] that magnifies high-salience areas while preserving global context:

$$v_w = \mathcal{T}(v_a, \texttt{cross-attention}\,(x, v_a)). \tag{5}$$

By presenting a "clarified" version of existing evidence without external annotations, $v_w$ forces the model to acknowledge strengthened visual support. $v_l$ is constructed by stochastic degradation in the anchored view $v_a$. In this way, we establish a formal evidentiary hierarchy: $v_w \succ v_a \succ v_l$, which serves as the core supervisory structure, teaching the model to positively use evidence for faithfulness.

### 3.2 Evidence-aware Preference Optimization

With the ordered visual triplet in place, OPPO extends standard DPO in two stages. First, it introduces an evidence-aware preference term that compares the *same faithful response* across different evidence conditions. Second, it adds fine-grained textual regularizers so that this evidence preference is expressed at the answer-bearing span level and throughout the token-generation trajectory.

Naive DPO optimizes whether a chosen response should outrank a rejected one under a fixed image. OPPO adds a complementary question: *should the same faithful response become more preferred when the supporting visual evidence is stronger?* We answer this with foci-guided preference optimization, which anchors the chosen response $y_w$ to the ordered visual chain introduced above:

$$\mathcal{L}_{FPO} = -\mathbb{E}_{\mathcal{D}}\big[\log\sigma\big(\beta_1\ell(y_w, v_w) - \beta_1\ell(y_w, v_a)\big) + \log\sigma\big(\beta_2\ell(y_w, v_a) - \beta_2\ell(y_w, v_l)\big)\big], \quad (6)$$

where $\ell(y, v) = \log\frac{\pi_\theta(y|x,v)}{\pi_{ref}(y|x,v)}$. Rather than only learning that $y_w$ should beat $y_l$, it learns that the *same* faithful answer should receive a larger preference margin when the visual evidence is stronger. The first term enforces $v_w \succ v_a$, encouraging the model to rely more heavily on amplified question-relevant evidence. The second term enforces $v_a \succ v_l$, discouraging the model from assigning higher confidence to answers that can be supported mainly by language priors. Together, the cooperation of the two terms makes visual evidence sensitivity part of the preference alignment objective.

**Fine-grained textual regularization.** Evidence-level preference alone is still too coarse to shape the generation reliably. In practice, hallucinations are often localized: a single entity, number, or answer span is wrong, while the rest of the response is partially correct. Moreover, even when the final answer looks plausible, token-level drift during autoregressive decoding can reveal that the model is gradually moving away from grounded evidence, and easily triggering reward hacking [12, 49]. To address these, we introduce decomposed optimization to facilitate fine-grained reward attribution:

***(a) Pivot-aware calibration.*** For critical span $y_s$, we think it should be assigned more reward. Following [80, 13], we perform a sequence-level comparison between the chosen $y_w$ and rejected $y_l$, identify the critical span $y_s$, i.e., the answer-bearing segment where the chosen and rejected trajectories diverge most strongly. Then, boost its contribution via a weighted log-probability term:

$$\mathcal{L}_{span} = -\mathbb{E}\big[\log\sigma\big(\beta\ell_{span}(y_w, v_a) - \beta\ell_{span}(y_l, v_a)\big)\big], \quad (7)$$

where $\ell_{span}(y, v_a) = \frac{1}{s}\big(\ell(y, v_a) + \sum_{y_i \in y}\ell(y_{<i}, v_a)\big)$, and $s = |y| + |y_s|$ serves as the length-normalization factor. By explicitly calibrating the pivot span, the term improves credit assignment.

***(b) Token-level calibration.*** As the autoregressive process is token by token, we think alignment on the token level is natural and makes MLLM keep diversity [82]. To further construct a fine-grained constraint, different from only response-level preference optimization (Eq. 3) in the previous methods, we enforce token-level supervision for consistency across the full generation trajectory:

$$\mathcal{L}_{token} = \mathbb{E}_{(y_w, y_l, x, v_a)\sim\mathcal{D}}\big[\beta[\tau(y_w, v_a)]_\perp - \beta\tau(y_l, v_a)\big], \quad (8)$$

where $[\cdot]_\perp$ denotes stop-gradient operation, $\tau(y, v_a = \sum_t \mathbb{D}_{\mathrm{KL}}[\pi_\theta(\cdot|x, v_a, y_{<t})\|\pi_{ref}(\cdot|x, v_a, y_{<t})]$ aggregates the sequential KL divergence along the generation path. Specifically, this token-level calibration penalizes accumulative per-token drift and keeps the aligned policy close to a stable reference trajectory during alignment training, which is especially important in complex objects or text-rich scenarios, where a locally wrong token can invalidate the entire generation.

The final `OPPO` objective establishes a comprehensive optimization framework by integrating standard response preference with evidence-oriented preference and a system of fine-grained textual regularization that constrains the model's generation space:

$$\mathcal{L}_{OPPO} = \mathcal{L}_{DPO} + \mathcal{L}_{DePO} + \gamma_1\mathcal{L}_{FPO}, \quad (9)$$

where $\mathcal{L}_{DePO} = \mathcal{L}_{span} + \gamma_2 * \mathcal{L}_{token}$, and $\gamma_1, \gamma_2$ balance the influence of evidence-aware preference learning and fine-grained textual regularization. Conceptually, standard $\mathcal{L}_{DPO}$ teaches the model *which answer* to consistently prefer, $\mathcal{L}_{FPO}$ teaches it *how that preference should scale with visual evidence strength*, and $\mathcal{L}_{DePO}$ teaches it *where and how* to express that preference during generation.

Table 2: Quantitative evaluation of hallucination performance. Hal. means hallucination rate, lower is better for ↓ marked metrics. **Bold** and underlined denote the best and the second value, respectively.

| Methods | POPE | | MMHal | | AMBER (Generative) | | | | OBJHal | | HallusionBench | | Avg. Hal. ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc↑ | Pre↑ | Score↑ | Hal.↓ | CHAIR↓ | Cover↑ | Hal.↓ | Cog↓ | $CHAIR_s$↓ | $CHAIR_i$↓ | qAcc↑ | fAcc↑ | |
| MemVR(ICML25) | 85.1 | 89.1 | 2.33 | 54.3 | 8.6 | 50.3 | 39.7 | 4.6 | 48.6 | 16.8 | 5.2 | 11.8 | 47.0 |
| HALVA (ICLR25) | 87.2 | 78.9 | 2.12 | 59.1 | 6.9 | 52.8 | 33.2 | 3.5 | 47.3 | 14.6 | 21.4 | 25.1 | 46.2 |
| RLAIFV(CVPR25) | 88.1 | 88.0 | 2.89 | 42.8 | 3.0 | 50.3 | 16.5 | 1.0 | 13.7 | 4.2 | 26.9 | 30.1 | 29.7 |
| LLaVA-1.5 | 84.9 | 89.1 | 2.18 | 59.2 | 8.8 | 50.1 | 40.4 | 4.7 | 54.7 | 26.5 | 3.9 | 11.5 | 49.8 |
| + DPO | 87.6 | 79.8 | 2.14 | 65.8 | 6.5 | 55.5 | 34.5 | 2.3 | 58.1 | 7.2 | 7.1 | 7.6 | 50.1 |
| + mDPO | 87.8 | 82.0 | 2.39 | 65.2 | 4.4 | 52.4 | 24.5 | 2.4 | 35.7 | 9.8 | 6.8 | 9.5 | 44.9 |
| + CHiP(ICLR25) | **88.1** | **92.2** | 2.32 | 56.2 | 2.9 | 57.3 | 19.9 | 1.9 | **25.3** | **6.2** | **7.3** | 9.8 | 38.1 |
| + OPPO(ours) | 87.9 | 91.9 | **2.54** | **52.0** | **2.3** | **59.3** | **14.9** | **1.8** | 28.9 | 6.5 | 7.2 | **10.1** | **33.5** |
| Qwen2.5-VL | 86.3 | 85.8 | 3.29 | 27.5 | 4.6 | 54.6 | 21.1 | 1.3 | 40.7 | 8.6 | 39.6 | 34.9 | 24.3 |
| + DPO | 86.8 | 86.2 | 3.31 | 27.0 | 4.4 | 54.9 | 20.5 | 1.3 | 38.5 | 8.2 | 39.8 | 35.1 | 23.7 |
| + mDPO | 86.6 | 85.9 | 3.30 | 27.2 | 4.5 | 54.7 | 20.8 | 1.3 | 39.2 | 8.4 | 39.7 | 35.0 | 24.0 |
| + CHiP(ICLR25) | 86.5 | 86.8 | 3.32 | 26.5 | 4.2 | 54.8 | 20.3 | 1.3 | 37.2 | 7.9 | 39.8 | 35.8 | 23.4 |
| + OPPO(ours) | **87.0** | 87.1 | **3.38** | 25.4 | **4.0** | **55.8** | 19.2 | **1.2** | **35.8** | **7.7** | **40.2** | **36.5** | **22.3** |
| LLaVA-NeXT | 85.5 | 90.3 | 3.50 | 40.6 | 8.7 | 61.1 | 49.7 | 4.2 | 11.3 | 6.6 | 4.4 | 8.9 | 45.2 |
| + DPO | 85.2 | 84.3 | 3.41 | 45.8 | 6.2 | **61.2** | 38.3 | 3.1 | 9.0 | 5.9 | **5.7** | 12.1 | 42.1 |
| + mDPO | 87.8 | 82.0 | 3.49 | 40.6 | 4.2 | 57.9 | 28.6 | **1.8** | 8.7 | 5.4 | 4.8 | 10.4 | 34.6 |
| + CHiP(ICLR25) | **88.0** | 91.8 | 3.51 | 39.6 | 4.0 | 57.6 | 27.2 | 1.9 | 5.4 | 4.2 | 5.3 | 12.7 | 33.5 |
| + OPPO(ours) | 87.9 | **92.5** | **3.54** | **39.5** | **3.9** | 58.0 | **26.4** | 1.9 | **4.3** | **3.0** | 5.6 | **13.0** | **32.8** |

# 4 Experiments

## 4.1 Experimental Settings

**Training Data.** We choose to mix the part of RLAIF-V [81] and TextVQA-train [59] datasets with total 8000 training samples as our training data, where for RLAIF-V dataset, as it has preference pairs we directly randomly sample a subset, and for the TextVQA-train data, we same first random sample a subset, then use the RLAIF-V method constructs chosen and rejected preference pairs. Accordingly, we prepare the correct responses with both descriptive (*e.g.*, `Describe the image in detail`) and non-descriptive (*e.g.*, `<Question>, Please answer in one word`) instructions.

**Datasets and Evaluation.** To rigorously assess the effectiveness of our proposed method, we conduct a comprehensive set of experiments across OBJHal (CHAIR) benchmark [53], POPE [37], MMHal-Bench [60], HallusionBench [16], AMBER [70], TextHalu-Bench [58], KIE-HVQA [20], and general benchmark MMBench[44], MMMU [22], LLaVA-Wild [2]. More details are in the Appendix C.

**Implementation Details.** We train LLaVA-1.5-7B, Qwen2.5-VL-7B, and LLaVA-Next-8B with a learning rate of 5e-7 and a batch size of 32, and take about two hours on 4 H20 GPUs per epoch, for a total of 2 epochs. All settings of baseline methods follow the default configurations from the original papers. We set warmup ratio to 0.03, and set $\beta = 0.5$, $\gamma_2 = 0.1$. The enhanced images are offline built based on the original image of the forward attentional warping process, for our 8K samples take few hours. The rejected images are built based on the original input image of the forward diffusion process at 500 steps. For fairness, all compared preference-optimization baselines are re-evaluated in the same pipeline with greedy decoding (temperature $= 0$), and the same judge model is employed whenever a benchmark requires model-based evaluation. More details are in Appendix E.

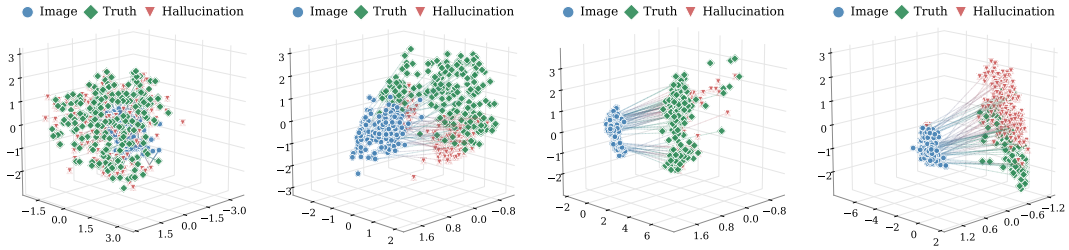
**Baselines.** We primarily compare `OPPO` with standard DPO, mDPO, and CHiP [13]. While other MLLMs cannot be directly compared due to differences in base models, preference data, and alignment methods, we provide MemVR [91], RLAIF-V [81], HALVA [54]'s results for reference, where MemVR is a representative efficient training-free paradigm, RLAIF-V, HALVA need training.

## 4.2 Experimental Results

We evaluate `OPPO` as a general-purpose, evidence-aware alignment objective across generalized hallucination, text-scene hallucination, and general benchmarks. Results are listed in Table 2, 3, 4.

**Generalized Hallucination**. Under matched-backbone comparisons, Table 2 shows that OPPO delivers consistent and competitive gains over strong preference-optimization baselines. On LLaVA-1.5 [42], Qwen2.5VL [2], LLaVA-NeXT [41], OPPO achieves the best performance, while also improving key grounding-sensitive metrics such as MMHal score, AMBER hallucination rate, and OBJHal CHAIR, which illustrate learning preferences of visual evidence strength is effective on

[2]https://huggingface.co/datasets/lmms-lab/llava-bench-in-the-wild




(a) Vanilla (LLaVA-NeXT) (b) Vanilla with mDPO (c) Vanilla with CHiP (d) Vanilla with OPPO (ours)

Figure 5: PCA visualization of representation distributions on AMBER dataset (*200 samples*). (a) Vanilla: significant overlap between image, hallucination, and truth embeddings. (b) DPO: separates truth from hallucination but fails to distinguish image features. (c) CHiP: distinguishes images but shows poor separation between truth and hallucination. (d) OPPO: yields the discriminative clusters.

generalized hallucination settings. Beyond average scores, OPPO exhibits favorable object-level grounding behavior. While methods like DPO and mDPO often trade precision for recall on OBJHal, OPPO reduces both $\text{CHAIR}_s$ and $\text{CHAIR}_i$, especially on LLaVA-NeXT, where it lowers $\text{CHAIR}_i$ to 3.0. On HallusionBench, OPPO also achieves the best fAcc among the compared alignment baselines.

**Text Scene Hallucination**. Beyond generalized visual tasks, we investigate the impact of preference optimization on fine-grained text perception using the KIE-HVQA and TextHalu benchmarks. A persistent challenge in multimodal alignment is the *"alignment tax"*, where mitigating hallucinations easily degrades the model's raw recognition capabilities. Table 3 shows that OPPO better counters this degradation than competing post-training strategies, retaining stronger accuracy and similarity on KIE-HVQA while remaining competitive on TextHalu. Particularly, text scene benchmarks are more sensitive to local mistakes, and post-training methods can easily improve refusal or caution while damaging recognition quality. OPPO is effective here because its evidence-aware objective encourages the model to capture stronger visual evidence for faithful generation, rather than simply becoming more conservative.

Table 3: The results on text hallucination benchmarks.

| Method | KIE-HVQA | | TextHalu-Bench | | TextVQA |
|---|---|---|---|---|---|
| | Acc.↑ | Sim.↑ | S-F1↑ | U-F1↑ | Acc.↑ |
| Baseline | 43.47 | 59.67 | 19.4 | 33.2 | 64.9 |
| + DPO | <u>28.97</u> | <u>42.52</u> | 20.3 | 35.0 | 65.2 |
| △(*vs.* base) | -14.50 | -17.15 | +0.9 | +1.8 | +0.3 |
| + mDPO | 27.76 | 40.57 | 21.2 | 34.3 | <u>66.2</u> |
| △(*vs.* base) | -15.71 | -19.10 | +1.8 | +1.1 | +1.3 |
| + CHiP | 28.40 | 41.70 | **21.7** | **35.6** | 65.9 |
| △(*vs.* base) | -15.07 | -17.97 | +2.3 | +2.4 | +1.0 |
| + OPPO | **31.57** | **47.39** | <u>21.4</u> | <u>34.9</u> | **66.3** |
| △(*vs.* base) | -11.90 | -12.28 | +2.0 | +1.7 | +1.4 |

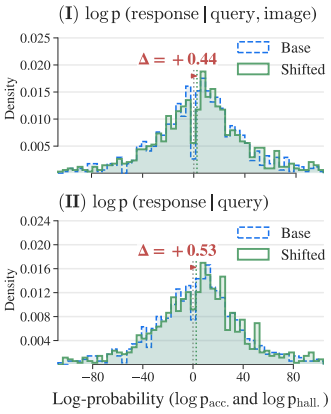
Table 4: Results on general-purpose benchmark.

| Method | MMBench | MMMU | LLaVAwild | Avg. |
|---|---|---|---|---|
| Baseline | 68.1 | 31.8 | 74.7 | 58.2 |
| + DPO | 67.2 | 30.9 | 75.5 | 57.9 |
| △(*vs.* base) | -0.9 | -0.9 | +0.8 | -0.3 |
| + mDPO | <u>67.8</u> | <u>31.5</u> | <u>75.9</u> | <u>58.4</u> |
| △(*vs.* base) | -0.3 | -0.3 | +1.2 | +0.2 |
| + CHiP | **68.3** | 31.0 | 75.8 | <u>58.4</u> |
| △(*vs.* base) | +0.2 | -0.8 | +1.1 | +0.2 |
| + OPPO | 67.5 | **32.3** | **76.1** | **58.6** |
| △(*vs.* base) | -0.6 | +0.5 | +1.4 | +0.4 |

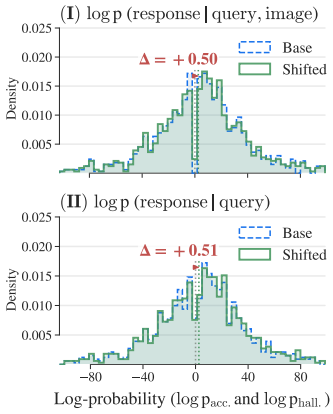
**General Capability.** To ensure that hallucination-specific tuning does not compromise the model's fundamental reasoning, we report performance on general-purpose benchmarks. As show in Table 4, `OPPO` maintains or even slightly enhances performance on these tasks compared to the baseline. For example, on MMMU, which requires complex domain knowledge, `OPPO` avoids the performance drop observed in CHiP, suggesting that our objective preserves the model's internal parameterized knowledge.

**Representation Analysis.** We visualize representation geometry via PCA [48] on AMBER data. In practice, we extract the average hidden state of the last layer for only image, image with ground-truth, image with wrong answer as inputs, then map them to the same latent space by PCA. As shown in Figure 5, the baseline LLaVA-NeXT exhibits highly entangled features, hindering discrimination between facts and fabrications. Among comparators, mDPO separates text types but suffers a modality gap, while CHiP isolates images but fails to disentangle truthful versus hallucinatory texts. In contrast, OPPO produces a clearer separation between image, truthful, and hallucinatory states, which is qualitatively consistent with the intended effect of evidence-aware alignment.

**Shifts of Distribution Gaps.** We analyze how preference optimization shifts the log-probability distributions between accurate and hallucinatory samples. As illustrated in Figure 6, both DPO and OPPO widen the distribution gap $\Delta$ in vision-conditioned scenarios (I). In the response-contrast setting, OPPO achieves a more pronounced shift of +0.50 compared to +0.44 for DPO, suggesting stronger discrimination when visual evidence is present. We also examine the textual-only conditioned case (II) to estimate residual reliance on language priors. Here OPPO yields a slightly smaller shift

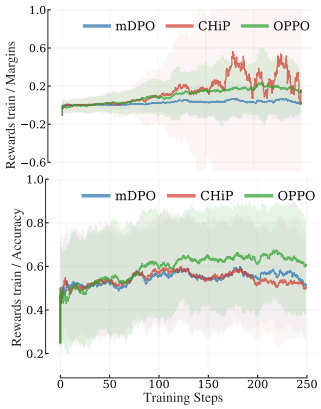



(a) Response-Contrast DPO




(b) Response-Contrast OPPO

Figure 6: Comparison of distribution-gap shifts. The x-axis denotes normalized log-probability, and density denotes the normalized empirical distribution. (I), and (II) illustrate the distributions of vision-conditioned, and textual-only-conditioned generations.




(c) Training Dynamics

Figure 7: Reward margins and reward accuracies over 250 training steps. OPPO exhibits better stability than baselines.

than DPO, +0.51 *vs.* +0.53, which is favorable as it suggests that the model is relying less on textual priors alone and more on actual visual evidence when deciding whether a response is trustworthy.

**Reward Metrics Visualization.** Beyond static representation analysis, we further investigate training dynamics. We track both metrics over 250 training steps, as illustrated in Figure 7. Here, *reward margin* denotes the implicit-reward difference assigned to the positive and negative elements in a preference pair, and *reward accuracy* denotes the fraction of pairs whose margin has the correct sign. The reward margin trajectories reveal distinct learning characteristics. mDPO follows a flat and conservative path, struggling to establish a significant margin between positive and negative pairs. While CHiP achieves higher margins, it exhibits substantial volatility, evidenced by the erratic fluctuations throughout the training process. In contrast, OPPO demonstrates a robust and steady ascent, which without the variance observed in CHiP, indicating a more stable learning signal. Besides, for the reward accuracy in Fig. 7 (bottom), we can observe that OPPO consistently outperforms both baselines, achieving higher accuracy levels early in training and maintaining this advantage until convergence, demonstrating superior effectiveness and algorithmic stability during training.

Table 5: Ablation study on hallucination-related benchmarks.

| Methods | POPE | MMHal | AMBER | OBJHal | HallusionBench | |
|---|---|---|---|---|---|---|
| | Acc↑ | Score↑ | CHAIR↓ | $CHAIR_s$↓ | qAcc↑ | fAcc↑ |
| LLaVA-NeXT | 85.5 | 3.50 | 8.7 | 11.3 | 4.4 | 8.9 |
| + OPPO | **87.9** | **3.54** | **3.9** | **4.3** | **5.6** | **13.0** |
| △ (*vs.* base) | +2.4 | +0.04 | -4.8 | -7.0 | +1.2 | +4.1 |
| w/o $L_{token}$ | 87.0 | 3.51 | 5.1 | 6.6 | 5.4 | 12.7 |
| w/o $L_{span}$ | 86.8 | 3.50 | 5.4 | 7.0 | 5.1 | 11.4 |
| w/o $L_{DePO}$ | 87.1 | 3.52 | 4.2 | 5.7 | 5.4 | 12.6 |
| w/o $L_{FPO}$ | 86.5 | 3.42 | 4.9 | 9.0 | 5.2 | 11.0 |

**Ablation Study.** We conduct a comprehensive ablation study on LLaVA-NeXT to dissect the contribution of each component in OPPO, as summarized in Table 5. Our findings reveal that the evidence-aware preference is the most critical driver for hallucination reduction. Removing $L_{FPO}$ results in the sharpest performance drop across nearly all metrics, *i.e.*, $CHAIR_s$ from 4.3 to 9.0, which confirms our hypothesis that explicitly modeling preferences over visual evidence strength is essential for guiding the model away from language priors. The fine-grained regularization framework provide vital stability during the alignment process, $L_{token}$ ensures trajectory-level consistency, the removal of $L_{span}$ significantly impacts performance, which suggests that anchoring preferences to specific answer-bearing spans is crucial for preventing the model from losing track of visual details during long-form generation.

**Hyperparameter Analysis.** We investigate the sensitivity of OPPO to its two primary hyperparameters: $\beta_1$, which controls the strength of the evidence-aware preference, and $\beta_2$, which governs the fine-grained textual regularization. As shown in Table 6, we observe that increasing $\beta_1$ from 0.01 to 0.1 significantly reduces the hallucination rate across all benchmarks, with the average score dropping from 37.90 to 31.93. This trend underscores the

Table 6: Results on different hyperparameters.

| Method | MMHal | AMBER | OBJHal | Avg. ↓ |
|---|---|---|---|---|
| | Hal. ↓ | Hal. ↓ | $CHAIR_s$↓ | |
| $\beta_1 = 0.01, \beta_2 = 0.1$ | 58.4 | 22.1 | 33.2 | 37.90 |
| $\beta_1 = 0.05, \beta_2 = 0.1$ | 54.8 | 18.5 | 30.6 | 34.63 |
| $\beta_1 = 0.1, \beta_2 = 0.1$ | 52.0 | 14.9 | 28.9 | **31.93** |
| $\beta_2 = 0.01, \beta_1 = 0.1$ | 57.3 | 25.6 | 38.4 | 40.43 |
| $\beta_2 = 0.05, \beta_1 = 0.1$ | 53.9 | 17.2 | 31.5 | 34.20 |




Figure 8: Qualitative cases showing improved visual grounding with OPPO compared with baselines.

importance of the evidence-level objective in forcing the model to prioritize visual cues. Besides, a very low $\beta_2$ leads to a noticeable performance degradation, suggesting that fine-grained textual grounding is necessary to prevent the model from drifting during preference optimization.

**Robustness to Attention Priors.** To further verify whether magnifying flawed attention priors in $v_w$ causes error amplification, we address this empirically and structurally. First, modern MLLMs possess strong baseline grounding; our analysis shows only $\sim$5.4% of attention peaks in our training set severely misalign with ground-truth bounding boxes. Second, we test an adversarial variant, OPPO†, which applies random spatial shifts to the heatmaps to forcibly magnify irrelevant regions. As shown in Table 7, while performance drops compared to standard OPPO, it avoids catastrophic collapse, which demonstrates the effectiveness of our strategy.

Table 7: Robustness to attention priors.

| Method | POPE↑ | MMHal↑ | OBJHal↓ |
|---|---|---|---|
| LLaVA-NeXT | 85.5 | 3.50 | 11.3 |
| + mDPO | 87.8 | 3.49 | 8.7 |
| + OPPO† | 86.4 | 3.48 | 7.5 |
| + OPPO (naive) | **87.9** | **3.54** | **4.3** |

As shown in Fig. 8, OPPO produces more visually grounded predictions across multiple failure modes, including scene-text recognition, object presence verification, counting, and local attribute reasoning.

## 5 Related Work

Recent efforts to mitigate MLLM hallucination include specialized fine-tuning [17], preference optimization like RLHF [80] and DPO [52], and inference-time interventions such as OPERA [28] and MemVR[91]. In text-rich scenarios, models often suffer from non-semantic hallucinations [46] despite improvements in resolution [38]. While recent works address text-specific issues through layer activation correction [58] or refusal strategies [20], they often introduce latency or domain-specific constraints. Unlike existing coarse-grained preference optimization [61, 56, 74, 69, 43], our work proposes an evidence-aware preference optimization framework that also transfers to demanding visual grounding settings. Please refer to the Appendix A for a comprehensive literature review.

## 6 Limitation

For the performance drop on KIE-HVQA, we believe it is partly due to the nature of this benchmark: its prompts are substantially longer and more instruction-heavy than those in standard OCR-style evaluation, and the benchmark is primarily designed to assess very large models with stronger long-context instruction-following ability. Under this setting, post-training a 7B-scale model for hallucination mitigation may introduce a mild instruction-following tax, which can in turn reduce performance on tasks that rely heavily on parsing long and complex prompts, even when visual grounding itself is improved. Our empirical study is still limited to a small number of model families and a fixed set of hyperparameters, so broader cross-family validation, stronger robustness analysis, and more extensive sensitivity studies remain important directions for future work.

## 7 Conclusion

We revisit multimodal hallucination through the lens of evidence utilization. Motivated by the empirical finding that strengthening query-relevant attended evidence improves grounded generation, we propose OPPO, an evidence-aware preference optimization framework that aligns response preference with ordered visual evidence strength. Across hallucination and general-purpose benchmarks, OPPO consistently outperforms strong baselines under matched settings, while our theoretical analysis suggests a positive lower bound on local visual sensitivity along the constructed evidence path, which demonstrates evidence-aware alignment as a promising direction for mitigating hallucination.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62506318); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

## References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

[5] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[6] Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6000–6010. IEEE, 2025.

[7] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, and Sungzoon Cho. Do you keep an eye on what i ask? mitigating multimodal hallucination via attention-guided ensemble decoding. *arXiv preprint arXiv:2505.17529*, 2025.

[9] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[10] Dwip Dalal, Gautam Vashishtha, Utkarsh Mishra, Jeonghwan Kim, Madhav Kanda, Hyeonjeong Ha, Svetlana Lazebnik, Heng Ji, and Unnat Jain. Constructive distortion: Improving mllms with attention-guided image warping. *arXiv preprint arXiv:2510.09741*, 2025.

[11] Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*, 2021.

[12] Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

[13] Jinlan Fu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, See-Kiong Ng, et al. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024.

[15] Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16563–16577, 2025.

[16] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[17] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

[18] Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pre-training. *arXiv preprint arXiv:2509.26625*, 2025.

[19] Yujin Han, Hao Chen, Andi Han, Zhiheng Wang, Xinyu Lin, Yingya Zhang, Shiwei Zhang, and Difan Zou. Self-contradiction as self-improvement: Mitigating the generation-understanding gap in mllms. *arXiv e-prints*, pages arXiv–2507, 2025.

[20] Zhentao He, Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, and Minghui Qiu. Seeing is believing? mitigating ocr hallucinations in multimodal large language models. *NeurIPS*, 2025.

[21] Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R Fung. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16427–16444, 2025.

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[23] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834, 2025.

[24] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264, 2024.

[25] Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, Xiaoqin Zhang, Ling Shao, Shijian Lu, and Dacheng Tao. Visual instruction tuning towards general-purpose multimodal large language model: A survey. *International Journal of Computer Vision*, 133(11):8151–8189, 2025.

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

[27] Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*, 2025.

[28] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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, 2024.

[29] Zhiwei Jin, Xiaohui Song, Nan Wang, Yafei Liu, Chao Li, Xin Li, Ruichen Wang, Zhihao Li, Qi Qi, Long Cheng, et al. Andesvl technical report: An efficient mobile-side multimodal large language model. *arXiv preprint arXiv:2510.11496*, 2025.

[30] Woojun Jung, Jaehoon Go, Mingyu Jeon, Sunjae Yoon, and Junyeong Kim. Visual funnel: Resolving contextual blindness in multimodal large language models. *arXiv preprint arXiv:2512.10362*, 2025.

[31] Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2585–2595, 2023.

[32] Younggun Kim, Ahmed S Abdelrahman, and Mohamed Abdel-Aty. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 761–771, 2025.

[33] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

[35] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 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, 2024.

[36] Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, et al. Mitigating object hallucinations in mllms via multi-frequency perturbations. *arXiv preprint arXiv: 2503.14895*, 2025.

[37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[38] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773, 2024.

[39] Zhihao Li, Haozhang Yuan, CL Philip Chen, Tong Zhang, et al. $P^2$-dpo: Grounding hallucination in perceptual processing via calibration direct preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[40] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4038–4049, 2025.

[41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2024.

[43] Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *arXiv preprint arXiv:2506.11712*, 2025.

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

[45] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024.

[46] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2(5):6, 2023.

[47] Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025.

[48] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[49] Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37:134387–134429, 2024.

[50] Dexter Neo and Tsuhan Chen. Vord: Visual ordinal calibration for mitigating object hallucinations in large vision-language models. *arXiv preprint arXiv:2412.15739*, 2024.

[51] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

[54] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv preprint arXiv:2405.18654*, 2024.

[55] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[56] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[57] Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*, 2025.

[58] Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser-Nam Lim, Harry Yang, and Nicu Sebe. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. *NeurIPS*, 2025.

[59] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.

[60] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[61] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

[62] Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Yangfan He, Kuan Lu, Hao Feng, Yang Li, et al. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*, 2024.

[63] Xiaole Tang, Xiang Gu, Xiaoyi He, Xin Hu, and Jian Sun. Degradation-aware residual-conditioned optimal transport for unified image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):1–16, 2025.

[64] Xiaole Tang, Xiaoyi He, Jiayi Xu, Xiang Gu, and Jian Sun. Learning continuous wasserstein barycenter space for generalized all-in-one image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[65] Xiaole Tang, Jiayi Xu, Xiang Gu, Yan Yang, and Jian Sun. Binding multiple modalities via multimodal wasserstein barycenter. *arXiv preprint arXiv:2609.33800*, 2026.

[66] Zhenchen Tang, Songlin Yang, Bo Peng, Zichuan Wang, and Jing Dong. Revisiting mllm based image quality assessment: Errors and remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9475–9483, 2026.

[67] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[68] Yen-Linh Vu, Dinh-Thang Duong, Truong-Binh Duong, Anh-Khoi Nguyen, Thanh-Huy Nguyen, Le Thien Phuc Nguyen, Jianhua Xing, Xingjian Li, Tianyang Wang, Ulas Bagci, et al. Describe anything model for visual question answering on text-rich images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7484–7494, 2025.

[69] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088, 2024.

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

[71] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[72] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

[73] Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551, 2025.

[74] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 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*, pages 13258–13273, 2024.

[75] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*, 2024.

[76] Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1571–1590, 2025.

[77] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10610–10620, 2025.

[78] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[79] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[80] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, 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*, pages 13807–13816, 2024.

[81] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[82] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 58348–58365, 2024.

[83] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025.

[84] Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*, 2024.

[85] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv e-prints*, pages arXiv–2406, 2024.

[86] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

[87] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[88] Xin Zou, Ruimeng Liu, Chang Tang, Zhenglai Li, Xinwang Liu, Kunlun He, and Wanqing Li. Learning disentangled representations for generalized multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[89] Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *Advances in Neural Information Processing Systems*, 38:39800–39832, 2026.

[90] Xin Zou, Chang Tang, Xiao Zheng, Zhenglai Li, Xiao He, Shan An, and Xinwang Liu. Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3550–3559, 2023.

[91] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025.

# A Related Work

**Mitigating Generalized Hallucination in MLLMs.** Extensive research has investigated the origins of hallucinations in MLLMs [79, 87, 4]. Early efforts focused on fine-grained modality alignment [53] and mitigating co-occurrence biases [31] in small-scale models. Recent strategies include specialized fine-tuning [17], RLHF [80], DPO [52], and post-hoc revisors like LURE [87], which edits potential hallucinations. Besides, there are attentional intervention methods like OPERA [28] that mitigate hallucinations without extra data or training [84, 75], yet they introduce higher inference latency. Alternatively, Contrastive Decoding (CD) methods, such as VCD [35] and ICD [72], adjust the logit distribution to suppress hallucinations [9, 7, 50]. Nevertheless, CD-based approaches can introduce potential noise into the distribution, leading to inconsistent performance improvements. More recently, MemVR [91] employs memory-space visual retracing for output verification, but such iterative refinement also comes at the cost of increased inference time. To address the computational overhead of inference-time interventions, recent advancements have pivoted towards efficient training-time alignment, particularly leveraging DPO to fundamentally reduce hallucinatory tendencies. Early attempts like HA-DPO [86] formulate hallucination mitigation as a preference ranking problem, constructing positive and negative pairs to directly penalize hallucinatory content. Building on this, V-DPO [74] argues that over-reliance on language priors is a primary cause of hallucination and introduces vision-guided preference learning to better anchor generations to the image context. Moving towards more granular control, HDPO [15] categorizes and targets specific failure modes, such as multimodal conflicts and long-context degradation, by constructing targeted preference data, while [73] utilize fine-grained AI feedback to synthesize high-quality preference datasets via a detect-then-rewrite pipeline. From a theoretical perspective, recent studies identify optimization bottlenecks in standard DPO when applied to MLLMs. OPA-DPO [77] reveals that the distribution shift from off-policy data hinders effective alignment, and proposes an on-policy strategy to bridge this gap. Similarly, SymMPO [43] introduces a symmetric preference optimization objective with margin consistency to ensure rigorous theoretical alignment, thereby achieving more robust hallucination mitigation. $P^2$DPO [39] integrates perceptual processing with calibration-based preference optimization to further ground model generations in vision and mitigate hallucination.

**OCR Hallucination and Its Mitigation.** Text-rich scenarios present unique grounding challenges, where models often succumb to "semantic hallucination", generating plausible but visually incorrect text driven by language priors rather than visual evidence [46]. Comprehensive evaluations in benchmarks like OCRBench [45] -v2[14] and TextSquare [62] reveal that standard MLLMs struggle significantly with dense text. While scaling input resolution [38, 78] improves recognition, specific hallucination issues persist. Very recently, [58] addressed this by identifying conflicts between semantic and visual cues, proposing a training-free intervention to correct layer activations, but such inference-time mechanisms inevitably increase latency. Besides, [20] tackled OCR hallucinations in degraded documents by incorporating uncertainty-aware refusal strategies. Nevertheless, their method focuses on abstention rather than correction and is primarily tailored to specific domains like invoices. Beyond existing studies, we establish a comprehensive benchmark for evaluation, and propose the first oriented pickup preference optimization algorithm to mitigate OCR hallucination.

**Preference Optimization.** The alignment of MLLMs with human intent has evolved significantly, transitioning from standard Supervised Fine-Tuning (SFT) to more robust Preference Optimization paradigms. Traditional approaches predominantly employ Reinforcement Learning from Human Feedback (RLHF) [51], which typically leverages Proximal Policy Optimization (PPO) [55] to maximize expected rewards under KL-divergence constraints. However, PPO is often plagued by training instability, hyperparameter sensitivity, and the high computational overhead of maintaining multiple models (e.g., actor, critic, and reward models). To mitigate these inefficiencies, Group Relative Policy Optimization (GRPO) [56] has emerged as a compelling alternative, estimating baselines from group scores to eliminate the dependency on an explicit value function. Parallelly, Direct Preference Optimization (DPO) [52] has revolutionized the field by deriving a closed-form solution that reparameterizes the reward function, thereby transforming the complex RL problem into a stable, offline binary classification objective. In the multimodal landscape [64, 63, 65], recent works such as RLHF-V [80] and LLaVA-RLHF [61] have successfully adapted these techniques to suppress hallucinations by penalizing non-factual responses. Nevertheless, these existing methods generally rely on coarse-grained rejection at the response or segment level. They treat the visual input as a static condition rather than an active optimization target, lacking the targeted mechanism to enforce fine-grained visual grounding [90, 89, 88].

Table 8: Feature-level comparison with closely related hallucination-mitigation and multimodal preference-optimization methods. △ denotes partial or implicit coverage. OPPO is distinguished by converting query-aware visual evidence into an ordered preference axis and coupling it with fine-grained textual calibration.

| Method | Policy-level evidence gap | Query-aware evidence | Stronger/ weaker views | Ordered evidence pref. | Visual pref. opt. | Span/token calibration | Main mechanism |
|---|---|---|---|---|---|---|---|
| DPO [52] | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Response-level preference under fixed inputs |
| mDPO [69] | △ | △ | ✗ | ✗ | ✓ | ✗ | Conditional multimodal preference over response pairs |
| V-DPO [74] | △ | △ | △ | ✗ | ✓ | ✗ | Vision-guided DPO to reduce language-prior hallucination |
| CHiP [13] | △ | △ | ✗ | ✗ | ✓ | △ | Cross-modal hierarchical preference optimization |
| SymMPO [43] | △ | ✗ | ✗ | ✗ | ✓ | ✗ | Symmetric preference margins for hallucination mitigation |
| OPERA [28] | △ | ✓ | ✗ | ✗ | ✗ | ✗ | Inference-time attention penalty and retrospection allocation |
| MLLMs Know Where to Look [83] | △ | ✓ | △ | ✗ | ✗ | ✗ | Training-free attention-guided perception of small details |
| Seeing but Not Believing [47] | ✓ | ✓ | △ | ✗ | ✗ | ✗ | Diagnostic analysis and inference-time evidence highlighting |
| **OPPO (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Ordered evidence pickup with span- and token-level calibration |

**Feature-level comparison.** To make the novelty boundary explicit, Table 8 contrasts OPPO with the most relevant diagnostic, inference-time, and preference-optimization approaches. The key difference is structural: OPPO does not merely identify attended evidence or build visual preference pairs; it orders multiple semantically consistent views by evidence strength and aligns the policy to scale its preference for the same faithful response along this evidence axis, further stabilized by span- and token-level calibration.

## B Theoretical Analysis of OPPO

We analyze the evidence-aware component of OPPO, since this is the term that directly encodes ordered visual-evidence preferences. For a fixed prompt $x$ and a faithful response $y_w$, define the log-ratio score

$$\ell_\theta(v) \;:=\; \log \frac{\pi_\theta(y_w \mid x, v)}{\pi_{\text{ref}}(y_w \mid x, v)}. \tag{10}$$

For any ordered pair of views $u \succ v$, define the pairwise evidence margin

$$\Delta_{u,v}(\theta) \;:=\; \gamma_{u,v}\big(\ell_\theta(u) - \ell_\theta(v)\big), \qquad \gamma_{u,v} > 0, \tag{11}$$

and the corresponding pairwise evidence-ranking loss

$$\mathcal{L}_{u\succ v}(\theta) \;:=\; -\log \sigma\big(\Delta_{u,v}(\theta)\big). \tag{12}$$

In OPPO, the evidence-aware objective is

$$\mathcal{L}_{\text{EV}}(\theta) = \mathcal{L}_{v_w \succ v_a}(\theta) + \mathcal{L}_{v_a \succ v_l}(\theta), \tag{13}$$

where $(v_w, v_a, v_l)$ denote the stronger-evidence, anchored, and weaker-evidence views, respectively.

**Remark.** Eq. (10) is exactly the quantity used by the evidence-aware preference term in the main method. Therefore, the analysis below directly characterizes the optimization behavior of OPPO's ordered evidence objective.

**Assumption B.1** (Semantic invariance of the view chain)**.** The three constructed views $(v_w, v_a, v_l)$ preserve the same semantic target for the query $x$, i.e., the faithful response $y_w$ remains valid across all three views, while only the effective strength of supporting visual evidence changes.

**Assumption B.2** (Weak-view prior proxy)**.** The weaker-evidence view $v_l$ substantially attenuates the query-relevant evidence, so that $\ell_\theta(v_l)$ is more dominated by language prior than by grounded visual support. This assumption is approximate and is used only to interpret the optimization direction on the weak view.

**Assumption B.3** (Smoothness)**.** For fixed $(x, y_w)$, the map $v \mapsto \ell_\theta(v)$ is continuously differentiable on every line segment joining two views in $\{v_w, v_a, v_l\}$.

**Proposition B.4** (Score-level monotonicity of the pairwise evidence loss)**.** *For any ordered pair $u \succ v$, the pairwise loss $\mathcal{L}_{u\succ v}$ satisfies*

$$\frac{\partial \mathcal{L}_{u\succ v}}{\partial \ell_\theta(u)} = -\gamma_{u,v}\big(1 - \sigma(\Delta_{u,v})\big) < 0, \qquad \frac{\partial \mathcal{L}_{u\succ v}}{\partial \ell_\theta(v)} = \gamma_{u,v}\big(1 - \sigma(\Delta_{u,v})\big) > 0. \tag{14}$$

*Hence, gradient descent on $\mathcal{L}_{u\succ v}$ pushes the preferred-view score $\ell_\theta(u)$ upward and the dispreferred-view score $\ell_\theta(v)$ downward.*

*Proof.* Let $\Delta = \Delta_{u,v}(\theta) = \gamma_{u,v}(\ell_\theta(u) - \ell_\theta(v))$. Since

$$\mathcal{L}_{u\succ v} = -\log \sigma(\Delta),$$

we have

$$\frac{d}{d\Delta}\big(-\log \sigma(\Delta)\big) = -(1 - \sigma(\Delta)).$$

Applying the chain rule gives

$$\frac{\partial \mathcal{L}_{u\succ v}}{\partial \ell_\theta(u)} = -(1 - \sigma(\Delta)) \cdot \gamma_{u,v},$$

and

$$\frac{\partial \mathcal{L}_{u\succ v}}{\partial \ell_\theta(v)} = -(1 - \sigma(\Delta)) \cdot (-\gamma_{u,v}) = \gamma_{u,v}(1 - \sigma(\Delta)).$$

Since $\gamma_{u,v} > 0$ and $0 < \sigma(\Delta) < 1$, the first derivative is strictly negative and the second is strictly positive. Therefore, gradient descent increases $\ell_\theta(u)$ and decreases $\ell_\theta(v)$. □

**Lemma B.5** (First-order margin amplification under parameter updates). *Assume $\Delta_{u,v}(\theta)$ is twice continuously differentiable with respect to $\theta$. Let*

$$\theta^+ = \theta - \eta \nabla_\theta \mathcal{L}_{u \succ v}(\theta), \qquad \eta > 0. \tag{15}$$

*Then the pairwise margin satisfies the first-order expansion*

$$\Delta_{u,v}(\theta^+) = \Delta_{u,v}(\theta) + \eta\big(1 - \sigma(\Delta_{u,v}(\theta))\big) \, \|\nabla_\theta \Delta_{u,v}(\theta)\|^2 + O(\eta^2). \tag{16}$$

*In particular, for sufficiently small $\eta$, if $\nabla_\theta \Delta_{u,v}(\theta) \neq 0$, then one gradient step strictly increases the margin $\Delta_{u,v}$.*

*Proof.* From Eq. (12),

$$\nabla_\theta \mathcal{L}_{u \succ v}(\theta) = -(1 - \sigma(\Delta_{u,v}(\theta)))\nabla_\theta \Delta_{u,v}(\theta).$$

Therefore,

$$\theta^+ - \theta = \eta\big(1 - \sigma(\Delta_{u,v}(\theta))\big)\nabla_\theta \Delta_{u,v}(\theta).$$

Applying the first-order Taylor expansion of $\Delta_{u,v}$ around $\theta$ yields

$$\Delta_{u,v}(\theta^+) = \Delta_{u,v}(\theta) + \nabla_\theta \Delta_{u,v}(\theta)^\top (\theta^+ - \theta) + O(\|\theta^+ - \theta\|^2).$$

Substituting the update expression above gives

$$\Delta_{u,v}(\theta^+) = \Delta_{u,v}(\theta) + \eta\big(1 - \sigma(\Delta_{u,v}(\theta))\big) \, \|\nabla_\theta \Delta_{u,v}(\theta)\|^2 + O(\eta^2),$$

which proves the claim. □

**Corollary B.6** (Relative suppression of prior-supported confidence on the weak view). *Consider the anchored-vs-weak pair $v_a \succ v_l$. Minimizing $\mathcal{L}_{v_a \succ v_l}$ strictly increases the gap*

$$\ell_\theta(v_a) - \ell_\theta(v_l). \tag{17}$$

*Under Assumption B.2, this comparatively suppresses confidence that can be maintained under weak, prior-dominated evidence. Moreover, if $\ell_\theta(v_a)$ is locally unchanged to first order, then $\ell_\theta(v_l)$ decreases strictly after a gradient step.*

*Proof.* The first statement follows directly from Proposition B.4 and Lemma B.5. Under Assumption B.2, the score on $v_l$ is interpreted as being increasingly dominated by language prior once grounded evidence is weakened. Therefore, enlarging the anchored-vs-weak gap suppresses prior-supported confidence *relative* to the anchored grounded score. If $\ell_\theta(v_a)$ is locally fixed, then the increase of the gap in Eq. (17) must come from a strict decrease in $\ell_\theta(v_l)$. □

**Lemma B.7** (Positive directional sensitivity along the enhancement direction). *Let*

$$\delta_+ := v_w - v_a. \tag{18}$$

*Under Assumption B.3,*

$$\ell_\theta(v_w) - \ell_\theta(v_a) = \int_0^1 \nabla_v \ell_\theta(v_a + t\delta_+)^\top \delta_+ \, dt. \tag{19}$$

*Consequently, if*

$$\ell_\theta(v_w) - \ell_\theta(v_a) \geq \varepsilon_+ > 0, \tag{20}$$

*then there exists a point $\xi_+$ on the line segment $[v_a, v_w]$ such that*

$$\nabla_v \ell_\theta(\xi_+)^\top \delta_+ = \ell_\theta(v_w) - \ell_\theta(v_a) \geq \varepsilon_+, \tag{21}$$

*and therefore*

$$\|\nabla_v \ell_\theta(\xi_+)\| \geq \frac{\varepsilon_+}{\|\delta_+\|}. \tag{22}$$

*Proof.* Define the scalar function

$$g_+(t) := \ell_\theta(v_a + t\delta_+), \qquad t \in [0,1].$$

By Assumption B.3, $g_+$ is continuously differentiable. Applying the fundamental theorem of calculus,

$$\ell_\theta(v_w) - \ell_\theta(v_a) = g_+(1) - g_+(0) = \int_0^1 g'_+(t)\,dt = \int_0^1 \nabla_v \ell_\theta(v_a + t\delta_+)^\top \delta_+ \,dt,$$

which proves Eq. (19). Since the integrand is continuous, the mean value theorem for integrals implies that there exists $t^\star \in (0,1)$ such that

$$\nabla_v \ell_\theta(v_a + t^\star \delta_+)^\top \delta_+ = \ell_\theta(v_w) - \ell_\theta(v_a).$$

Let $\xi_+ = v_a + t^\star \delta_+$. Then Eq. (21) follows from Eq. (20). Finally, by Cauchy–Schwarz,

$$\nabla_v \ell_\theta(\xi_+)^\top \delta_+ \le \|\nabla_v \ell_\theta(\xi_+)\|\ \|\delta_+\|,$$

which implies Eq. (22). □

**Proposition B.8** (Transitivity of attained evidence margins)**.** *Suppose the two OPPO evidence margins satisfy*

$$\Delta_{v_w,v_a}(\theta) \ge m_1 > 0, \qquad \Delta_{v_a,v_l}(\theta) \ge m_2 > 0. \tag{23}$$

*Then*

$$\ell_\theta(v_w) > \ell_\theta(v_a) > \ell_\theta(v_l), \tag{24}$$

*and*

$$\ell_\theta(v_w) - \ell_\theta(v_l) \ge \frac{m_1}{\gamma_{v_w,v_a}} + \frac{m_2}{\gamma_{v_a,v_l}}. \tag{25}$$

*Proof.* By the definition of the margins in Eq. (11),

$$\ell_\theta(v_w) - \ell_\theta(v_a) = \frac{\Delta_{v_w,v_a}(\theta)}{\gamma_{v_w,v_a}} \ge \frac{m_1}{\gamma_{v_w,v_a}} > 0,$$

and similarly

$$\ell_\theta(v_a) - \ell_\theta(v_l) = \frac{\Delta_{v_a,v_l}(\theta)}{\gamma_{v_a,v_l}} \ge \frac{m_2}{\gamma_{v_a,v_l}} > 0.$$

This proves Eq. (24). Adding the two inequalities gives Eq. (25). □

**Theorem B.9** (Local non-vanishing visual sensitivity along the evidence path)**.** *Under Assumption B.3, suppose Eq. (23) holds. Let*

$$\delta := v_w - v_l. \tag{26}$$

*Then there exists a point $\xi$ on the line segment $[v_l, v_w]$ such that*

$$\|\nabla_v \ell_\theta(\xi)\| \ge \frac{1}{\|v_w - v_l\|}\left(\frac{m_1}{\gamma_{v_w,v_a}} + \frac{m_2}{\gamma_{v_a,v_l}}\right) > 0. \tag{27}$$

*Equivalently, any solution that achieves positive ordered evidence margins cannot be visually insensitive everywhere along the constructed evidence path from $v_l$ to $v_w$.*

*Proof.* Define

$$g(t) := \ell_\theta(v_l + t\delta), \qquad t \in [0,1].$$

By Assumption B.3, $g$ is continuously differentiable. By the one-dimensional mean value theorem, there exists $t^\star \in (0,1)$ such that

$$g'(t^\star) = g(1) - g(0) = \ell_\theta(v_w) - \ell_\theta(v_l). \tag{28}$$

Let $\xi = v_l + t^\star \delta$. Then

$$\nabla_v \ell_\theta(\xi)^\top (v_w - v_l) = \ell_\theta(v_w) - \ell_\theta(v_l). \tag{29}$$

By Proposition B.8,

$$\ell_\theta(v_w) - \ell_\theta(v_l) \ge \frac{m_1}{\gamma_{v_w,v_a}} + \frac{m_2}{\gamma_{v_a,v_l}}.$$

Applying Cauchy–Schwarz to Eq. (29) yields

$$\|\nabla_v \ell_\theta(\xi)\|\ \|v_w - v_l\| \ge \ell_\theta(v_w) - \ell_\theta(v_l) \ge \frac{m_1}{\gamma_{v_w,v_a}} + \frac{m_2}{\gamma_{v_a,v_l}}.$$

Rearranging proves Eq. (27). □

**Interpretation.** Theorem B.9 is intentionally local: it does *not* claim that OPPO globally eliminates hallucination. Instead, it shows that once the ordered evidence margins are achieved, the optimized model must exhibit non-zero visual sensitivity at least somewhere along the constructed path from weak to strong evidence. This is exactly the behavior OPPO is designed to encourage.

# C Benchmarks and Evaluation Metrics.

**Object HalBench (ObjHal)** [53] is a widely used benchmark for evaluating object hallucination. To improve evaluation stability, the benchmark includes 8 diverse prompts and is tested on 300 instances. **Metrics**: Following [80, 69], we report both the *response-level hallucination rate* (`R.`) and *mention-level hallucination rate* (`M.`).

**MMHal-Bench** [60] is a question-answering benchmark that covers 8 question categories and 12 object topics. **Metrics:** It uses GPT-4 to assess response quality (`Ova.`) and hallucination rates (`R.`).

**HallusionBench** [16] evaluates visual illusions and knowledge hallucinations, featuring 346 images and 1129 questions. It was the GPT4-assisted evaluation. **Metrics:** Question Pair Accuracy (`qA`), Figure Accuracy (`fA`), and All Accuracy (`aA`).

**AMBER** [70] was designed to be evaluated without LLM assistance. Following previous works [69], we only consider the generative tasks. **Metrics:** (a) CHAIR [53] (`CHAIR`); (b) Object coverage of responses (`Cover`); (c) Response-level hallucination (`Hal`); (d) Human cognition hallucination (`Cog`).

**Polling based Object Probing Evaluation (POPE)** [37] is a VQA-based metric proposed to assess hallucinations in MLLMs. This metric evaluates the MLLM's response to the prompt "Is [object] in this image?" To emphasize that this is a binary VQA task, we appended the prompt with "Please answer yes or no." To select objects referenced in the question prompt, we followed three different sampling options: random, popular, and adversarial. We evaluated performance across all options.

**TextHalu-Bench** [58] is a benchmark designed to evaluate scene-text hallucination, with a particular focus on cases where semantic priors can override faithful visual recognition. It contains carefully curated samples spanning both *semantic* and *non-semantic* text cases, and covers visually challenging scenarios such as occlusions, low-contrast text, and uncommon fonts. The benchmark includes two subtasks: *Spotting*, which requires direct text extraction from the image, and *Understanding*, which evaluates whether the recognized text is correctly grounded for downstream question answering. **Metrics:** Following the official setup, we report the F1 score for the *Spotting* subtask (`S-F1`) and the F1 score for the *Understanding* subtask (`U-F1`).

**KIE-HVQA** [20] is a benchmark for evaluating OCR hallucination under degraded document conditions. It focuses on visually-grounded question answering over key information extraction (KIE) scenarios, with test samples drawn from real-world document types such as identity cards and invoices, and augmented with simulated degradations including blur, occlusion, and low contrast. The benchmark is designed to measure whether a model can distinguish reliable visual evidence from ambiguous or unreadable regions instead of over-relying on linguistic priors. **Metrics:** Following the benchmark protocol, we report answer accuracy on question-critical text regions (`Acc.`) and text similarity (`Sim.`), where the latter reflects the normalized string-level agreement between the extracted answer and the ground truth.

**MMBench** [44] is a systematically designed multiple-choice benchmark for holistic evaluation of multimodal models. It covers a broad range of perception, reasoning, and knowledge abilities, and provides both English and Chinese versions to support bilingual and apples-to-apples comparisons across models. **Metrics:** We report the overall multiple-choice accuracy (`Acc.`).

**MMMU** [22] is a college-level multimodal benchmark designed to test expert-level perception, knowledge, and reasoning. It contains about 11.5K multimodal questions collected from exams, quizzes, and textbooks, spanning 30 subjects across six core disciplines, and involving highly heterogeneous image types such as charts, diagrams, tables, maps, and scientific figures. **Metrics:** We report the overall answer accuracy (`Acc.`).

**LLaVA-Wild** (*LLaVA-Bench in-the-Wild*) is a benchmark for evaluating general-purpose multimodal chat ability in open, diverse real-world scenarios. It contains 24 images and 60 manually curated questions covering daily-life visual chat tasks, and groups them into conversation, detailed description, and complex reasoning categories. **Metrics:** We are following the official evaluation protocol.

## D More Results

**Ablation on View Generation Strategies.** To validate our design choices for constructing the ordered visual triplet $v_w \succ v_a \succ v_l$, we conduct a detailed ablation study on the generation strategies for the stronger-evidence view ($v_w$) and the weaker-evidence view ($v_l$), as detailed in Table 9.

Table 9: Ablation on visual view construction strategies. Experiments are conducted on LLaVA-NeXT. Our default choices (Warping + Stochastic Degradation) achieve the best trade-off.

| Construction Strategies | | POPE↑ | MMHal↑ | OBJHal↓ |
|---|---|---|---|---|
| Stronger View ($v_w$) | Weaker View ($v_l$) | | | |
| *Baseline (LLaVA-NeXT w/o OPPO)* | | 85.5 | 3.50 | 11.3 |
| Cropping | Stochastic Deg. | 86.8 | 3.49 | 5.8 |
| Red Circle (Prompting) | Stochastic Deg. | 86.5 | 3.46 | 6.5 |
| Warping (Ours) | Gaussian Blur | 87.2 | 3.51 | 5.2 |
| Warping (Ours) | Cutout (Masking) | 86.3 | 3.44 | 6.1 |
| Warping (Ours) | Stochastic Deg. (Ours) | **87.9** | **3.54** | **4.3** |

For the **stronger-evidence view** ($v_w$), we compare our attention-guided *Warping* with two alternatives: *Cropping* (extracting the bounding box of the attended region) and *Visual Prompting* (drawing a red bounding box or circle around the target). While all methods improve over the baseline, Warping is markedly superior. *Cropping* forcibly discards the global visual context, which is often necessary for resolving complex queries or spatial relationships. *Visual Prompting*, on the other hand, introduces artificial pixel artifacts that shift the image distribution away from natural scenes, making the alignment less generalizable. Warping elegantly avoids both issues by smoothly magnifying local salience while retaining peripheral context.

For the **weaker-evidence view** ($v_l$), we compare our *Stochastic Degradation* against uniform *Gaussian Blur* and hard *Cutout* (masking the attended region with black pixels). *Cutout* performs poorly because it completely removes the visual evidence. When the evidence is entirely absent, penalizing the model for utilizing language priors becomes an ill-posed objective, as guessing is its only remaining option. This breaks the continuous spectrum assumption of our preference chain. *Gaussian Blur* applies a uniform low-pass filter, which preserves too much coarse semantic structure, making it insufficiently "weak" for OCR-heavy or fine-grained tasks. Our *Stochastic Degradation* strikes the optimal balance by progressively corrupting the fine-grained query-relevant cues while preserving the overall image statistics, providing a smoother and more robust gradient for preference learning.

**Attention Necessity under Localization Quality.** A remaining question is whether OPPO truly benefits from meaningful query-aware localization, or whether its gains mainly come from generic visual perturbation. Since our stronger-evidence view $v_w$ is constructed by amplifying query-relevant regions identified from cross-attention, a natural test is to ask whether OPPO becomes more effective when the underlying attention prior more accurately covers the answer-bearing region. To this end, we conduct a localization-quality study on a held-out ROI-annotated evaluation set. For each sample, we compute the ROI coverage of the base model's attention map, defined as the fraction of normalized attention mass falling inside the ground-truth answer-bearing region, and partition the samples into three bins of equal size: Low, Mid, and High coverage. We then compare the downstream performance of DPO, OPPO†, and OPPO under the same matched-backbone setting. If OPPO mainly works because of query-aware evidence pickup, its advantage over DPO and random-warp should increase with localization quality, rather than remain constant across bins. This analysis complements the robustness result in Table 10, instead of only testing whether OPPO collapses under noisy priors, we ask whether better localization systematically amplifies the gains of evidence-aware preference optimization. We observe a consistent monotonic trend: the gain of OPPO over DPO increases from +1.1 in the low-coverage bin to +2.8 in the mid-coverage bin and +4.2 in the high-coverage bin. In contrast, the random-warp variant remains consistently weaker, especially when localization quality is high. This indicates that OPPO does not simply benefit from generic visual perturbation; rather, it becomes increasingly effective when the attention prior better captures the answer-bearing evidence. Notably, OPPO still improves over DPO even in the low-coverage regime, suggesting that localization quality is not a strict prerequisite for usefulness, but a key factor governing the size of the improvement.

Table 10: Localization quality vs. OPPO gains. Samples are partitioned into three bins according to the base model's attention coverage over the answer-bearing ROI. We report POPE accuracy under the matched-backbone LLaVA-NeXT setting. As localization quality improves, the gain of OPPO over both DPO and random-warp increases monotonically, indicating that OPPO benefits specifically from better query-aware evidence localization rather than from generic visual perturbation alone.

| Coverage bin | Coverage | Base Acc. | DPO Acc. | OPPO† Acc. | OPPO Acc. | Δ(vs. DPO) | Δ(vs. OPPO†) |
|---|---|---|---|---|---|---|---|
| Low | 0.18 | 84.6 | 83.8 | 84.5 | 84.9 | +1.1 | +0.4 |
| Mid | 0.47 | 85.4 | 85.0 | 86.3 | 87.8 | +2.8 | +1.5 |
| High | 0.79 | 86.5 | 86.8 | 88.4 | 91.0 | +4.2 | +2.6 |
| Overall | 0.48 | 85.5 | 85.2 | 86.4 | 87.9 | +2.7 | +1.5 |

**Error Analysis.** Following [10], we analyzed 150 randomly sampled VQA tasks from the GQA and TextVQA evaluation sets, identifying 60 errors for the baseline LLaVA model and 39 for OPPO. These errors were classified into six failure modes: (1) Fine-Grained Details (visually minute targets), (2) Hallucination (fabricating non-existent details), (3) Misaligned Attention (focusing on incorrect objects), (4) Size (object scale misjudgments), (5) Semantically Correct (accurate but paraphrased answers), and (6) Compositional Reasoning (complex multi-object relationships). Tab. 11 demonstrates that OPPO yields notably fewer errors in fine-grained and compositional tasks. Nevertheless, warping can sometimes obscure the peripheral context necessary for global reasoning, and noisy underlying attention may lead to performance degradation. Finally, because warping preserves relative proportions despite altering absolute sizes, errors in size-related tasks are effectively limited.

Table 11: Comparison of Error Counts between Base MLLM and AttWarp across Different Categories (# Error ↓)

| **Category** | **Base** | **OPPO** |
|---|---|---|
| Fine-Grained Details | 12 | 3 |
| Hallucination | 11 | 7 |
| Misaligned Attention | 4 | 3 |
| Size | 18 | 17 |
| Semantically Correct | 8 | 6 |
| Compositional Reasoning | 7 | 4 |

# E Implement Details

## E.1 Attention-Guided Stronger-View Construction: AttCrop and AttWarp

Given an image-question pair $(v_a, x)$, we first extract a query-conditioned attention score matrix

$$A \in \mathbb{R}_{\geq 0}^{H \times W}, \tag{30}$$

where $A_{ij}$ measures the relevance of spatial location $(i, j)$ to the query. In practice, $A$ is obtained by aggregating cross-modal attention weights from the base MLLM and then upsampling them to the image resolution.[3]

For numerical stability, we normalize $A$ into a probability map:

$$\widetilde{A}_{ij} = \frac{A_{ij}}{\sum_{u=1}^{H} \sum_{v=1}^{W} A_{uv} + \varepsilon}, \qquad \sum_{i=1}^{H} \sum_{j=1}^{W} \widetilde{A}_{ij} = 1, \tag{31}$$

where $\varepsilon > 0$ is a small constant.

**Marginal attention profiles.** We further compute the row- and column-wise marginal attention distributions:

$$p_y(i) = \sum_{j=1}^{W} \widetilde{A}_{ij}, \qquad p_x(j) = \sum_{i=1}^{H} \widetilde{A}_{ij}, \tag{32}$$

where $\sum_{i=1}^{H} p_y(i) = 1$ and $\sum_{j=1}^{W} p_x(j) = 1$.

[3] For AttWarp, this follows the attention-guided image warping pipeline of Dalal et al. [10].

### E.1.1 AttCrop

AttCrop constructs a stronger-evidence view by cropping the attended region and resizing it back to the original input resolution. To obtain a deterministic attention-guided crop, we first form the cumulative distributions

$$F_y(i) = \sum_{u=1}^{i} p_y(u), \qquad F_x(j) = \sum_{v=1}^{j} p_x(v). \tag{33}$$

Given a mass-retention hyperparameter $\rho \in (0, 1)$, we define the crop box as the smallest axis-aligned rectangle that preserves the central $\rho$ mass along both axes:

$$i_{\min} = \min\left\{ i : F_y(i) \geq \frac{1-\rho}{2} \right\}, \qquad i_{\max} = \max\left\{ i : F_y(i) \leq 1 - \frac{1-\rho}{2} \right\}, \tag{34}$$

$$j_{\min} = \min\left\{ j : F_x(j) \geq \frac{1-\rho}{2} \right\}, \qquad j_{\max} = \max\left\{ j : F_x(j) \leq 1 - \frac{1-\rho}{2} \right\}. \tag{35}$$

The cropped image is then

$$v_{\text{crop}} = v_a[i_{\min} : i_{\max},\, j_{\min} : j_{\max}], \tag{36}$$

and the final AttCrop view is obtained by resizing the crop back to the input resolution:

$$v_w^{\text{crop}} = \text{Resize}(v_{\text{crop}},\, H,\, W)\,. \tag{37}$$

In words, AttCrop enlarges the most attended region by discarding peripheral content and reallocating all available pixels to the retained box.

### E.1.2 AttWarp

Unlike AttCrop, AttWarp preserves the full image content and redistributes spatial resolution according to the query-conditioned attention map. Following Dalal et al. [10], we perform *rectilinear* warping based on the row and column marginals in Eq. (32).

We first convert the marginals into cumulative distribution functions:

$$C_y(i) = \sum_{u=1}^{i} p_y(u), \qquad C_x(j) = \sum_{v=1}^{j} p_x(v). \tag{38}$$

Since $C_y$ and $C_x$ are monotone non-decreasing, they define valid 1D transport maps. We denote their inverse CDFs by

$$T_y = C_y^{-1}, \qquad T_x = C_x^{-1}. \tag{39}$$

To express the warp continuously, let $(u, v) \in [0, 1]^2$ be normalized coordinates on the target warped image. The source coordinates are defined as

$$\phi_y(u) = T_y(u), \qquad \phi_x(v) = T_x(v), \tag{40}$$

where $T_y(u)$ and $T_x(v)$ are interpreted via linear interpolation between discrete grid locations. The warped image is then produced by bilinear sampling:

$$v_w^{\text{warp}}(u, v, c) = v_a(\phi_y(u),\, \phi_x(v),\, c)\,, \qquad c \in \{1, 2, 3\}. \tag{41}$$

Intuitively, regions with larger marginal attention mass occupy a larger span in the warped image, while low-attention regions are compressed. Because the mapping is rectilinear and defined separately along the vertical and horizontal axes, AttWarp preserves the regular grid structure required by standard vision encoders while keeping all original image content.

Both AttCrop and AttWarp use the same query-conditioned attention prior $A$ but differ in how they allocate visual budget. AttCrop strengthens evidence by *discarding* low-attention context and zooming into the attended box, whereas AttWarp strengthens evidence by *redistributing* pixel density toward attended regions while preserving the full scene. In our experiments, AttWarp generally yields a better trade-off as it magnifies local evidence without removing potentially useful global context.

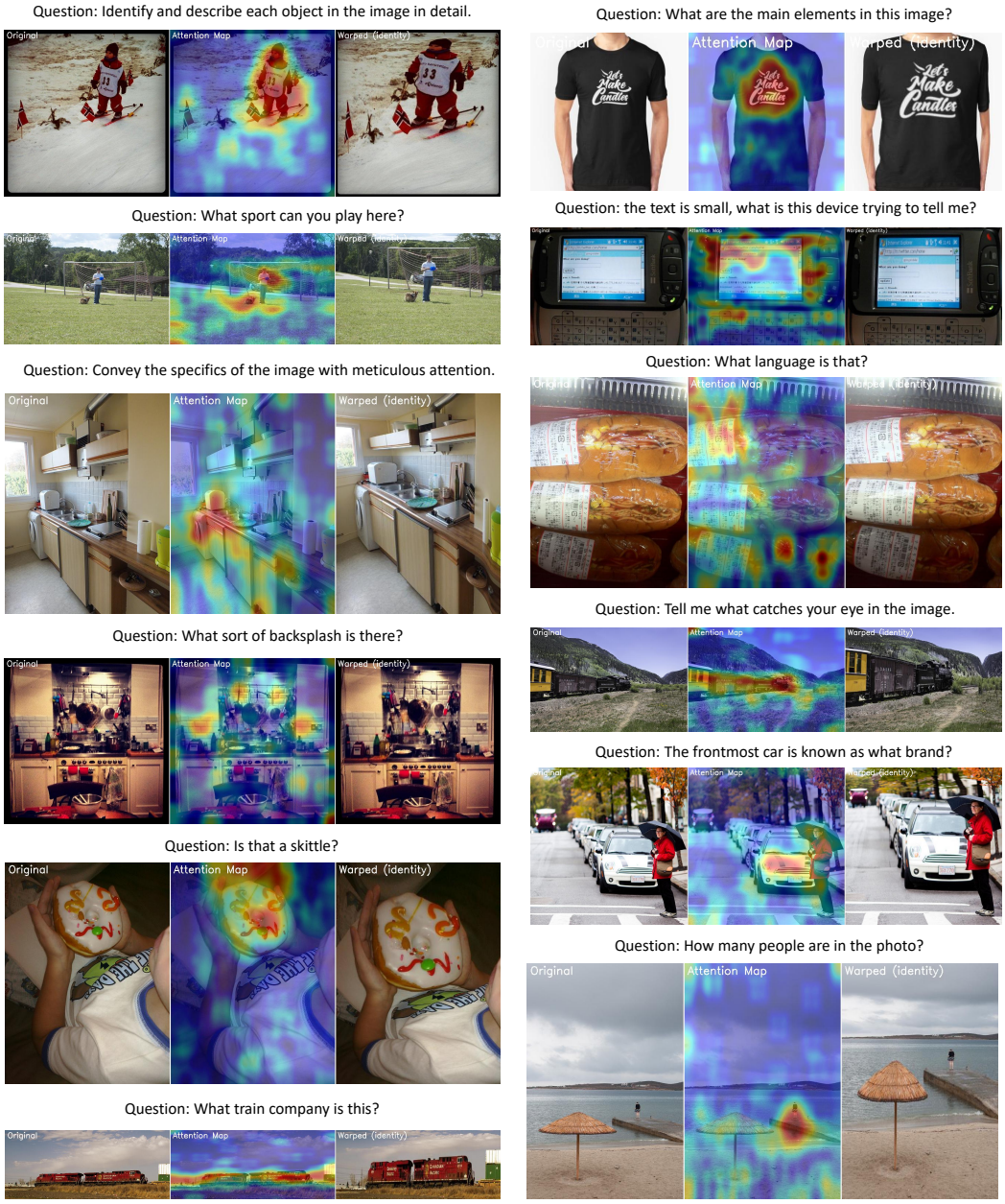



Figure 9: Qualitative results of AttWarp. We visualize the attention maps and the corresponding warped images across various VQA tasks. AttWarp effectively aligns the model's visual focus with task-relevant regions (*e.g.*, the Skittle on the donut or the train logo).

## F Impact Statements

This work improves the visual groundedness of multimodal large language models and may benefit applications such as document understanding, visual question answering, and other settings where reducing hallucination is important for reliability. By aligning model preference with visual evidence strength, our method can help make model outputs more faithful to the input image.

However, stronger fine-grained visual grounding and OCR ability may also introduce dual-use risks, such as sensitive text extraction or surveillance-related misuse. Moreover, reduced hallucination does not imply error-free behavior, and deployment in high-stakes settings should therefore remain cautious. We encourage responsible use with appropriate safeguards, domain-specific evaluation, and clear restrictions on misuse.