DRIFTDPO: MECHANISM-GUIDED PREFERENCE OPTIMIZATION FOR HALLUCINATION MITIGATION IN DEEP-FUSION VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object hallucination in vision-language models (VLMs) is widely attributed to language-prior dominance, motivating contrastive decoding (CD) methods that suppress this prior. However, CD consistently worsens hallucination on modern deep-fusion VLMs, suggesting the prevailing theory does not hold for these architectures. Through token-level probing of internal representations, we discover that the theory is not merely incomplete but *inverted*: on deep-fusion VLMs, hallucination is driven by visual-mean alignment rather than language-prior dominance, and this mechanism is architecture-gated—active in deep-fusion models, absent in shallow-projection models, and reversed in next-generation ones. Based on this finding, we propose DRIFTDPO, which derives its preference signal entirely from the model's own hidden states, requiring no external annotation. A per-token analysis reveals that the caption-level signal is biased by function-word dominance, necessitating a direction correction step to align the training objective with the object-level mechanism. DRIFTDPO achieves state-of-the-art hallucination reduction on multiple benchmarks while *increasing* coverage at zero inference overhead—the only method to simultaneously achieve all three. Mechanism verification confirms that training genuinely internalizes the signal into model weights rather than superficially suppressing it, revealing a coupling property unique to self-supervised rewards.

1 INTRODUCTION

Object hallucination—vision-language models (VLMs) generating visually unsupported objects—impedes deployment in safety-critical applications such as autonomous driving and medical image analysis (Rohrbach et al., 2018; Li et al., 2023; Liu et al., 2024a; Bai et al., 2024; Rawte et al., 2023), and this gap becomes a practical barrier as VLMs spread into real-world systems (Brown et al., 2020; Touvron et al., 2023a;b), making effective mitigation a prerequisite for trustworthy deployment.

The dominant training-free mitigation is contrastive decoding (CD) (Leng et al., 2024; Wang et al., 2024), which suppresses the language prior by subtracting the logits of a degraded input. It rests on the *language-prior-dominance* theory—hallucinations arise because statistical priors override visual evidence—which was validated on shallow-projection architectures such as LLaVA-1.5 (Liu et al., 2023; 2024c), where visual features enter through a lightweight projection layer. Numerous CD variants (Zhao et al., 2024; Chen et al., 2024a; Zhou et al., 2024) and training-based methods (Yu et al., 2024; Sun et al., 2024; Zhao et al., 2023) build on this theory.

However, a problem that has been overlooked is that the language-prior-dominance theory has *never been tested* on modern deep-fusion architectures (Qwen2.5-VL (Bai et al., 2025), InternVL (Chen et al., 2024c;b)) that now predominate in practice. On these models, CD methods consistently *worsen* hallucination rather than reduce it, with similar reversals observed across independent deep-fusion architectures. This systematic failure indicates that the underlying mechanism assumption does not hold—the field has been treating the wrong disease.

The central finding, revealed through token-level probing of internal representations, is that the hallucination mechanism is *architecture-gated*: on deep-fusion VLMs, hallucinated object tokens show *lower* language-prior probability but *higher* visual-mean alignment—the cosine similarity between

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

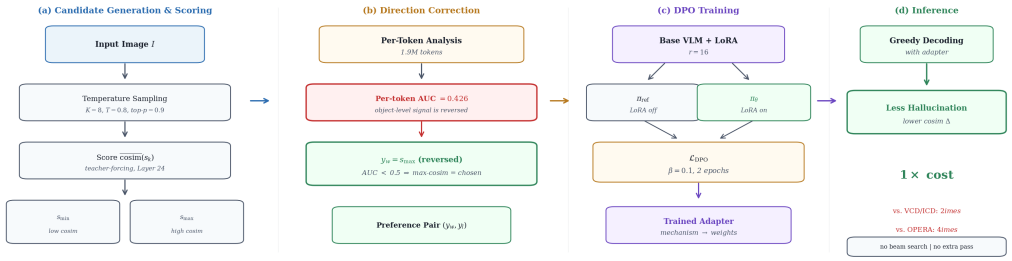


Figure 1: Overview of DRIFTDPO. (a) Candidates are generated via temperature sampling and scored by caption-level cosim. (b) A per-token analysis reveals the object-level signal is *reversed* (AUC<0.5), requiring direction correction. (c) DPO training internalizes the mechanism into LoRA weights. (d) Inference uses standard greedy decoding at zero overhead.

generation-step hidden states and the visual token mean. The theory is thus not merely incomplete but *inverted*: CD suppresses the language prior, but since hallucination is actually driven by visual-mean alignment, CD amplifies the same mechanism causing it. This mechanism varies by architecture: active in deep-fusion models, absent in shallow-projection models, and reversed in next-generation ones. This establishes that hallucination has no single universal cause, and effective mitigation must begin by diagnosing the operative mechanism for the target architecture.

Because visual-mean alignment is self-measurable from the model's own hidden states and causally linked to hallucination, it can serve as a *self-supervised* reward—no external annotation is needed. We instantiate this as DRIFTDPO, which constructs preference pairs by ranking sampled captions by their alignment score and trains with DPO (Rafailov et al., 2023). A per-token analysis reveals that the caption-level signal is biased by function-word dominance, and the object-level signal is *reversed*, requiring a direction correction step. At inference, the trained model uses standard greedy decoding with zero overhead—the mechanism is fully internalized into weights. Figure 1 illustrates the complete pipeline.

Our contributions are:

1. **Mechanism identification**: We show that the language-prior-dominance theory is *inverted* on deep-fusion VLMs and identify visual-mean alignment as the operative driver. The mechanism is *architecture-gated*: active in deep-fusion, absent in shallow-projection, and reversed in next-generation architectures. This finding reframes hallucination as architecture-specific rather than universal.
2. **DRIFTDPO**: The *first* method to derive its preference signal solely from the model's internal mechanism, requiring no external annotation. With direction correction from per-token analysis, it achieves state-of-the-art hallucination reduction on multiple benchmarks while *increasing* coverage at zero inference overhead—the only method to simultaneously achieve all three.
3. **Mechanism verification**: Training *consumes* the internal signal, demonstrating genuine internalization rather than superficial suppression; non-monotonic data scaling reveals a coupling property unique to self-supervised rewards.

2 RELATED WORK

2.1 HALLUCINATION MECHANISM AND THEORY

The dominant theory—language-prior dominance—was established by VCD (Leng et al., 2024) and ICD (Wang et al., 2024) on LLaVA-1.5 (Liu et al., 2023), a shallow-projection architecture where visual features are injected via a lightweight linear projection. Kaul et al. (2024) observed low correlation between discriminative and generative hallucination, and surveys (Liu et al., 2024a; Bai et al., 2024; Rawte et al., 2023) catalogued hallucination types while assuming language-prior dominance as universal. However, these studies were conducted exclusively on shallow-projection

architectures; the theory has not been tested on modern deep-fusion models, where visual and linguistic representations are entangled through shared transformer layers (Chen et al., 2024c; Bai et al., 2025). *In contrast, our work directly probes the internal mechanism on deep-fusion architectures, revealing that the theory is inverted and identifying visual-mean alignment as the operative driver.*

2.2 Hallucination Mitigation Methods

Existing mitigation methods fall into two categories. **Training-free methods** operate at inference time: VCD (Leng et al., 2024) and ICD (Wang et al., 2024) suppress language priors via contrastive decoding; OPERA (Huang et al., 2024) penalizes attention sinks; HALC (Chen et al., 2024a) uses adaptive focal-contrast decoding; Woodpecker (Yin et al., 2023) relies on external detectors; M3ID (Liu et al., 2024b) uses multi-modal instruction decoding. These impose 2–4$\times$ inference overhead and, as we show, backfire on deep-fusion architectures. **Training-based methods** use preference optimization: RLHF-V (Yu et al., 2024), LLaVA-RLHF (Sun et al., 2024), and HA-DPO (Zhao et al., 2023) all depend on external supervision from human feedback or GPT-4, creating a scalability bottleneck. *Unlike all prior methods,* DriftDPO *requires no external annotation and achieves zero inference overhead by internalizing the mechanism signal into model weights.*

2.3 Mechanistic Interpretability

Probing has been used to analyze internal representations (Belinkov, 2022; Tenney et al., 2019; Hewitt & Manning, 2019; Clark et al., 2019; Rogers et al., 2020); Burns et al. (2023) discovered latent knowledge without supervision and Elhage et al. (2021) provided a mathematical framework for transformer circuits. However, it has rarely been applied to identify the *causal* mechanism of hallucination in VLMs, nor to derive a self-supervised reward from internal states. Concurrent work interprets object hallucination through attention lenses (Jiang et al., 2025) and steers decoding from hidden visual information (Li et al., 2025); *we instead turn the probed signal into a self-supervised reward internalized into the weights. Our work bridges this gap: we use probing not merely to describe internal states but to identify the operative hallucination driver, validate its causal role, and internalize it as a self-supervised training signal.*

3 Motivation: The Visual-Mean Drift Mechanism

The CD reversal on deep-fusion VLMs (§3.1) is a *symptom* that the underlying mechanism assumption is wrong. To diagnose the *cause*, we proceed in four steps: (1) document the CD failure, (2) identify the operative mechanism via probing, (3) confirm it is architecture-gated, and (4) validate causality.

3.1 The CD Reversal: A Symptom of Wrong Diagnosis

Contrastive decoding adjusts the output distribution by subtracting the logits of a *degraded* forward pass from the original visual forward pass. Formally, VCD (Leng et al., 2024) perturbs the image and ICD (Wang et al., 2024) distorts the instruction, both producing a contrastive logit $\mathbf{z}_{\text{cd}} \in \mathbb{R}^{|V|}$. The adjusted logit at step t is:

$$\tilde{\mathbf{z}}_t = (1+\alpha)\,\mathbf{z}_{\text{vis}}^{(t)} - \alpha\,\mathbf{z}_{\text{cd}}^{(t)} = \mathbf{z}_{\text{vis}}^{(t)} + \alpha\,\underbrace{(\mathbf{z}_{\text{vis}}^{(t)} - \mathbf{z}_{\text{cd}}^{(t)})}_{\Delta\mathbf{z}_t} \tag{1}$$

where $\mathbf{z}_{\text{vis}}^{(t)}$ is the logit from the original (visual) forward pass, α controls the contrastive strength, and $\Delta\mathbf{z}_t$ is the *visual contribution*—how much the visual input changes the prediction relative to the degraded input. An adaptive plausibility constraint further masks implausible tokens.

The language-prior-dominance theory predicts that hallucinated tokens are driven by the language prior (high p_{prior}, low visual contribution $\Delta\mathbf{z}$), so CD's subtraction should suppress them. However, on deep-fusion VLMs, CD consistently produces the opposite effect: VCD (+15.9%) and ICD (+13.3%) both *increase* hallucination on Qwen2.5-VL-3B (Table 1(a)). This reversal is systematic across independent deep-fusion architectures (InternVL3.5-4B, Qwen2.5-VL-7B), ruling out implementation error. The systematic nature of this failure implies that the visual contribution $\Delta\mathbf{z}$ does

Table 1: Mechanism probing evidence. (a) CD methods worsen hallucination on deep-fusion VLMs, a systematic reversal. (b) Three-signal probing on Qwen2.5-VL-3B (Layer 24): cosim is the only signal with AUC > 0.5, identifying visual-mean alignment as the operative driver. (c) cosim AUC across architectures: the mechanism is active in deep-fusion models, absent in shallow-projection, and reversed in next-generation.

(a) CD Reversal			(b) Three-Signal Probing		
Method	CHAIRi	Δ	Signal	AUC	Interpretation
Greedy	0.1810	—	p_{prior}	0.439	Prior *lower* at obj. tokens
VCD	0.2097	+15.9%	cosim	**0.589**	**Alignment *higher***
ICD	0.2051	+13.3%	Δlogit	0.422	Not predictive

(c) Architecture Gating: cosim AUC Across Models

Model	Architecture	cosim AUC	CD effect
Qwen2.5-VL-3B	Deep-fusion	0.589	worse
InternVL3.5-4B	Deep-fusion	0.618	worse
LLaVA-1.5-7B	Shallow-projection	0.446	effective
Qwen3-VL-4B	Deep-fusion (new gen.)	0.390	reversed

not separate hallucinated tokens from correct ones in the way the language-prior theory assumes, so CD amplifies rather than suppresses the objects it is meant to remove. This motivates a deeper question: *if language-prior dominance is not the driver, what is?*

3.2 Identifying the Operative Mechanism

To diagnose the cause, we probe three candidate signals on Layer 24 hidden states, following the probing methodology of (Belinkov, 2022; Burns et al., 2023). For each generation step t, let $\mathbf{z}_{\text{vis}}^{(t)}$ denote the full logit and $\mathbf{z}_{\text{prior}}^{(t)}$ the logit from a *vision-masked* forward pass (where visual token embeddings are zeroed). This masked pass is a *proxy* for the language prior—zeroing the embeddings does not remove visual information from the shared layers—so we read p_{prior} only alongside the visual logit (Eq. 1). We compute: (1) $p_{\text{prior}}(y_t) = \text{softmax}(\mathbf{z}_{\text{prior}}^{(t)})(y_t)$: the language-prior probability of the generated token; (2) $\text{cosim}(t)$: the cosine similarity between the hidden state $\mathbf{h}_t^{(l)}$ and the visual mean $\bar{\mathbf{v}} = \frac{1}{|\text{vis}|}\sum_{i\in\text{vis}} \mathbf{h}_i^{(l)}$ (Eq. 2); (3) $\Delta\text{logit}(y_t) = \mathbf{z}_{\text{vis}}^{(t)}(y_t) - \mathbf{z}_{\text{prior}}^{(t)}(y_t)$: the visual contribution, directly corresponding to $\Delta\mathbf{z}_t$ in the CD formula (Eq. 1). We label each token as hallucinated object vs. background and measure one-vs-rest AUC.

Finding 1 (Mechanism identification). Table 1(b) reveals that the language-prior-dominance assumption is *inverted*: the prior is *lower* at object tokens (p_{prior} AUC $= 0.439$), meaning object generation is not driven by language-prior strength. The operative signal is visual-mean alignment (cosim, AUC $= 0.589$)—the only signal identifying object tokens above chance. The visual contribution Δlogit is also not predictive (AUC $= 0.422$ on the same object-vs-background comparison—hallucinated objects are not simply the tokens receiving the largest logit-level visual boost), indicating that the mechanism is a representational alignment rather than a logit-level effect. This directly explains the CD reversal (Eq. 1): CD adds $\alpha \cdot \Delta\mathbf{z}_t$ to the logit, assuming hallucinated tokens have low $\Delta\mathbf{z}$. But since hallucination is driven by cosim (high alignment with the visual mean), CD amplifies the same mechanism causing it. We attribute the inversion to deep-fusion's entanglement of visual and linguistic representations: because visual tokens are injected into the same hidden space as text and processed through shared transformer layers (Chen et al., 2024c; Bai et al., 2025), the language prior is no longer the dominant generative force—visual-mean alignment is.

3.3 Architecture Gating: The Mechanism Is Not Universal

Finding 2 (Architecture-gated mechanism). Table 1(c) reveals the mechanism is *architecture-gated*: active in deep-fusion models (Qwen2.5-VL AUC=0.589, InternVL3.5 AUC=0.618), absent in shallow-projection (LLaVA-1.5 AUC=0.446, protective), and reversed in Qwen3-VL-4B

(AUC=0.390). We attribute this divergence to fusion depth: shallow projection (Liu et al., 2023) keeps visual and linguistic representations separable (language prior dominates, CD effective); deep fusion (Chen et al., 2024c; Bai et al., 2025) entangles them (cosim dominates, CD counterproductive). The AUC gating criterion provides a practical diagnostic: DRIFTDPO applies when AUC > 0.5 (mechanism active). The mechanism also extends to video (NExT-QA, AUC=0.560; Appendix G).

3.4 PER-TOKEN ANALYSIS: THE CAPTION-LEVEL SIGNAL IS BIASED

The caption-level mean treats all tokens equally, but object tokens are only 3.4% of generated tokens; 96.6% are function words. A per-token analysis reveals that the per-token AUC (hallucinated vs. correct) is **0.426—below chance**: correct objects have *higher* cosim than hallucinated ones, *reversed* from the caption-level AUC (0.589), because the mean is dominated by function tokens. The full distribution is in Appendix C (Table 5).

Implication for preference direction. The standard construction (min-cosim = chosen) is correct at the caption level but *incorrect* at the object level: it penalizes correct objects, causing coverage suppression. Reversing the direction—choosing max-cosim captions—should improve object-level quality, confirmed in §5.2 and §5.6.

3.5 CAUSAL VALIDATION

We validate that the mechanism is *causal*, not merely correlational, by intervening on the signal: selecting the lowest-$\overline{\mathrm{cosim}}$ of B=4 candidates (CoSim-Beam) achieves CHAIRi = 0.1592 (-12.0% vs. greedy). Since the mechanism is causal and self-measurable, it can serve as a self-supervised reward—the premise for DRIFTDPO (§4).

4 METHOD

4.1 OVERVIEW

The mechanistic finding in §3—that hallucination on deep-fusion VLMs is driven by visual-mean alignment, an architecture-specific, causal, and self-measurable mechanism—yields DRIFTDPO. The core idea is to use the mechanism signal (cosim) as both the preference label and the implicit training objective, thereby internalizing the alignment-minimization goal into model weights so that standard greedy decoding naturally produces less hallucinated outputs.

DRIFTDPO proceeds in three stages (Figure 1): (1) *preference data construction with direction correction*—for each image, generate K candidate captions, score each by caption-level cosim, determine the preference direction from per-token analysis, and form (chosen, rejected) pairs; (2) *DPO training*—train a LoRA adapter to maximize the log-likelihood margin between chosen and rejected outputs; (3) *inference*—standard greedy decoding with the adapter, requiring no modification to the decoding procedure. The key design principle is that no external supervision is needed at any stage.

4.2 SIGNAL DEFINITION

Let $\mathbf{h}_t^{(l)} \in \mathbb{R}^d$ denote the hidden state at layer l for generation step t. The visual mean is the average hidden state over the $|\mathrm{vis}|$ visual token positions in a prefill-only forward pass, $\bar{\mathbf{v}} = \frac{1}{|\mathrm{vis}|} \sum_{i \in \mathrm{vis}} \mathbf{h}_i^{(l)}$, and the visual-mean alignment at token t is the cosine similarity

$$\mathrm{cosim}(t) = \frac{\mathbf{h}_t^{(l)}}{\|\mathbf{h}_t^{(l)}\|} \cdot \frac{\bar{\mathbf{v}}}{\|\bar{\mathbf{v}}\|} \tag{2}$$

This quantity measures how aligned the generation-step representation is with the average visual representation. As established in §3.2, elevated cosim at object-generation tokens is the operative hallucination driver on deep-fusion models. The caption-level signal averages over all generated tokens, $\overline{\mathrm{cosim}}(s) = \frac{1}{|s|} \sum_{t=1}^{|s|} \mathrm{cosim}(t; s)$, computed in a single teacher-forcing pass over caption s. We use Layer 24 (l=24) throughout, as it achieves the highest probing AUC (§3.2).

4.3 Preference Data Construction with Direction Correction

The goal is to create (chosen, rejected) pairs where the chosen caption has lower hallucination risk than the rejected one. The construction is fully automatic—no human annotation or ground-truth object labels are needed.

Candidate generation. For each image I with prompt p, we sample K=8 candidates $s_k \sim \pi_\theta(\cdot \mid I, p)$ with temperature T=0.8 and top-p=0.9. Sampling rather than beam search (Graves, 2012; Vijayakumar et al., 2016) is necessary here: beam-search candidates are too similar in cosim (mean spread 0.0039) to give a learnable preference signal, whereas sampling yields 4× larger variation (0.015; Appendix D, Figure 2).

Scoring and pair formation. Each candidate is scored by $\overline{\mathrm{cosim}}(s_k)$ in a single teacher-forcing pass; the visual mean $\bar{\mathbf{v}}$ is computed once per image and reused, so scoring costs $O(K)$ passes per image. We form one preference pair per image from the extremes of the cosim distribution and discard pairs whose gap falls below τ=0.005, retaining 99.2% of pairs.

Direction correction. The per-token analysis (§3.4) shows that the caption-level signal is biased—96.6% of tokens are function words and the object-level signal is *reversed* (per-token AUC = 0.426 $<$ 0.5)—so the standard pair formation (min-cosim = chosen) penalizes correct objects, causing coverage suppression. DriftDPO therefore fixes the direction once from the per-token AUC: $y_w = \arg\min_k \overline{\mathrm{cosim}}(s_k)$ if AUC $>$ 0.5 (standard), and $y_w = \arg\max_k \overline{\mathrm{cosim}}(s_k)$ otherwise (reversed; used for Qwen2.5-VL and InternVL3.5, both AUC $<$ 0.5). Appendix F (Algorithm 1) gives the full procedure.

4.4 DPO Training

Reward design: mechanism as implicit self-supervised reward. The engineering corollary from §3 is that the self-measurable mechanism can serve as its own reward. In standard RLHF-based hallucination methods (Yu et al., 2024; Sun et al., 2024; Zhao et al., 2023; Ouyang et al., 2022; Christiano et al., 2017; Schulman et al., 2017), the reward is explicitly defined by human annotators or external models (GPT-4, reward classifiers). In DriftDPO, the reward is *implicitly* defined by the cosim ranking: the preference pair (y_w, y_l) encodes the constraint that y_w should have lower hallucination risk, and DPO optimizes the model to assign higher likelihood to y_w. No explicit reward function is needed because the mechanism signal has been validated as a causal hallucination driver (§3.5).

The reward is thus *self-supervised*, coming from the model's own hidden states and removing the scalability bottleneck of human- or GPT-4-labeled preferences, and *mechanism-grounded*, derived from a validated causal theory rather than a surface-level heuristic such as caption length or object count.

DPO objective. The preference dataset $\mathcal{D}$ is used to train a LoRA adapter via the DPO objective (Rafailov et al., 2023). For a preference pair (x, y_w, y_l), the sequence-level log-probability $\log \pi_\theta(y \mid x) = \sum_t \log \pi_\theta(y_t \mid x, y_{<t})$ is computed under both the policy π_θ (LoRA-enabled) and the reference π_{ref} (LoRA-disabled). The DPO loss (Eq. 3) maximizes the log-likelihood margin between chosen and rejected:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)}\Big[\log \sigma\Big(\beta\big(r_\theta(x, y_w) - r_\theta(x, y_l)\big)\Big)\Big] \tag{3}$$

where $r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ is the implicit reward, β controls the KL penalty strength, and σ is the sigmoid. The model learns to assign higher reward to lower-hallucination outputs, where "lower hallucination" is defined by the cosim ranking rather than external annotation.

LoRA configuration. LoRA (Hu et al., 2022; Dettmers et al., 2023; Liu et al., 2022) is applied with rank r=16, α_{LoRA}=32 to all attention and MLP projections, yielding 37.15M trainable parameters (0.98% of 3.75B). The visual encoder and LM head are frozen, ensuring the mechanism signal

is computed from the same base representations during both data construction and training. Full hyperparameters in Appendix A.

4.5 INFERENCE

At test time DRIFTDPO uses standard greedy decoding with the LoRA adapter loaded—no beam search, no cosim scoring, no extra forward passes: the insight is fully internalized, so greedy decoding naturally produces less hallucinated captions at negligible latency ($<$1%), in contrast to the 4$\times$ teacher-forcing that CoSim-Beam needs.

5 EXPERIMENTS

The experiments address four questions: (Q1) Does internalizing the mechanism signal outperform inference-time interventions? (§5.2) (Q2) Does the AUC gating criterion generalize across architectures? (§5.3) (Q3) Does the improvement transfer to unseen tasks? (§5.4) (Q4) Does training internalize the mechanism, and what governs the training dynamics? (§5.5–§5.7)

5.1 SETUP

Models. To test whether the mechanism is architecture-specific rather than model-specific, we evaluate across five VLMs spanning three architectural families. Qwen2.5-VL-3B-Instruct (Bai et al., 2025) serves as the primary model for mechanism analysis and method development. Qwen2.5-VL-7B tests cross-scale transfer within the same family. InternVL3.5-4B (Chen et al., 2024c;b) tests cross-architecture transfer to a different deep-fusion design. LLaVA-1.5-7B (Liu et al., 2023; 2024c) represents the shallow-projection family where visual features are injected via a lightweight projection layer, serving as the architecture-gating contrast. Qwen3-VL-4B tests a next-generation architecture where the mechanism is reversed.

Benchmarks. We evaluate on four benchmarks covering different hallucination types and task formats. COCO CHAIR (Rohrbach et al., 2018; Lin et al., 2014) measures object hallucination in open-ended captioning on MSCOCO (Lin et al., 2014) val2017, reporting instance-level CHAIRi, sentence-level CHAIRs, and coverage (number of mentioned objects). POPE (Li et al., 2023) evaluates discriminative object recognition via yes/no questions under the adversarial setting, testing whether improvements transfer from generative to discriminative tasks. MMHal-Bench (Sun et al., 2024) assesses hallucination density in open-ended QA. AMBER (Wang et al., 2023) provides cross-benchmark validation on a different image set.

Baselines. We compare with two categories of methods. *Inference-time methods* operate without training: VCD (Leng et al., 2024) and ICD (Wang et al., 2024) apply contrastive decoding by subtracting logits from degraded inputs; M3ID (Liu et al., 2024b) uses multi-modal instruction decoding; OPERA (Huang et al., 2024) modifies beam search to penalize attention sinks; HALC (Chen et al., 2024a) uses adaptive focal-contrast decoding; CoSim-Beam applies our mechanism signal as a beam search scoring criterion. *Training-based methods* (HA-DPO (Zhao et al., 2023), RLHF-V (Yu et al., 2024)) are not directly compared because they require GPT-4 or human-annotated preference data, making fair comparison under identical data conditions infeasible. DRIFTDPO's distinguishing advantage is that it requires no external annotation.

Training. We train a LoRA adapter (Hu et al., 2022) (rank 16, α_{LoRA}=32) on all attention and MLP projections of Qwen2.5-VL-3B (37.15M trainable parameters, 0.98% of 3.75B) with DPO (Rafailov et al., 2023) (β=0.1, 2 epochs, batch size 8; $\sim$2 hours on one L40s). Full hyperparameters in Appendix A.

5.2 CAN INTERNALIZATION OUTPERFORM INFERENCE-TIME INTERVENTION?

Table 2 separates methods into three groups. CD methods (VCD, ICD, M3ID) increase hallucination by +13–25%, confirming the CD reversal (§3.1): they subtract $\Delta\mathbf{z}$, but on deep-fusion models hallucinated objects are not the tokens with the largest $\Delta\mathbf{z}$ (Table 1(b)), so CD does not suppress

Table 2: COCO CHAIR on Qwen2.5-VL-3B. DRIFTDPO achieves −22.2% CHAIRi at 1× cost while *increasing* coverage (+29%)—the only method to do so.

Method	CHAIRi↓	ΔCHAIRi	CHAIRs↓	Coverage↑	Inference
Greedy	0.1810	—	0.1067	6.7	1×
VCD (Leng et al., 2024)	0.2097	+15.9%	0.1010	6.3	2×
ICD (Wang et al., 2024)	0.2051	+13.3%	0.0915	5.9	2×
M3ID (Liu et al., 2024b)	0.2269	+25.4%	0.1137	6.6	2×
OPERA (Huang et al., 2024)	0.1602	−11.5%	0.1009	7.1	4×
HALC (Chen et al., 2024a)	0.1584	−12.5%	0.0858	6.6	2×
CoSim-Beam	0.1592	−12.0%	0.0990	6.9	4×
DRIFTDPO (w/o dir. corr.)	0.1597	−11.8%	**0.0629**	4.2	1×
DRIFTDPO	**0.1408**	**−22.2%**	0.0939	**8.66**	**1**×

them. Inference-time methods (OPERA, HALC, CoSim-Beam) reduce hallucination by 11–12.5% at 2–4× overhead, but all reduce or barely maintain coverage (−6 to +1%). They reduce hallucination by generating fewer objects. DRIFTDPO breaks this pattern: −22.2% CHAIRi at 1× cost, with coverage increasing +29% (6.7→8.66). It reduces hallucination by generating *more* objects with lower per-object error rate (18.1%→14.1%).

The w/o dir. corr. variant isolates direction correction: −11.8% CHAIRi but coverage collapses (−38%), the same pattern as inference-time methods. Without correction, the caption-level cosim is dominated by function words (96.6%), so the model learns to suppress objects. Under full training this collapses to coverage=0 (§5.6).

5.3 DOES THE AUC GATING CRITERION GENERALIZE ACROSS ARCHITECTURES?

If the mechanism is architecture-gated, the cosim AUC should predict both CD failure and DRIFTDPO applicability before either is applied. Appendix J confirms this across five VLMs: all four deep-fusion models (AUC > 0.5) worsen under CD, while LLaVA-1.5 (AUC $= 0.446$) improves (−7.5%). InternVL3.5-4B provides the strongest validation—its per-token AUC (0.461) predicts the direction reversal that yields −47.4% under correction, whereas the standard direction increases hallucination by +18.7%—confirming the two-stage diagnostic: the caption-level AUC decides *whether* to apply DRIFTDPO, the per-token AUC decides *which direction*.

5.4 DOES THE IMPROVEMENT TRANSFER TO UNSEEN TASKS?

To test whether DRIFTDPO overfits to COCO objects, we evaluate on two unseen tasks. On POPE—a discriminative yes/no task—accuracy improves by +3.4pp and F1 by +3.6pp (3B), while the uncorrected direction variant even degrades F1 (−0.5pp) (Table 7, Appendix H). On MMHal-Bench—open-ended QA—hallucination density drops −23.7% (3B) and −13.3% (7B) (Table 8, Appendix I). The transfer from generative training to discriminative evaluation indicates cosim captures general visual grounding, not COCO-specific recognition.

AMBER serves as a negative control for direction correction. Without correction, DRIFTDPO worsens CHAIR_i by +17.0% due to coverage collapse—the same failure as on COCO. With correction, it improves by −5.2% while maintaining coverage. This confirms direction correction addresses a general signal bias, not a dataset artifact. Table 4 confirms zero inference overhead (0.97×).

5.5 DOES TRAINING INTERNALIZE OR SUPPRESS THE MECHANISM?

Training could either *suppress* the mechanism (avoiding high-cosim outputs while the underlying states persist) or *internalize* it (shifting the distribution so the signal becomes redundant). Suppression would preserve cosim separability between correct and hallucinated tokens; internalization would eliminate it.

Figure 3 supports internalization. Before training, correct and hallucinated tokens are separable (AUC=0.440, d=−0.234). After training, distributions converge (AUC→0.510, d→−0.009): the model no longer produces high-cosim hallucinated tokens. This convergence coincides with cov-

erage increasing (+29%) and per-object hallucination decreasing (-22.2%), ruling out suppression (which would reduce coverage).

Because cosim is computed from the model's own hidden states, DPO training modifies the states the reward depends on, so the signal self-erodes—a coupling that external rewards (Ouyang et al., 2022; Christiano et al., 2017), whose reward function is fixed, do not have—giving self-supervised mechanism rewards a natural stopping point (AUC reaching 0.5).

5.6 Why Does the Standard Direction Fail?

The per-token analysis (§3.4) predicts the standard direction (min-cosim = chosen) should fail: it penalizes correct objects (which have higher cosim), incentivizing the model to avoid objects entirely. This is counterintuitive—standard DPO with external rewards never produces coverage=0—so it tests the mechanism theory.

The ablation confirms this (Figure 4). Under identical data and hyperparameters, the standard direction achieves higher val_acc (0.953 vs. 0.653) but CHAIRi=0 with coverage=0: the model outputs no objects. This is correct optimization of the wrong objective: minimizing the function-word-dominated caption-level cosim rewards object-free captions, whereas the reversed direction achieves both low hallucination (0.1408) and high coverage (8.66).

This shows that for self-supervised rewards from aggregate statistics, the aggregation level must match the target phenomenon. Caption-level aggregation obscures the object-level signal; per-token analysis is needed to diagnose and correct this.

5.7 Why Does More Data Hurt?

RLHF scaling laws (Gao et al., 2023) predict monotonic improvement with more data, but DriftDPO violates this: its best CHAIRi comes from a small fraction of the constructed pairs, while training on the full set barely improves over greedy (-0.6%). The degradation is monotonic and validation accuracy *increases* (88%→97.4%), so it is not conventional overfitting. The cause is the coupling of §5.5: with more pairs the DPO update reshapes the hidden states from which cosim is computed, so the reward drifts away from the states it was measured on and the model converges to shorter, object-sparse captions. Appendix L reports the full data-scaling and β ablation, and Appendix M shows that DriftDPO and CoSim-Beam are complementary (distribution vs. selection).

6 Conclusion

This work shows that hallucination in VLMs is not governed by a universal cause: the decade-long default—language-prior dominance—is architecture-specific and, on deep-fusion models, *inverted*—the operative driver is visual-mean alignment. Effective mitigation must therefore begin by diagnosing the mechanism of the target architecture. DriftDPO shows that this self-measurable signal can serve as a fully self-supervised reward, achieving SOTA hallucination reduction at zero inference overhead while increasing coverage. Its non-monotonic scaling reveals a coupling between self-supervised rewards and the model states that define them, a characteristic absent in externally-supervised rewards; as architectures evolve, the diagnosis must be re-derived.

Limitations. Our study bounds the scope of hallucination mitigation rather than settling it. The diagnosis is empirical: we identify visual-mean drift by probing and validate it causally through intervention, but we do not derive it from first principles, so *why* deep fusion gives rise to this drift remains an open theoretical question. The self-supervised reward inherits the limits of the signal it is built on—it is validated for the architectures and attribute space studied here, and other internal signals may dominate elsewhere. Our evidence is also concentrated on object-centric, closed-vocabulary evaluation, with limited open-vocabulary coverage. Finally, the coupling we observe between the reward and the model states suggests that self-supervised mechanism rewards call for their own stopping criteria rather than standard data-scaling recipes; what transfers is the methodology—diagnose the operative mechanism, turn it into a reward, internalize it—not a universally valid signal.

AI USE STATEMENT

We used generative AI tools (GLM) only for English polishing and LaTeX/bibliography formatting, not for any task requiring disclosure under the ICLR 2027 AI Policy for Authors, and we take full responsibility for the final content.

REFERENCES

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.

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.

Yonatan Belinkov. *Probing Neural Language Models*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *ICML*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. In *ACL Workshop BlackboxNLP*, 2019.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *ICML*, 2023.

Alex Graves. Sequence transduction with recurrent neural networks. In *ICML*, 2012.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *NAACL*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Qidong Huang, Xiao Liu, Yuwen Zhang, Tianze Ren, Zihao Zhao, Yao Xu, Zhiyang Li, Yifan Zhu, Lifu Tu, Tianyu Shi, Sishuo Yan, Yifan Liu, Junjie Hu, Wei Lin, Wanli Ouyang, and Nan Dang. OPERA: Over-trust problem and mitigation in vision-language models. In *CVPR*, 2024.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, 2025.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. THRONE: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *CVPR*, 2024.

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 *CVPR*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *ICML*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dian Chao, Yucong Zhou, Wei Huang, Bin Xiao, Libo Long, Henglei Lv, Shujin Lin, Xinyu Tian, Yong Guo, and Shaohui Lin. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Hao Liu, Wilson Yan, and Pieter Abbeel. M3ID: Multi-modal in-context instruction decoding for vision-language models. In *ICLR*, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Genglin Liu, Jianfeng Gao, and Yong Jae Lee. LLaVA-1.5: Improved baselines with visual instruction tuning. In *CVPR*, 2024c.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *ACL*, 2022.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

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. In *NeurIPS*, 2023.

Vipula Rawte, Amit Sheth, and Apala Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*, 2024.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you know about BERT? An analysis of contextual and structural information. In *EMNLP*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

Zhaoyang Yu, Shaohan Huang, Xiaozhen Ren, Zhen Hou, Haizhen Huang, Weiwei Deng, Feng Zhao, Daoyuan Chen, Yiru Li, Shiqi Guo, Fangyu Zou, Junfeng Tian, Yuchi Liu, Yingshui Tan, Jianwei Zhang, Jianxin Sun, Yuxuan Cai, Baosa Liang, Rui Liu, Junfeng Yao, Zhenxiang Huang, Yulong Li, Lin Ma, and Wei Zhao. RLHF-V: Towards fine-grained alignment with natural language feedback. In *CVPR*, 2024.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. *arXiv preprint arXiv:2402.08680*, 2024.

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.

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. In *ICLR*, 2024.