Staying VIGILant: Mitigating Visual Laziness via Counterfactual Visual Alignment in MLLMs

Xi Xiao[1*], Chen Liu[2*], Chih-Ting Liao[3], Yunbei Zhang[4], Qizhen Lan[5], Yuxiang Wei[6], Lin Zhao[7], Janet Wang[4], Jianyang Gu[8†], Muchao Ye[9], Tianyang Wang[1†], and Hao Xu[10]

[1] University of Alabama at Birmingham, Birmingham, AL, USA
[2] Yale University, New Haven, CT, USA
[3] University of New South Wales, Sydney, NSW, Australia
[4] Tulane University, New Orleans, LA, USA
[5] The University of Texas Health Science Center, Houston, TX, USA
[6] Georgia Institute of Technology, Atlanta, GA, USA
[7] Northeastern University, Boston, MA, USA
[8] The Ohio State University, Columbus, OH, USA
[9] University of Iowa, Iowa City, IA, USA
[10] Harvard University, Cambridge, MA, USA

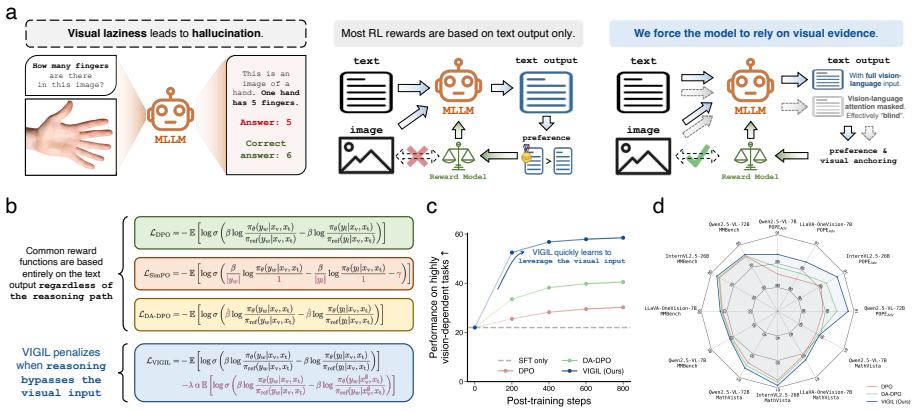

Fig. 1: Overview of VIGIL for visually grounded MLLMs. (a) A longstanding limitation of MLLMs is *visual laziness*, namely the heavy reliance on textual knowledge priors over visual evidence that can lead to hallucination. As a remedy, we introduce VIGIL, a post-training RL framework that promotes dependence on visual evidence. **(b)** Rather than altering text-based rewards, VIGIL penalizes insufficient reliance on visual information. **(c)** Qwen2.5-VL trained with VIGIL rapidly learns to leverage visual inputs during generation. **(d)** VIGIL consistently improves hallucination mitigation and multimodal reasoning across model scales, architectures, and benchmarks.

Abstract. Multimodal large language models (MLLMs) extend large language models (LLMs) with visual perception, enabling joint reasoning

*Equal Contribution. †Corresponding authors.

over images and text. Despite inheriting strong reasoning capabilities from LLMs, they remain prone to hallucinations that contradict their visual inputs. Mechanistic studies indicate that this weakness stems from *visual laziness*: MLLMs encode the correct visual evidence internally, but overly rely on strong language priors during response. Existing alignment methods, such as direct preference optimization, primarily optimize outcome-level rewards based on text. This introduces an optimization bias toward linguistic shortcuts, leading to responses that often contradict the visual evidence. To address this, we propose Visual Information Gain In aLignment (`VIGIL`), a reinforcement-learning (RL) post-training framework that shifts the focus from numerical reward fitting to causal visual grounding. `VIGIL` introduces a geometric constraint that explicitly maximizes the mutual information between the visual input and the generated response. We achieve this by penalizing "blind confidence" instances where the model remains improperly certain even when textual-visual attention is masked to create a counterfactual blind state. Extensive experiments show that `VIGIL` consistently outperforms recent alignment methods across hallucination and reasoning benchmarks without compromising text-only capabilities. Our approach matches the full-data performance of state-of-the-art methods using only 25% of the preference data and even demonstrates emergent spatial grounding capabilities without explicit bounding box supervision.

Keywords: Multimodal Alignment · Hallucination Mitigation

1 Introduction

Multimodal large language models (MLLMs) are transitioning from basic perception tools to sophisticated reasoning engines capable of interpreting complex visual data [58]. Modern architectures, such as Qwen2-VL [53], LLaVA-OneVision [22] and InternVL2.5 [6], can conduct complex reasoning, including solving mathematical problems and analyzing financial charts. However, *hallucination* remains a persistent barrier to their reliability. In the multimodal context, hallucination is not simply a perceptual error, but rather represents a failure of visual grounding where models generate descriptions that lack support from the physical visual input.

The root cause of this unreliability is often *visual laziness* [59]. Modern MLLMs typically combine a visual encoder with a pre-trained, text-centric LLM. Because the LLM component is trained on vast text corpora, it harbors strong *language priors*. When evaluating complex multimodal inputs, the model's reasoning often defaults to these inherent priors rather than relying on visual evidence. Recent studies [2, 21, 37, 59, 66] reveal a striking disparity: MLLMs often contain the correct visual features in their latent space, but still generate incorrect answers because linguistic correlations hijack the decoding process.

Existing approaches to mitigate this issue have clear trade-offs. One direction relies on architectural modifications, such as the introduction of external tools or modular designs [19] to separate reasoning from perception, which complicates

end-to-end learning. Another direction uses online reinforcement learning for dynamic visual token selection [32], introducing significant computational overhead and training instability. Meanwhile, standard outcome-based alignment methods, such as direct preference optimization (DPO) [47] and its difficulty-aware variants [46], optimize the policy by minimizing a preference loss on the final text output. They penalize incorrect tokens but do not constrain how the model derives its answer. If MLLMs arrive at the correct answer through language priors rather than visual evidence, standard DPO still rewards the output, inadvertently reinforcing these optimization shortcuts and preserving visual laziness.

Taken together, we argue that building robust multimodal alignment and reducing visual laziness requires explicitly enforcing reliance on *visual evidence* rather than *language priors*. Our core intuition is simple: to teach a model to use visual information, we must expose the consequences of being blind. Motivated by this insight, we introduce Visual Information Gain In aLignment (`VIGIL`), a post-training framework that maximizes the Visual Information Gain (VIG) between the visual input and the generated response.

Specifically, during optimization, we construct a counterfactual blind state ($x_{\mathrm{v}}^{\emptyset}$) by masking visual attention, preventing the model from accessing visual evidence. The policy is then optimized to maximize the divergence between the standard seeing state and the introduced counterfactual blind state. This divergence acts as a geometric constraint on the model's response distribution, penalizing blind confidence cases where the model remains confident even without visual input. As a result, `VIGIL` separates visually grounded reasoning from predictions driven purely by language priors, ensuring that high-confidence outputs are causally anchored to visual features (x_{v}).

Our contributions are summarized as follows:

1. **A counterfactual alignment framework:** We propose `VIGIL` to mitigate multimodal hallucinations by incorporating counterfactual visual decoupling into the preference optimization loop and explicitly maximizing the Visual Information Gain.
2. **Data and computational efficiency:** Operating entirely offline, `VIGIL` is stable and computationally lightweight. It matches the performance of competitive baselines using only 25% of their preference data.
3. **Consistent gains across scales and architectures:** Across diverse architectures ranging from 7B to 72B parameters, `VIGIL` consistently improves performance, with gains increasing for larger and stronger foundation models.
4. **Emergent spatial grounding:** Without any explicit bounding box supervision, `VIGIL` improves zero-shot referring expression comprehension, indicating that spatial grounding can emerge without explicit localization supervision.

2 Preliminaries

In this section, we provide an extensive theoretical analysis of visual hallucination in MLLMs, which originates from a phenomenon we term *visual laziness*.

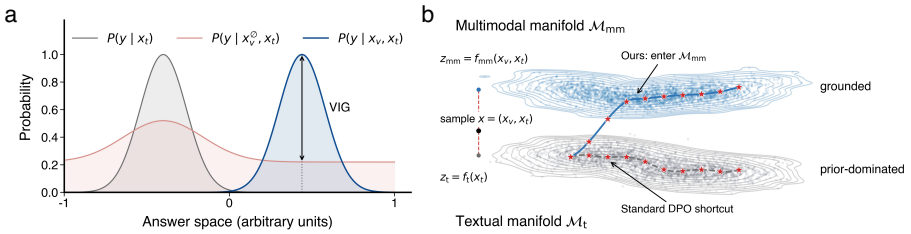

Fig. 2: Visual laziness in MLLMs. (a) Degenerate posterior and vanishing visual gain. When predictions are dominated by language priors $P(y|x_{\mathrm{t}})$, the multimodal posterior $P(y|x_{\mathrm{v}}, x_{\mathrm{t}})$ collapses toward the vision-blind prior $P(y|x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})$, yielding a diminished visual information gain (VIG). **(b)** Optimization geometry of visual shortcuts. A geometric view illustrates an optimization shortcut that stays on the textual manifold $\mathcal{M}_{\mathrm{t}}$ and converges to a prior-dominated region, instead of entering the multimodal manifold $\mathcal{M}_{\mathrm{mm}}$ for grounded reasoning, motivating our `VIGIL` framework. Refer to [26,30,31,39,49] for the definition of manifold.

2.1 Definition of Visual Laziness

Modern MLLMs often combine a visual encoder with a pre-trained LLM. During reasoning, although the visual encoder captures rich spatial features, an MLLM often overly relies on the inherent *language priors* learned during massive text-only pre-training [59]. This phenomenon, which we term *visual laziness*, causes models to favor outputs using exclusively textual information despite having access to relevant visual clues.

2.2 Information-Theoretic Analysis of Visual Laziness

Give a multimodal preference dataset $\mathcal{D} = \{(x_{\mathrm{v}}, x_{\mathrm{t}}, y_w, y_l)\}$, where x_{v} is the visual input, x_{t} is the textual instruction, and (y_w, y_l) are the preferred and disfavored responses (w and l respectively stands for "winning" and "losing"), standard Direct Preference Optimization (DPO) [47] optimizes the policy π_θ by minimizing the negative log-likelihood of the preference:

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x_{\mathrm{v}},x_{\mathrm{t}})}{\pi_{\mathrm{ref}}(y_w|x_{\mathrm{v}},x_{\mathrm{t}})} - \beta \log \frac{\pi_\theta(y_l|x_{\mathrm{v}},x_{\mathrm{t}})}{\pi_{\mathrm{ref}}(y_l|x_{\mathrm{v}},x_{\mathrm{t}})}\right)\right] \quad (1)$$

where π_{ref} serves as the reference policy to prevent excessive drift. We argue that the objective in Eq. 1 allows for an *optimization shortcut* that exacerbates visual laziness. To understand this, we analyze the model's behavior through the lens of conditional probability and information redundancy.

Language Prior Dominance. A hallucination occurs when the model's reasoning is dominated by the unimodal prior $P(y|x_{\mathrm{t}})$ rather than the multimodal posterior $P(y|x_{\mathrm{v}}, x_{\mathrm{t}})$. If a preferred response y_w can be inferred from the text alone, such as in instances of strong linguistic correlation, the model can minimize the loss by simply reinforcing its *language priors* without truly grounding the answer in the visual content. Mathematically, this manifests as an approximation:

$$\pi_\theta(y|x_{\mathrm{v}}, x_{\mathrm{t}}) \approx \pi_\theta(y|x_{\mathrm{t}}) \quad (2)$$

In this regime, the visual input x_v becomes statistically redundant in the optimization trajectory. As illustrated in Fig. 2, visual laziness manifests itself as both a degenerate posterior with vanishing VIG (Fig. 2a, see formal definition in Section 3.1) and an optimization shortcut that stays in the textual manifold (Fig. 2b).

Perception-Reasoning Decoupling. When visual evidence x_v is treated as redundant, the model's reasoning foundation collapses. For example, a model might predict that an apple is red based on common textual associations even when the image depicts a green apple, illustrating how a strong *language prior* could override the visual signal, resulting in a decoupling between perception and reasoning. We argue that this failure mode reflects a fundamental limitation of existing multimodal alignment methods. While prior approaches have improved performance through sample reweighting [46] or increased architectural complexity [19], they do not explicitly encourage the model to rely on visual evidence when forming its predictions. This observation motivates our proposed `VIGIL` framework. We posit that robust grounding requires maximizing the information contributed specifically by x_v, thereby decoupling the multimodal manifold $\mathcal{M}_{mm}$ from the pure textual manifold $\mathcal{M}_t$.

2.3 Outcome Rewards vs. Dependency Alignment

The recent success of group relative policy optimization (GRPO) [11] shows the effectiveness of outcome-based reinforcement learning for scaling complex reasoning, provided that reliable verifiers are available. GRPO primarily trains models through outcome-level rewards computed by rule-based or model-based verifiers [11]. However, mitigating *visual laziness* in multimodal models presents a different challenge. In the multimodal domain, a model can often produce a plausible, and sometimes even correct, response by relying on strong *language priors* without truly anchoring its prediction in visual evidence [37, 55, 59, 65].

Importantly, correctness is not a sufficient proxy for grounding. Prior work shows that models may achieve correct answers while relying on incorrect intermediate programs or inconsistent reasoning, essentially being "right for the wrong reasons" [10, 28, 43]. Outcome-only rewards can thus inadvertently reinforce *causal misattribution* [41], where the model reaches the right answer for purely linguistic reasons, preserving the very textual shortcuts we aim to remove.

Because our goal is to correct the model's insufficient reliance on visual evidence rather than merely improve the final response, `VIGIL` leverages the pairwise structure of DPO [47] to directly compare the seeing state against a counterfactual blind state within a unified log-likelihood objective. This comparison explicitly suppresses blind confidence by rewarding responses based on reliance on the visual input. In contrast, GRPO optimizes rewards over sampled responses in a group, and does not naturally encode such counterfactual state comparisons. Incorporating equivalent supervision would require constructing and evaluating additional blind trajectories for each sample, substantially increasing compute and memory. DPO therefore provides a highly efficient and stable offline

optimization objective for decoupling the multimodal manifold from the textual manifold. More detailed discussions and empirical results on GRPO are provided in Appendix 6.4.

3 Method

To address *visual laziness*, we propose `VIGIL`, a framework that explicitly anchors MLLM reasoning to visual evidence. It introduces a geometric constraint to maximize the mutual information between visual inputs and model responses, preventing the model from falling into language-driven optimization shortcuts.

3.1 Grounding via Information Maximization

We posit that an ideal MLLM should not merely generate high-probability tokens conditioned on *language priors*, but should instead maximize the information dependency between the visual input x_{v} and the generated response y.

Visual Information Gain (VIG). To quantify this dependency, we draw inspiration from *Maximum Mutual Information* (MMI) principles [24] and introduce VIG. We define VIG as the Pointwise Mutual Information (PMI) between the visual input x_{v} and the response y, conditioned on the instruction x_{t}:

Definition 1 (Visual Information Gain). *For a given multimodal input tuple* $(x_{\mathrm{v}}, x_{\mathrm{t}})$ and response y, the Visual Information Gain is defined as:

$$\mathrm{VIG}(y, x_{\mathrm{v}}|x_{\mathrm{t}}) = \log \frac{\pi_\theta(y|x_{\mathrm{v}}, x_{\mathrm{t}})}{\pi_\theta(y|x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})} \tag{3}$$

where $x_{\mathrm{v}}^{\emptyset}$ denotes the counterfactual attention-masked blind state, functionally equivalent to performing a *do-intervention* on the visual modality [41].

VIG measures the reduction in uncertainty about y provided *solely* by the visual evidence x_{v}. A VIG near zero indicates that the model is relying entirely on *language priors*, a core symptom of visual laziness.

The Constrained Optimization Problem. Standard DPO maximizes the margin between preferred (y_w) and disfavored (y_l) responses. However, as discussed in Sec. 2, maximizing the reward alone does not guarantee grounding. We instead formulate multimodal alignment as a *constrained optimization problem*, where we maximize the preference reward subject to the constraint that the **Conditional Mutual Information** $I(y; x_{\mathrm{v}}|x_{\mathrm{t}})$ exceeds a threshold δ:

$$\max_{\pi_\theta} \quad \mathbb{E}_{\mathcal{D}}\left[\mathcal{J}_{\mathrm{DPO}}(\pi_\theta)\right] \quad \text{s.t.} \quad \underbrace{\mathbb{E}_{\mathcal{D}}\left[\mathrm{VIG}(y_w, x_{\mathrm{v}}|x_{\mathrm{t}})\right]}_{\approx I(y_w; x_{\mathrm{v}}|x_{\mathrm{t}})} \geq \delta \tag{4}$$

This constraint encourages the generated response y_w contains sufficient information derived from x_{v}, thereby decoupling the multimodal manifold from the pure *textual manifold* and preventing optimization from collapsing onto language-only shortcuts.

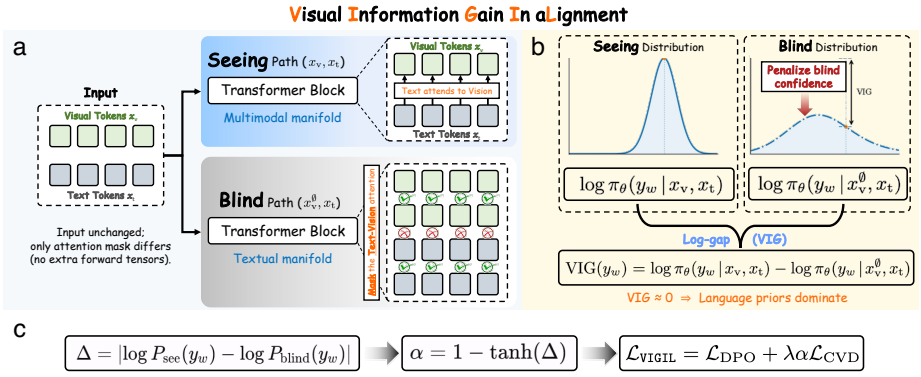

Fig. 3: Overview of VIGIL. (a) Dual-path forward pass. For each training sample, we run two matched conditions with identical image tokens, text tokens, positional embeddings, and projector outputs. In the **seeing** path, text tokens can attend to visual tokens normally. In the counterfactual **blind** path, the input tensors remain unchanged, but text-query rows are masked from visual-token key/value columns in every transformer layer, preventing text positions from accessing visual evidence through attention and yielding the blind state $x_{\mathrm{v}}^{\emptyset}$. The two paths differ only in attention connectivity, without additional input construction or image corruption. **(b)** Geometric constraint via Visual Information Gain (VIG). We explicitly contrast the likelihood of the preferred response under seeing and blind states. VIGIL penalizes high **blind confidence** and enlarges the log-gap $\mathrm{VIG}(y_w) = \log \pi_\theta(y_w \,|\, x_{\mathrm{v}}, x_{\mathrm{t}}) - \log \pi_\theta(y_w \,|\, x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})$, encouraging predictions to depend on visual evidence rather than language priors. **(c)** Dynamic gating and unified objective. A gating factor α is computed from the seeing–blind gap and adaptively balances the standard DPO loss with the Counterfactual Visual Decoupling (CVD) constraint, yielding $\mathcal{L}_{\mathrm{VIGIL}} = \mathcal{L}_{\mathrm{DPO}} + \lambda\alpha\mathcal{L}_{\mathrm{CVD}}$.

Derivation of the Counterfactual Visual Decoupling (CVD) Objective. To solve Eq. 4, we introduce a Lagrange multiplier $\lambda \geq 0$, leading to the following objective:

$$\mathcal{L}(\pi_\theta, \lambda) = \mathcal{J}_{\mathrm{DPO}}(\pi_\theta) + \lambda \mathbb{E}\left[\log \pi_\theta(y_w|x_{\mathrm{v}}, x_{\mathrm{t}}) - \log \pi_\theta(y_w|x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})\right] \tag{5}$$

We instantiate this Lagrangian term as a preference optimization problem, where the seeing state $(x_{\mathrm{v}}, x_{\mathrm{t}})$ is preferred over the attention-masked blind state $(x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})$ for the same response y_w. Applying the Bradley-Terry model, we derive our Counterfactual Visual Decoupling (CVD) loss:

$$\mathcal{L}_{\mathrm{CVD}}(\pi_\theta) = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x_{\mathrm{v}}, x_{\mathrm{t}})}{\pi_{\mathrm{ref}}(y_w|x_{\mathrm{v}}, x_{\mathrm{t}})} - \beta \log \frac{\pi_\theta(y_w|x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})}{\pi_{\mathrm{ref}}(y_w|x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})}\right)\right] \tag{6}$$

This loss enforces a preference for the seeing state over the blind state by penalizing responses whose likelihood does not decrease when visual evidence is removed through counterfactual masking, while the reference model π_{ref} serves to normalize language priors. Importantly, the frozen reference policy is evaluated under the same attention masks as the policy model: the seeing term uses the seeing mask and the blind term uses the blind mask. Thus, Eq. (6) compares

matched seeing and blind states, rather than normalizing a blind policy term with a seeing reference term.

3.2 VIGIL Implementation

The VIGIL framework is implemented through three integrated steps:

Step 1: Efficient Visual Counterfactuals. To simulate the blind state $x_{\mathrm{v}}^{\emptyset}$, we perform *attention masking*. We zero out the attention weights between textual and visual tokens in the forward pass, ensuring the probability drop is strictly causal to the missing visual modality. Pseudocode is provided in Appendix 7.

Step 2: The Geometric Decoupling Loss. We utilize $\mathcal{L}_{\mathrm{CVD}}$ (Eq. 6) as a geometric constraint. This pushes the optimization trajectory to differentiate between grounded reasoning and prior-driven hallucination.

Step 3: Self-Adversarial Hard Negative Mining. To complement our geometric constraints, we generate *hallucination-like* negatives y_{hard} by prompting the reference model to include plausible but absent details [46]. These samples provide sharper gradients for fine-grained visual discrimination.

3.3 Unified Objective with Dynamic Gating

To balance semantic alignment and visual grounding, we introduce a dynamic gating factor α:

$$\alpha = 1 - \tanh(|\log \pi_\theta(y_w|x_{\mathrm{v}}, x_{\mathrm{t}}) - \log \pi_\theta(y_w|x_{\mathrm{v}}^{\emptyset}, x_{\mathrm{t}})|) \tag{7}$$

Here, α automatically reduces the regularization strength for samples that already exhibit strong visual grounding (i.e., a high VIG gap). The final VIGIL objective is defined as:

$$\mathcal{L}_{\texttt{VIGIL}} = \mathcal{L}_{\mathrm{DPO}}(y_w, y_{\mathrm{hard}}) + \lambda \cdot \alpha \cdot \mathcal{L}_{\mathrm{CVD}} \tag{8}$$

This unified objective ensures the model learns not just what to say, but why to say it based on visual evidence.

3.4 Facing the Sword of Damocles: The Fragile Equilibrium of Grounded Reasoning

The heavy reliance on *language priors* in Multimodal Large Language Models is often described as a double-edged sword, or more aptly, a "Sword of Damocles" hanging over the model's reliability. While these priors enable models to generate fluent and contextually plausible narratives [7], they exist as a fragile equilibrium that easily collapses when the reasoning task demands strict adherence to visual evidence rather than statistical correlation [59, 65]. Our VIGIL framework acknowledges this inherent tension by moving beyond simple error correction. By introducing the counterfactual blind state $x_{\mathrm{v}}^{\emptyset}$, we establish a regulatory mechanism that quantifies the risk of over-reliance on linguistic context. This creates what we term a **Geometric Information Bottleneck**, which ensures that

the model's policy π_θ is causally anchored in the pixel space. Unlike existing reweighting schemes that treat all samples with equal priority [17], `VIGIL` explicitly penalizes high confidence in the absence of visual support, transforming the passive vulnerability of priors into a controlled and grounded reasoning process. To sum up, the primary advantages of our approach are threefold. (1) It offers a **geometric intervention** that is more fundamental than numerical loss reweighting, providing a direct path for causal alignment. (2) It achieves **remarkable efficiency**, requiring neither additional reference models nor full-dataset fine-tuning to achieve state-of-the-art performance. (3) By stabilizing the balance between linguistic fluency and visual fidelity, `VIGIL` encourages the emergence of spatial grounding capabilities, suggesting that teaching a model to avoid hallucinations inherently deepens its understanding of the visual world.

4 Experiments

4.1 Implementation Details

We implement all alignment algorithms using the OpenRLHF framework [16]. Training is conducted on a high-performance computing cluster equipped with NVIDIA H100 (80GB) GPU to ensure efficient distributed execution. To manage the memory requirements of training a 72B parameter model with DPO, which necessitates maintaining both the policy and reference models simultaneously, we employ DeepSpeed ZeRO-3 [48] with CPU offloading for optimizer states. We also integrate FlashAttention-2 [9] to accelerate the processing of long-context multimodal sequences. For experiments at the 7B scale, including Qwen2.5-VL-7B [53] and LLaVA-OneVision-7B [22], we utilize standard Fully Sharded Data Parallel (FSDP [63] strategies on A100 nodes. We perform full parameter fine-tuning for all models to maximize learning plasticity. In alignment with recent findings in multimodal preference optimization [46], we use large global batch sizes to maintain training stability: 2048 for 72B models and 1024 for 7B models. All models are trained for exactly one epoch to prevent overfitting to the preference set. The learning rate is set to 5×10^{-7} using a cosine decay scheduler with a warmup ratio of 0.03. The KL penalty coefficient β is fixed at 0.1 across all DPO variants. For our proposed `VIGIL`, the dynamic gating factor α is computed on a per-batch basis. We implement counterfactual visual inputs efficiently by manipulating the attention mask during the forward pass rather than modifying physical input tensors, ensuring the blind forward pass incurs minimal computational overhead. To ensure a rigorous comparison, we reproduce all baselines, including SimPO [40], HA-DPO [64], and DA-DPO [46], on the same training data using the hyperparameters recommended in their official implementations.

4.2 Evaluation Benchmarks

We adopt a multi-layered evaluation protocol to assess the trade-off between visual faithfulness and reasoning intelligence. For hallucination evaluation, we

report F1-scores on the POPE benchmark [27] across its adversarial, popular, and random splits. As performance on POPE has begun to saturate for state-of-the-art models, we also prioritize AMBER [52], a generative benchmark requiring free-form descriptions verified against fine-grained annotations. This metric is sensitive to hallucination propagation issues in long-context generation. Following the protocol of DA-DPO [46], we include MMHal-Bench [50] to measure hallucination in realistic, open-ended VQA scenarios.

Meanwhile, to verify that our visual constraints do not degrade general intelligence, we evaluate models on multimodal reasoning benchmarks MathVista [38] and MMBench [35]. Additionally, we perform evaluations on text-only benchmarks, namely MMLU [15] and GSM8K [8], to show that `VIGIL` improves multimodal reasoning without compromising text-only capabilities (Appendix 9).

4.3 Baselines

To demonstrate the effectiveness of `VIGIL`, we benchmark against a diverse array of alignment strategies categorized as follows:

General Alignment Methods. We consider SFT as the performance baseline. For preference alignment, we employ standard DPO [47] and SimPO [40]. SimPO is a reference-free algorithm that allows us to determine if performance gains stem from our visual constraints or simply from improved optimization stability. Comparing `VIGIL` against SimPO helps identify whether a superior optimizer is sufficient for multimodal grounding or if explicit visual modeling is required.

Hallucination-Specific Optimization. We compare `VIGIL` against HA-DPO [64], which mitigates hallucination by training on synthetic negative captions generated through text rewriting. We also include the state-of-the-art DA-DPO [46], which addresses overfitting by reweighting preference pairs based on difficulty. Unlike `VIGIL`, which utilizes physical interventions through counterfactual visual inputs, DA-DPO relies on the implicit difficulty distribution of the dataset.

Inference-Time Intervention. To verify the advantages of training-time alignment over inference-time patching, we compare our method with visual contrastive decoding (VCD) [14]. VCD penalizes hallucination by contrasting logits from original and distorted visual inputs during decoding. While effective, VCD increases inference latency and is sensitive to hyperparameter settings (See Appendix 11). `VIGIL` instead exposes the model to analogous visual contrast during post-training, resulting in an inference-efficient policy.

4.4 Main Results

Model Size Scaling Behavior. We first assess the performance of `VIGIL` and competing methods as we scale the Qwen2.5-VL base model from 7B to 72B (Table 1). `VIGIL` achieves state-of-the-art results across all evaluated metrics. A notable insight from these results is the scaling behavior of the model: while `VIGIL` improves the 7B model by 4.1 percentage points on POPE_{Adv}, this gain increases to 5.3 percentage points for the 72B model. This performance supports the existence

Table 1: Main Results across Model Scales. We report performance on both Qwen2.5-VL-7B and 72B to demonstrate the scalability of VIGIL. Performance gains are more pronounced on the 72B model, supporting the hypothesis that stronger foundation models utilize counterfactual visual signals more effectively.

Base Model	Method	Hallucination					System-2 Reasoning		Avg Score↑
		$\text{POPE}_{\text{Rand}}$ ↑	POPE_{Pop} ↑	POPE_{Adv} ↑	$\text{AMBER}_{\text{Gen}}$ ↑	$\text{MMHal}_{\text{Score}}$ ↑	MathVista ↑	MMBench ↑	
Qwen2.5-VL-**7B** [53]	SFT only	85.1	83.0	80.5	41.2	34.5	48.2	70.5	63.3
	SFT + DPO [47]	86.2	84.5	82.8	42.5	36.8	48.0	71.2	64.6
	SFT + SimPO [40]	86.5 (+0.3)	84.8 (+0.3)	83.1 (+0.3)	42.8 (+0.3)	37.0 (+0.2)	49.1 (+1.1)	71.5 (+0.3)	64.9 (+0.3)
	SFT + DA-DPO [46]	87.0 (+0.8)	85.8 (+1.3)	84.2 (+1.4)	44.8 (+2.3)	38.5 (+1.7)	48.8 (+0.8)	72.0 (+0.8)	65.9 (+1.3)
	SFT + VIGIL **(Ours)**	**88.5** (+2.3)	**87.2** (+2.7)	**86.9** (+4.1)	**46.5** (+4.0)	**40.2** (+3.4)	**49.5** (+1.5)	**72.5** (+1.3)	**67.3** (+2.7)
Qwen2.5-VL-**72B** [53]	SFT only	86.5	84.2	82.1	45.3	38.2	54.5	76.8	66.8
	SFT + DPO [47]	87.8	85.9	84.5	46.8	40.5	54.1	77.2	68.1
	SFT + SimPO [40]	88.1 (+0.3)	86.2 (+0.3)	84.8 (+0.3)	47.0 (+0.2)	41.1 (+0.6)	55.2 (+1.1)	77.5 (+0.3)	68.6 (+0.5)
	SFT + HA-DPO [64]	88.5 (+0.7)	87.1 (+1.2)	86.2 (+1.7)	48.5 (+1.7)	42.8 (+2.3)	53.8 (−0.3)	76.5 (−0.7)	69.1 (+1.0)
	SFT + DA-DPO [46]	89.2 (+1.4)	88.0 (+2.1)	87.4 (+2.9)	49.8 (+3.0)	44.2 (+3.7)	55.4 (+1.3)	77.8 (+0.6)	70.3 (+2.2)
	SFT + VIGIL **(Ours)**	**91.5** (+3.7)	**90.2** (+4.3)	**89.8** (+5.3)	**52.2** (+5.4)	**46.0** (+5.5)	**56.6** (+2.5)	**78.5** (+1.3)	**72.1** (+4.0)

Table 2: Generalization across base model architectures. We validate VIGIL across diverse architectures: LLaVA-OneVision and InternVL2.5-26B. VIGIL consistently outperforms standard DPO and DA-DPO. Performance gains are calculated relative to the standard DPO baseline. For the MMHal benchmark, we report the hallucination rate to provide a fine-grained measure of grounding failure.

Base Model	Method	Hallucination			General Capabilities		
		POPE_{Adv} ↑	$\text{AMBER}_{\text{Hal}}$ ↓	$\text{MMHal}_{\text{HalRate}}$ ↓	MathVista ↑	MMBench ↑	SeedBench ↑
LLaVA-OneVision-7B [22] *(SigLIP + MLP)*	SFT only	80.5	35.4	2.76	51.2	72.1	71.5
	SFT + DPO [47]	82.8	34.3	2.61	50.8	72.5	71.8
	SFT + DA-DPO [46]	84.2 (+1.4)	28.0 (−6.3)	2.78 (+0.17)	51.5 (+0.7)	73.0 (+0.5)	72.4 (+0.6)
	SFT + VIGIL **(Ours)**	**86.9** (+4.1)	**25.1** (−9.2)	**2.45** (−0.16)	**52.8** (+2.0)	**73.8** (+1.3)	**73.1** (+1.3)
InternVL2.5-26B [6] *(Large ViT Encoder)*	SFT only	84.1	31.2	2.50	58.4	79.2	76.4
	SFT + DPO [47]	85.5	29.8	2.35	57.9	79.5	76.8
	SFT + DA-DPO [46]	86.8 (+1.3)	25.5 (−4.3)	2.40 (+0.05)	58.8 (+0.9)	80.1 (+0.6)	77.2 (+0.4)
	SFT + VIGIL **(Ours)**	**89.4** (+3.9)	**22.3** (−7.5)	**2.15** (−0.20)	**60.1** (+2.2)	**81.3** (+1.8)	**78.0** (+1.2)

of a post-training scaling law, suggesting that larger models are better positioned to interpret the visual counterfactuals introduced by our method. Whether this advantage arises from a higher knowledge capacity [1] or from a more favorable latent representation geometry [33] remains an interesting question for future study. In contrast, text-centric methods like SimPO show diminishing returns on multimodal tasks, indicating that optimization stability alone cannot compensate for the absence of modality-specific constraints.

Cross-Architecture Universality. We further evaluate the universality of our approach on different base model architectures and designs (Table 2). VIGIL consistently outperforms standard DPO across different architectures, including the projector-based LLaVA-OneVision-7B and the visual-centric InternVL2.5-26B [6], which utilizes a 6B vision encoder. For both architectures, VIGIL achieves an improvement of approximately 4.0 point on POPE_{Adv}. These results demonstrate that even models with powerful visual encoders remain susceptible to visual laziness, a failure mode that our counterfactual masking effectively addresses.

Table 3: Component ablation. We analyze the contribution of each module on Qwen2.5-VL-7B. The *Visual Anchor* ($x_v^{\emptyset}$) is the most critical component, providing the largest performance boost.

Method	$\texttt{POPE}_{\text{Rand}}$ ↑	$\texttt{AMBER}_{\text{Gen}}$ ↑	MathVista ↑
SFT + VIGIL **(Ours)**	**88.5**	**46.5**	**49.5**
– Visual Anchor ($x_v^{\emptyset}$)	85.3 (−3.2)	43.1 (−3.4)	49.0 (−0.5)
– Hard Negatives	86.5 (−2.0)	44.2 (−2.3)	48.8 (−0.7)
– Dynamic Gating (α)	87.4 (−1.1)	45.4 (−1.1)	49.2 (−0.3)
SFT + DPO [47]	86.2	42.5	48.0

Table 4: Alignment strategy comparison. Our explicit masking (Geometric Decoupling) strategy significantly outperforms implicit difficulty reweighting (DA-DPO) and filtering.

Strategy	Core Mechanism	$\texttt{POPE}_{\text{Rand}}$ ↑
DPO [47]	Equality	82.8
Filtering (10%)	Remove Easy	84.3 (+1.5)
DA-DPO [46]	Reweighting	85.7 (+2.9)
CVD-Masking	Physical Constraint	**88.5** (+5.7)

Table 5: Performance across Visual Dependency Buckets. Instead of generic difficulty, we categorize samples by *Visual Dependency Index (VDI)*. VIGIL achieves substantial gains on samples with high visual dependency, demonstrating it effectively fixes "blind" hallucinations where baselines fail. Performances are evaluated on $\texttt{POPE}_{\text{Rand}}$.

Method	VDI Bucket		
	Low (Text-heavy)	Medium (In between)	High (Fine-grained)
SFT + DPO [47]	88.5	84.1	78.2
SFT + DA-DPO [46]	89.0	86.5	82.4
SFT + VIGIL (Ours)	**89.2**	**88.8**	**87.5**

Table 6: Effect of VIGIL on localization. We evaluate Zero-shot Referring Expression Comprehension on RefCOCOg (val). Standard DPO often degrades spatial awareness. In contrast, VIGIL significantly improves localization accuracy, showcasing that our method physically anchors text to image regions.

Method	Hallucination	Localization
	$\texttt{POPE}_{\text{Adv}}$ ↑	RefCOCOg Acc@0.5 ↑
SFT only	80.5	45.2
SFT + DPO [47]	82.8	44.8 (−0.4)
SFT + DA-DPO [46]	84.2	46.1 (+0.9)
SFT + VIGIL **(Ours)**	**86.9**	**49.5** (+4.3)

4.5 Ablation Study

We conduct ablation studies on Qwen2.5-VL-7B to analyze the source of performance gains, focusing on component contributions, mechanism design, system robustness, efficiency, and a surprising study on emergent spatial grounding.

Impact of Key Components. The contribution of each module is shown in Table 3. Starting from the DPO baseline, the addition of hard negatives improves the POPE score by 2.0 percentage points, confirming the importance of high-quality preference data. The most significant improvement, a gain of 3.2 percentage points, is driven by the **Visual Anchor** ($x_v^{\emptyset}$). This result demonstrates that geometric contrast, specifically the ability to distinguish between the seeing and blind states, is essential for mitigating hallucinations. Moreover, the removal of dynamic gating results in a minor performance decrease, validating its role in stabilizing gradient flow during optimization.

Mechanism Analyses. We evaluate why VIGIL outperforms existing hard-negative mining techniques. Table 4 compares our explicit masking strategy with the implicit reweighting approach used in DA-DPO [46]. While reweighting provides a 2.9 point gain, the physical masking employed by VIGIL yields a 5.3 point increase. This suggests that physical intervention on the input geometry provides a more effective supervision signal than numerical adjustments to loss weights.

To further investigate this effect, Table 5 categorizes samples by their Visual Dependency Index (VDI), which measures the divergence in model predictions

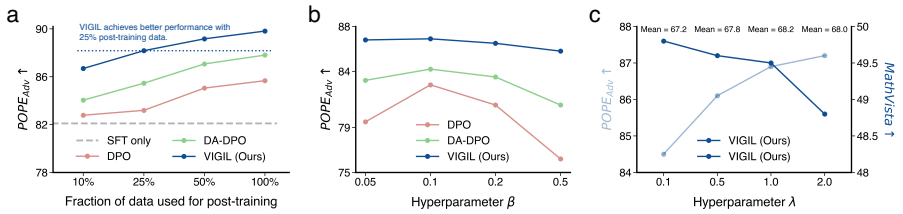

Fig. 4: Analyses of data efficiency and effects of hyperparameters. (a) Our proposed `VIGIL` is data-efficient. It achieves performance comparable to state-of-the-art methods using only 25% of the post-training data. **(b)** Our method is more robust to the KL penalty coefficient β than competing methods, and the results justifies the choice of $\beta = 0.1$. **(c)** The weighing coefficient $\lambda = 1.0$ leads to a good balance between hallucination mitigation and reasoning performance.

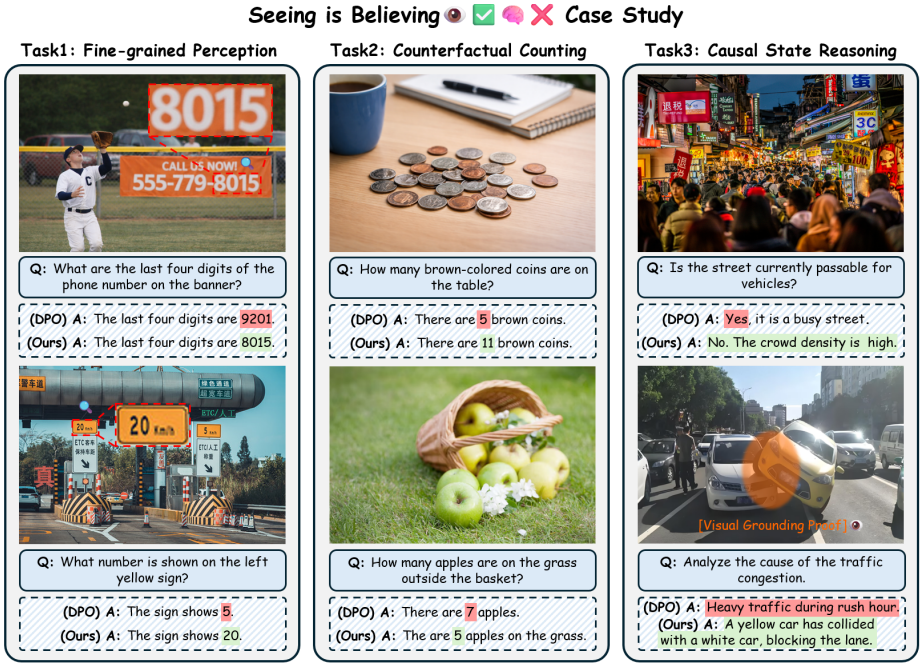

Fig. 5: Qualitative Comparison of Visual Grounding. MLLMs often exhibit *visual laziness* by relying on *language priors*, resulting in plausible but factually incorrect hallucinations (red). In contrast, `VIGIL` anchors reasoning to visual evidence x_v (green). **Task 1 (Left):** In fine-grained perception, `VIGIL` accurately identifies the digits 8015 as verified by the visual grounding box, whereas the DPO baseline hallucinates 9201. **Task 2 (Middle):** In counterfactual counting, the DPO baseline defaults to generic numbers based on priors, while `VIGIL` correctly identifies 11 coins and 5 apples. **Task 3 (Right):** In causal reasoning, instead of providing generic descriptions, `VIGIL` identifies specific physical states, such as vehicle collisions, to derive correct logical conclusions.

when the visual input is removed. Standard DPO performs poorly on high-

dependency samples, scoring 78.2, whereas VIGIL increases this to 87.5. These findings indicate that our method effectively guides the model to prioritize visual features when *language priors* are ambiguous.

Additional rigorous evaluations of VIGIL on its optimization trajectory through the lens of *Maximum Mutual Information* (MMI) [24] and *Rate-Distortion Theory* are shown in Appendix 12.

Robustness and Efficiency. Figure 4 illustrates the practical advantages of our framework. Regarding data efficiency, VIGIL trained on 25% of the dataset matches the performance of DA-DPO trained on the full dataset. Despite the overhead of masking, this efficiency reduces the total training wall-clock time by approximately 70% compared to full-data training (see Appendix 10). In terms of stability, VIGIL remains robust across a range of KL penalty coefficients, $\beta \in [0.05, 0.5]$, whereas standard DPO often exhibits instability at lower β values. We attribute this stability to the regularization effect of the visual anchor.

Emergent spatial grounding. Table 6 shows that VIGIL improves zero-shot RefCOCOg performance by 4.3 percentage points. Although the model is not explicitly trained on bounding box coordinates, it develops improved spatial localization capabilities to satisfy the counterfactual grounding objective.

4.6 Qualitative Analysis

To intuitively demonstrate how VIGIL mitigates *visual laziness*, we present a qualitative comparison against the standard DPO baseline in Figure 5. Both models are based on the Qwen2.5-VL-7B architecture. The results are categorized into three levels of difficulty: fine-grained perception, counterfactual counting, and causal state reasoning.

Fine-grained Perception. The first case illustrates the model's ability to resolve high-frequency visual details. In the baseball scenario, when asked for the phone number on the banner, the DPO baseline predicts 9201, a response that appears plausible based on inherent *language priors* but is factually incorrect. In contrast, VIGIL accurately identifies the digits 8015. The visual grounding proof, highlighted by the red bounding box, confirms that the model's attention is correctly anchored to the specific pixel region. This indicates that the model effectively utilizes visual evidence before generating the corresponding tokens.

Visual Counting. Counting serves as a benchmark for visual grounding. As shown in the second case, the baseline exhibits a strong dependency on *language priors*. For the brown coins and apples, the DPO baseline predicts 5 and 7, respectively. These numbers are statistically frequent in training corpora yet incorrect for these specific images. By incorporating geometric constraints, VIGIL suppresses prior-based guessing and correctly identifies 11 coins and 5 apples. This demonstrates that our method encourages the model to perform actual enumeration rather than relying on probabilistic retrieval from its language components.

Causal State Reasoning. The third case highlights the capability gap in reasoning about physical states. In the traffic accident scene, the baseline attributes the congestion to heavy traffic during rush hour, representing a generic guess that ignores visual reality. Conversely, `VIGIL` identifies the specific causal mechanism, stating that a yellow car has collided with a white car. Similarly, in the crowded street scene, our model correctly identifies high crowd density as the physical barrier to vehicle passage. These observations confirm that `VIGIL` enables grounded reasoning, where logical conclusions are derived from accurate visual premises rather than language-based associations.

5 Conclusion

In this paper, we propose `VIGIL`, a principled post-training framework that mitigates hallucinations in MLLMs by maximizing visual information gain. Unlike standard preference optimization that often leads to *visual laziness*, our approach introduces a geometric constraint through counterfactual visual decoupling. By penalizing model confidence in the blind state $x_{\mathrm{v}}^{\emptyset}$, we force the optimization trajectory to anchor reasoning in actual visual features x_{v} rather than inherent *language priors*. Our evaluations demonstrate that `VIGIL` consistently outperforms state-of-the-art baselines while exhibiting superior scaling behavior and data efficiency. Notably, the emergence of spatial grounding capabilities suggests that our geometric anchoring effectively aligns the model's latent perception with its explicit generation.

Despite its effectiveness, a potential limitation of our current framework lies in the simplistic construction of the counterfactual state via zero-out masking, which may not capture more nuanced cross-modal conflicts in highly cluttered environments. Future work could explore more granular counterfactual interventions, such as object-level spectral filtering or semantic-preserving transformations, to further refine the model's visual sensitivity. We hope our work encourages a shift toward causal and information-theoretic alignment in multimodal learning.

Acknowledgements

This research is supported in part by the U.S. National Science Foundation (OAC-2118240, HDR Institute: Imageomics).

References

1. Allen-Zhu, Z., Li, Y.: Physics of language models: Part 3.3, knowledge capacity scaling laws. In: International Conference on Learning Representations (2025) 11
2. Bai, T., Fan, Y., Jiantao, Q., Sun, F., Song, J., Han, J., Liu, Z., He, C., Zhang, W., Yuan, B.: Hallucination at a glance: Controlled visual edits and fine-grained multimodal learning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 2

3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024) 21
4. Cao, M., Zhao, H., Zhang, C., Chang, X., Reid, I., Liang, X.: Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. arXiv preprint arXiv:2505.20272 (2025) 21
5. Chen, S., Liu, J., Han, Z., Xia, Y., Cremers, D., Torr, P., Tresp, V., Gu, J.: True multimodal in-context learning needs attention to the visual context. In: Second Conference on Language Modeling (2025) 21
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 2, 11
7. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of machine learning research **24**(240), 1–113 (2023) 8
8. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021) 10
9. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations. vol. 2024, pp. 35549–35562 (2024) 9
10. Gandhi, M., Gul, M.O., Prakash, E., Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Measuring compositional consistency for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5046–5055 (2022) 5
11. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025) 5, 22
12. Guo, X., Tyagi, U., Gosai, A., Vergara, P., Park, J., Montoya, E.G.H., Zhang, C.B.C., Hu, B., He, Y., Liu, B., et al.: Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning. arXiv preprint arXiv:2510.12712 (2025) 21
13. Hao, H., Min, D., Zhang, Z., Zhang, Y., Xu, M., Ge, Y., Cheng, L.: Poise: Position-aware undetectable skill injection on llm agents. arXiv preprint arXiv:2606.07943 (2026) 21
14. He, Y., Sun, H., Qi, Q., Zhuang, Z., Ren, P., Wang, H., Nan, Y., Wang, J.: Mitigating object hallucination in large vision-language models via visual attention direct preference optimization. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025) 10, 27, 28
15. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: International Conference on Learning Representations (2021) 10
16. Hu, J., Wu, X., Shen, W., Liu, J.K., Zhu, Z., Wang, W., Jiang, S., Wang, H., Chen, H., Chen, B., et al.: Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (2025) 9
17. Hu, X., Ni, H., Zhang, Y., Hamm, J., Li, Z., Ding, Z.: Seeing clearly, reasoning confidently: Plug-and-play remedies for vision language model blindness. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2026) 9

18. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024) 21
19. Kang, W., Kuen, J., Ren, M., Wei, Z., Yan, Y., Liu, K.: Vgent: Visual grounding via modular design for disentangling reasoning and prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41160–41170 (2026) 2, 5, 21
20. Kil, J., Mai, Z., Lee, J., Chowdhury, A., Wang, Z., Cheng, K., Wang, L., Liu, Y., Chao, W.L.: Mllm-compbench: A comparative reasoning benchmark for multimodal llms. Advances in Neural Information Processing Systems **37**, 28798–28827 (2024) 21
21. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024) 2, 21
22. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2025) 2, 9, 11, 21
23. Li, H., Jiang, L., Xiao, X., Wang, T., Yi, H., Wu, B., Cai, D.: Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12737–12746 (October 2025) 21
24. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B.: A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 110–119 (2016) 6, 14, 27
25. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 22
26. Li, T., He, K.: Back to basics: Let denoising generative models denoise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 36115–36125 (2026) 4
27. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023) 10, 21
28. Liang, Y., Li, M., Fan, C., Li, Z., Nguyen, D., Cobbina, K.A., Bhardwaj, S., Chen, J., Liu, F., Zhou, T.: Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025) 5
29. Liao, C.T., Xiao, X., Meng, C., Chen, Z., Qiao, Y., Zhou, W., Wang, T., Zheng, X., Cao, X.: Spamem: Benchmarking dynamic spatial reasoning via perception-memory integration in embodied environments. arXiv preprint arXiv:2604.22409 (2026) 22
30. Liao, D., Liu, C., Christensen, B.W., Tong, A., Huguet, G., Wolf, G., Nickel, M., Adelstein, I., Krishnaswamy, S.: Assessing neural network representations during training using noise-resilient diffusion spectral entropy. In: 2024 58th Annual Conference on Information Sciences and Systems (CISS). pp. 1–6. IEEE (2024) 4

31. Liao, D., Liu, C., Sun, X., Tang, D., Wang, H., Youlten, S., Gopinath, S.K., Lee, H., Strayer, E.C., Giraldez, A.J., et al.: Rnagenscape: property-guided optimization and interpolation of mrna sequences with manifold langevin dynamics. arXiv preprint arXiv:2510.24736 (2025) 4
32. Lin, Z., Liu, Y., Yang, Y., Tao, L., Ye, D.: Adaptvision: Efficient vision-language models via adaptive visual acquisition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11923–11932 (2026) 3, 21
33. Liu, C., Sun, X., Xiao, X., Van Tassel, A., Xu, K., Reimann, K., Liao, D., Gerstein, M., Wang, T., Wang, X., Krishnaswamy, S.: Dispersion loss counteracts embedding condensation and improves generalization in small language models. In: International Conference on Machine Learning. PMLR (2026) 11
34. Liu, M., Zhang, Y., Liu, S., Wang, L., Zhang, W.: Wan-r1: Verifiable-reinforcement learning for video reasoning. arXiv preprint arXiv:2603.27866 (2026) 22
35. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024) 10
36. Liu, Y., Qin, L., Wang, S.: Small drafts, big verdict: Information-intensive visual reasoning via speculation. arXiv preprint arXiv:2510.20812 (2025) 21
37. Long, L., Oh, C., Park, S., Li, S.: Understanding language prior of LVLMs by contrasting chain-of-embedding. In: The Fourteenth International Conference on Learning Representations (2026) 2, 5
38. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: The Twelfth International Conference on Learning Representations (2024) 10
39. Meilă, M., Zhang, H.: Manifold learning: What, how, and why. Annual Review of Statistics and Its Application **11**(1), 393–417 (2024) 4
40. Meng, Y., Xia, M., Chen, D.: Simpo: Simple preference optimization with a reference-free reward. Advances in Neural Information Processing Systems **37**, 124198–124235 (2024) 9, 10, 11
41. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12700–12710 (2021) 5, 6, 28
42. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022) 21
43. Panagopoulou, A., Zhou, H., Savarese, S., Xiong, C., Callison-Burch, C., Yatskar, M., Niebles, J.C.: Viunit: Visual unit tests for more robust visual programming. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24646–24656 (2025) 5
44. Pearl, J.: Causality. Cambridge university press (2009) 22
45. Peng, F., Yang, X., Wang, Y., Xu, C.: Group-relative visual discrimination enhancement for unlocking intrinsic capability of mllms. IEEE Transactions on Circuits and Systems for Video Technology (2026) 21
46. Qiu, L., Ning, S., Zhang, C., Sun, J., He, X.: Da-dpo: Cost-efficient difficulty-aware preference optimization for reducing mllm hallucinations. Transactions on Machine Learning Research (2025) 3, 5, 8, 9, 10, 11, 12, 21
47. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances

in neural information processing systems **36**, 53728–53741 (2023) 3, 4, 5, 10, 11, 12, 21
48. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: international conference for high performance computing, networking, storage and analysis. pp. 1–16. IEEE (2020) 9
49. Sun, X., Liao, D., MacDonald, K., Zhang, Y., Huguet, G., Wolf, G., Adelstein, I., Rudner, T.G., Krishnaswamy, S.: Geometry-aware generative autoencoders for warped riemannian metric learning and generative modeling on data manifolds. In: The 28th International Conference on Artificial Intelligence and Statistics (2025) 4
50. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 13088–13110 (2024) 10
51. Wang, J., Zhang, Y., Ding, Z., Hamm, J.: Doctor approved: Generating medically accurate skin disease images through ai-expert feedback. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) 21
52. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397 (2023) 10
53. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 9, 11, 21, 27
54. Wang, Z., Xiong, F., Lin, L., Hu, X., Wang, Y., Wang, Y., Zhang, M., Chu, X.: Visually-guided policy optimization for multimodal reasoning. arXiv preprint arXiv:2604.09349 (2026) 21
55. Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., et al.: Perception-aware policy optimization for multimodal reasoning. arXiv preprint arXiv:2507.06448 (2025) 5
56. Xiao, X., Ma, C., Zhang, Y., Liu, C., Wang, Z., Li, Y., Zhao, L., Hu, G., Wang, T., Xu, H.: Not all directions matter: Towards structured and task-aware low-rank model adaptation. arXiv preprint arXiv:2603.14228 (2026) 21
57. Xu, T., Jing, H., Li, Y., Wei, Y., Feng, J., Chen, G., Gao, H., Zhang, T., Chen, F.: Defacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. arXiv preprint arXiv:2509.20912 (2025) 21
58. Zhang, D., Li, Z.Z., Zhang, M.L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Chen, X., Zhang, Y., et al.: From system 1 to system 2: a survey of reasoning large language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2
59. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: The Thirteenth International Conference on Learning Representations (2025) 2, 4, 5, 8, 21, 28
60. Zhang, Y., Ge, Y., Xu, W., Xu, Y., Hamm, J., Reddy, C.K.: Visual exclusivity attacks: Automatic multimodal red teaming via agentic planning. arXiv preprint arXiv:2603.20198 (2026) 22
61. Zhao, L., Jiang, X., Xiao, X., Fan, Q., Lu, L., Wang, Y., Lin, X., Camps, O., Zhao, P., Gu, J.: Hieramp: Coarse-to-fine autoregressive amplification for generative dataset distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41688–41698 (2026) 21

62. Zhao, L., Li, H., Ning, X., Jiang, X.: Thinimg: Cross-modal steganography for presenting talking heads in images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5553–5562 (2024) 21
63. Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., et al.: Pytorch fsdp: Experiences on scaling fully sharded data parallel. Proceedings of the VLDB Endowment **16**(12), 3848–3860 (2023) 9
64. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond multimodal hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025) 9, 10, 11, 21
65. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. In: The Twelfth International Conference on Learning Representations (2024) 5, 8
66. Zhu, D., Wang, Z., Xiao, X., Jiang, H., Vahidian, S., Chao, W.L., Berger-Wolf, T., Su, Y., Vatsavai, R., Gu, J.: Leveraging latent visual reasoning in silence. arXiv preprint arXiv:2605.18641 (2026) 2, 21