SEEING WHAT MATTERS: VISUAL-SELECTIVE ON-POLICY SELF-DISTILLATION FOR VLM HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) still suffer from object hallucination in open-ended generation. They often mention objects, attributes, or relations that are not grounded in the image. Preference-based methods such as DPO can reduce hallucinations, but open-ended caption preferences are expensive to collect and often entangle visual faithfulness with style, length, and level of detail. Meanwhile, directly turning object cues into hard pseudo captions can amplify the teacher's linguistic bias and hallucinations. We study how to extend on-policy self-distillation from verifiable language tasks to open-ended multimodal hallucination reduction. Instead of using privileged hints that directly verify an answer, we use training-time object cues to construct a non-verifying privileged visual teacher and distill its soft token distributions on responses sampled by the student itself. We further propose Visual-Selective On-Policy Self-Distillation (VS-OPSD). By contrasting the privileged teacher under the real image and a weak-image condition, VS-OPSD identifies visually sensitive tokens and applies distillation only at those positions. Experiments across two model families and two LLaVA model scales show that VS-OPSD reduces hallucination without preference pairs while largely preserving general multimodal capability. Controlled ablations attribute the gains to concentrating supervision on visually sensitive tokens rather than simply reducing the number of distilled tokens.

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.

1 INTRODUCTION

Large vision-language models (LVLMs) have demonstrated strong visual understanding and reasoning capabilities across tasks such as image captioning, visual question answering, and object localization (Liu et al., 2023; 2024a; Dai et al., 2023; Li et al., 2023a; Chen et al., 2023). However, visual hallucination remains a persistent challenge. Models may generate objects that do not exist in an image or incorrectly describe object attributes and relations (Li et al., 2023b; Wang et al., 2023). In open-ended image captioning, this problem reflects insufficient visual grounding in the generated content. Meanwhile, the same image can correspond to multiple semantically correct descriptions with different expressions. Open-ended generation therefore typically lacks a unique reference answer and is difficult to supervise directly through answer matching or correctness verifiers. This paper therefore investigates how to provide reliable and effective visual supervision for open-ended multimodal generation without specifying a unique target response in advance.

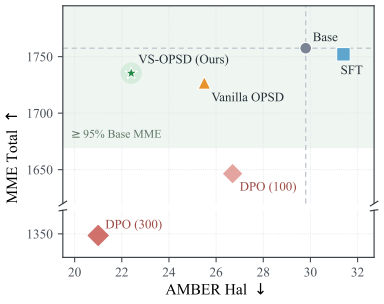

Figure 1: Overall performance overview. AMBER Hal and MME summarize hallucination and general capability. By distilling only 30% of response tokens, VS-OPSD reduces hallucination while retaining 98.7% of Base MME. DPO uses external RLHF-V preference data (Yu et al., 2024).

An important technical direction is to improve model visual faithfulness through post-training methods such as supervised fine-tuning (SFT) (Liu et al., 2024a; Dai et al., 2023) and preference optimization (DPO) (Rafailov et al., 2023; Zhao et al., 2025; Wang et al., 2024). SFT aligns a model with desired behaviors by learning from demonstration responses and has become a standard post-training approach. However, in open-ended multimodal generation, demonstrations cover only a limited subset of the valid output space and cannot fully characterize the diversity of visually faithful responses. DPO adjusts model output preferences through *chosen*/*rejected* preference pairs. Although it can align models with human preferences more effectively, preference judgments may be contaminated by noise from non-visual factors such as response length and style (Park et al., 2024; Wang et al., 2024). In addition, these methods are sensitive to data quality, while constructing high-quality data is expensive (Yu et al., 2024). Meanwhile, fixed off-policy responses can create a mismatch between training prefixes and the states visited by the model at inference time (Agarwal et al., 2024; Gu et al., 2024).

Online self-distillation methods provide an on-policy perspective. They operate on model rollout responses and perform token-level distillation using either an external teacher model (OPD) (Agarwal et al., 2024) or the same model conditioned on privileged information (OPSD) (Zhao et al., 2026). However, OPD requires a separate, often larger teacher and therefore increases the training burden (Agarwal et al., 2024; Zhao et al., 2026). The original OPSD formulation, meanwhile, focuses on verifiable mathematical reasoning, where verified solutions provide a clearly directed corrective signal (Zhao et al., 2026). In contrast, object cues in open-ended multimodal generation provide only partial visual facts, as they neither specify a complete description nor directly verify whether an answer is correct. Furthermore, standard OPSD typically distills all response tokens uniformly, even though not every token carries the same amount of visual information. Some noisy tokens may dilute the limited visual supervision and introduce distribution shifts unrelated to visual grounding.

To address these problems, we first use object cues to construct a non-verifying privileged teacher for open-ended multimodal generation and perform self-distillation on responses generated by the student itself. Building on this formulation, we propose **Visual-Selective On-Policy Self-Distillation (VS-OPSD)**. Specifically, we introduce a contrastive teacher that receives a weak image together with the same object cues, and estimate visual sensitivity through the token-level distribution discrepancy between the real-image teacher and the weak-image teacher. VS-OPSD selects the tokens with the highest visual sensitivity and applies sensitivity-weighted distillation, thereby concentrating supervision on positions where real visual evidence substantially changes the teacher's predictions.

We validate our method on LLaVA-1.5 at both 7B and 13B scales and on Qwen2.5-VL-3B. Under the matched COCO-500 setting, VS-OPSD distills only approximately 30% of the response tokens. On LLaVA-1.5-7B, it reduces AMBER Hal from 25.5 with Vanilla OPSD to 22.4 and lowers CHAIRs/CHAIRi on a separate COCO-val CHAIR evaluation from $49.6/15.5$ to $43.2/12.6$. Randomly selecting the same token budget or uniformly weighting the selected visually sensitive tokens does not yield comparable gains. The improvements also persist across weak-image constructions, which produce consistent token rankings and hallucination-reduction trends. Meanwhile, VS-OPSD retains 98.7% of the base model's MME score and remains close to the base model on ScienceQA and POPE. On Qwen2.5-VL-3B, it reduces AMBER Hal from 26.9 to 25.0, while on LLaVA-1.5-13B it reduces Hal from 23.8 to 21.1. These results provide evidence that the visual-selection principle transfers across the evaluated model family and scale settings.

Our main contributions are summarized as follows:

- **Open-ended privileged supervision.** We extend OPSD to open-ended multimodal hallucination mitigation and use object cues to construct a privileged visual teacher that neither specifies a complete answer nor directly verifies correctness.
- **Token-level visual selection.** We reveal that privileged visual supervision is highly non-uniform across response tokens and propose VS-OPSD, which identifies visually sensitive tokens through a contrast between real-image and weak-image teachers and selectively distills these tokens.
- **Systematic empirical analysis.** Under matched settings, VS-OPSD achieves better hallucination metrics than Vanilla OPSD using approximately 30% of the distillation tokens. Ablations, CHAIR evaluation on COCO-val, and general-capability evaluations further validate the role of visual selection and characterize the boundaries of the method.

2 PRELIMINARIES

2.1 FROM OFFLINE DISTILLATION TO OPSD

Let I denote an image, q a task prompt, and h privileged information available only during training. We use p_θ for the trainable student policy and p_0 for a frozen teacher initialized from the same pre-trained VLM. Conventional offline distillation (Hinton et al., 2015) matches their token distributions on the prefixes of a fixed response y:

$$\mathcal{L}_{\text{off}}(\theta) = \mathbb{E}_{(I,q,h,y)\sim\mathcal{D}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D(p_\theta(\cdot \mid I, q, y_{<t}) \,\|\, \text{sg}[p_0(\cdot \mid I, q, h, y_{<t})]) \right]. \tag{1}$$

Because y is supplied by a fixed dataset or an external teacher, whereas the model conditions on its own generated tokens at inference time, training and inference may visit different autoregressive states. OPSD (Zhao et al., 2026) instead samples a response from the current student policy,

$$\hat{y} \sim p_\theta(\cdot \mid I, q), \tag{2}$$

and asks the teacher to provide soft distributional supervision along the same student trajectory:

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{\substack{(I,q,h)\sim\mathcal{D} \\ \hat{y}\sim p_\theta(\cdot|I,q)}} \left[\frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} D(p_\theta(\cdot \mid I, q, \hat{y}_{<t}) \,\|\, \text{sg}[p_0(\cdot \mid I, q, h, \hat{y}_{<t})]) \right]. \tag{3}$$

Here, D is a token-level distribution divergence and sg denotes stop-gradient. OPSD does not require the teacher to generate a new target response. Instead, it provides dense soft supervision on prefixes actually visited by the student, thereby reducing the state-distribution mismatch induced by fixed target sequences. We instantiate D as the generalized Jensen–Shannon divergence (Lin, 1991).

2.2 WHY VISUAL-SELECTIVE DISTILLATION?

Applying OPSD directly to open-ended visual generation raises two key questions. ***Q1:*** *Without a verifiable answer, what should guide the teacher's correction?* ***Q2:*** *Even when visual supervision is available, which positions in the response should be corrected?*

How can we provide supervision without a verifiable answer? Existing OPSD-based methods (Zhao et al., 2026) primarily target verifiable tasks such as mathematical reasoning, where correct answers, reasoning chains, or verifier feedback specify a clear direction for correction. Open-ended image captioning, however, has no unique correct realization, while object annotations provide only partial visual facts that indicate which entities are present but do not determine the wording, length, or the attributes, actions, and relations. These cues can therefore neither serve as complete hard targets nor directly determine whether a student response is correct, limiting the direct transfer of OPSD based on verifiable privileged information to open-ended multimodal hallucination mitigation.

Which tokens require visual correction? Multimodal generation imposes an additional grounding requirement beyond language modeling, since generated content must remain consistent with the visual input. However, different response positions depend on visual evidence to different degrees. Content words describing entities, attributes, counts, actions, and spatial relations are often strongly grounded in the image, while function words and syntactic tokens rely more heavily on linguistic context. Vanilla OPSD distills all response positions uniformly and may therefore spread limited visual supervision over tokens that contribute little to visual grounding.

Exploratory diagnostic. To test whether visual correction is uniformly distributed, we run a token-level diagnostic on LLaVA-1.5-7B using 200 COCO-val images. We keep the teacher, object cues, and student-generated prefixes fixed, replace only the teacher's real-image input with a weak image, and compare the resulting next-token distributions. We define their forward KL divergence as *Visual-KL*, as formalized in Eq. 6. A larger value indicates that the teacher's prediction at that position depends more strongly on real visual evidence. As shown in Figure 2, across 200

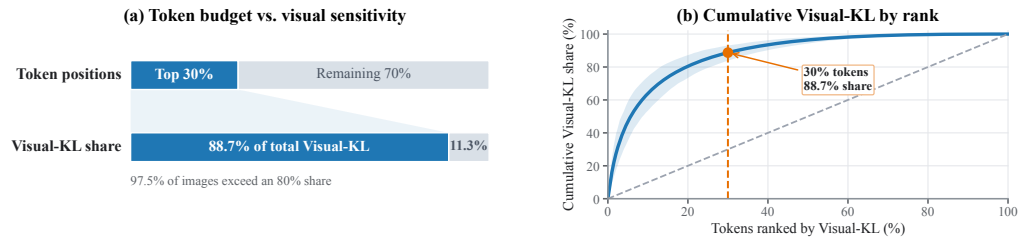

Figure 2: **A small token budget captures most visual signal.** (a) The top 30% of response tokens account for 88.7% of Visual-KL mass on average. (b) Mean cumulative Visual-KL by token rank, with the 10th–90th percentile band across 200 images.

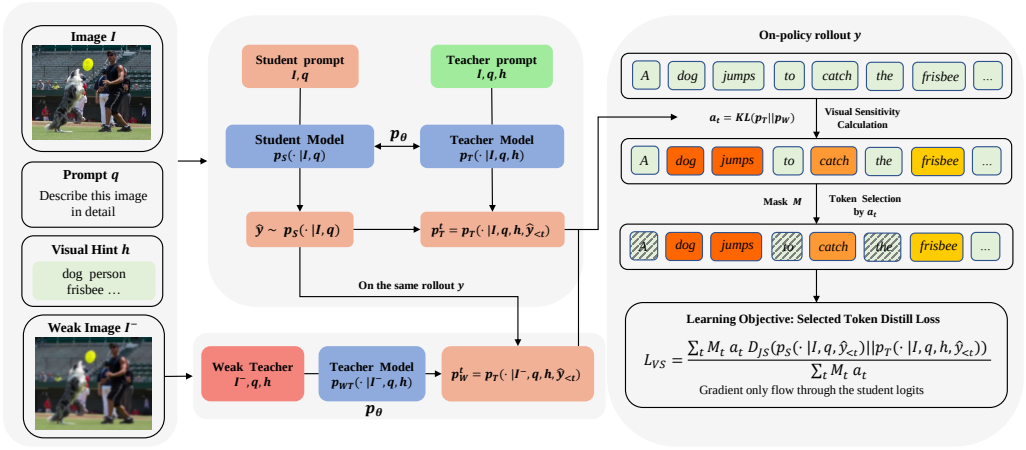

Figure 3: **Overview of VS-OPSD.** The student generates an on-policy response, while the privileged and weak-image teachers evaluate the same trajectory. Their discrepancy estimates token-level visual sensitivity, and distillation is applied only to selected visually sensitive tokens. Object cues and the weak-image branch are used only during training.

COCO images, the top 30% of tokens ranked by Visual-KL account for **88.7%** of the visual distribution shift on average, and this share exceeds 80% for 97.5% of the samples. This concentration suggests that distillation signals should be allocated selectively along on-policy trajectories rather than uniformly across all response tokens.

3 Method

Figure 3 summarizes VS-OPSD in three stages. First, we use object cues to construct training-time privileged context h, enabling the privileged teacher to provide soft supervision along student-generated on-policy trajectories. Second, we calculate the proposed Visual-KL, which measures each token's visual dependence by comparing the teacher distributions under the real and weak images. Finally, we use Visual-KL to select visually sensitive tokens, applying sensitivity-weighted distillation at these positions. Object cues and the weak-image branch are used only during training.

3.1 Non-Verifying Privileged Visual Teacher

Constructing privileged context. For each training image, we extract deduplicated object categories from object-level annotations to form

$$h(I) = \{c_1, \ldots, c_K\}. \tag{4}$$

In our COCO implementation, we remove invalid or very small instances, rank categories by object area, and retain at most $K_{\max}$ categories. The cue set contains neither a reference caption nor

attributes, relations, or a complete response. It supplies partial visual facts without specifying a target sequence. The student receives the standard prompt "Describe this image in detail." The privileged teacher prepends the training-only cue: "Given that the image contains the following objects: $c_1, \ldots, c_K$, describe this image in detail."

Matching distributions on the same student trajectory. Following Eq. 2, the student first samples $\hat{y} \sim p_\theta(\cdot \mid I, q)$. At each prefix $\hat{y}_{<t}$, the student and privileged teacher distributions are

$$\begin{aligned} p_S^t &= p_\theta(\cdot \mid I, q, \hat{y}_{<t}), \\ p_T^t &= p_0(\cdot \mid I, q, h, \hat{y}_{<t}). \end{aligned} \tag{5}$$

Only the student parameters are updated. The teacher does not decode an independent pseudo-caption. It only re-evaluates states the student has already visited. Consequently, the object cues modify a soft distribution on the student's own trajectory rather than collapsing open-ended generation into a single hard target. Compared with SFT on cue-generated captions, this design preserves the student's current wording and state distribution while injecting additional visual facts through h. At inference time, only the student branch is retained, and h is removed.

3.2 Token-Level Visual Sensitivity via Controlled Intervention

Isolating the effect of real visual evidence. Privileged information does not provide equally useful visual correction at every response position. We therefore apply a strong Gaussian blur to construct a weak image I^-. This preserves the global layout of the original image while suppressing edges, textures, and fine-grained object evidence. The weak teacher shares the frozen parameters, prompt, object cues, and student prefix of the privileged teacher. Only its visual input differs:

$$p_W^t = p_0(\cdot \mid I^-, q, h, \hat{y}_{<t}), \qquad a_t = \mathrm{KL}\big(p_T^t \big\| p_W^t\big)\,. \tag{6}$$

This comparison is a controlled visual intervention: because the two forward passes differ only in I versus I^-, a_t measures how strongly restoring real visual evidence changes the teacher's prediction. A high a_t indicates greater dependence on the image, whereas a low value indicates that object cues, language priors, or syntactic context largely determine the prediction. Unlike generic teacher–student divergence, this score does not mix visual dependence with student capability or policy shift. It measures visual sensitivity, not token correctness.

3.3 Visual-Selective Distillation

Concentrating the distillation budget. We rank the response positions by $\{a_t\}_{t=1}^{|\hat{y}|}$ and retain the top-ρ fraction. Let τ_ρ be the corresponding threshold. The selection mask and token weight are

$$M_t = \mathbf{1}\{a_t \geq \tau_\rho\}\,, \qquad \omega_t = M_t a_t. \tag{7}$$

Here, $\mathbf{1}\{\cdot\}$ denotes the indicator function. This allocation has two levels: M_t performs hard selection and determines which positions participate in distillation, while a_t softly determines the relative correction strength within the selected set. The resulting objective is

$$\mathcal{L}_{\text{VS-OPSD}} = \frac{\sum_{t=1}^{|\hat{y}|} \omega_t D_{\mathrm{JS}}(p_S^t \| \mathrm{sg}(p_T^t))}{\sum_{t=1}^{|\hat{y}|} \omega_t + \epsilon}, \tag{8}$$

where D_{JS} denotes the Jensen–Shannon divergence with an equal-weight mixture, and ϵ ensures numerical stability. The weak-image teacher determines *where* and *how* strongly to distill. The actual soft target remains the privileged teacher distribution p_T^t under the real image. Thus, VS-OPSD does not learn the weak-image distribution. If $\rho = 1$ and a_t is replaced by a constant, Eq. 8 reduces to Vanilla OPSD.

In summary, VS-OPSD preserves the on-policy trajectory and soft supervision of OPSD while reallocating token-level correction through a controlled visual intervention. It requires no external teacher, gold answer, verifier, preference pair, or rejected response. The additional teacher branch is used only during training and therefore does not alter inference-time parameters or computation.

4 EXPERIMENTS

We organize the experiments along three dimensions: the effectiveness of non-verifying object cues for open-ended generation, the contribution of visual token selection, and robustness to weak-image construction, sampling randomness, and cue noise. We further analyze the token-level signal captured by Visual-KL. Additional experiments are provided in the appendix.

4.1 EXPERIMENTAL SETUP

Models, data, and baselines. We primarily evaluate VS-OPSD on LLaVA-1.5-7B (Liu et al., 2023), use Qwen2.5-VL-3B (Team, 2025) for cross-architecture validation, and LLaVA-1.5-13B for validation at a larger model scale. Matched experiments use 500 images sampled from COCO train2014 (Lin et al., 2014). For each image, we extract at most five object categories from instance annotations as training-only privileged cues. No reference caption is provided. We compare SFT, Vanilla OPSD, and VS-OPSD using the same images and cues. DPO and prior hallucination-mitigation methods are included only as external references because their training data and protocols differ.

Training details. We use 4-bit LoRA (Hu et al., 2021; Dettmers et al., 2023) with rank 64 and scaling factor 128. Training uses a per-device batch size of 1, 8 gradient- accumulation steps, and a learning rate of 5×10^{-6}. Student responses are sampled with temperature 0.7, top-$p = 0.9$, and at most 128 new tokens. Unless stated otherwise, all matched methods are trained for 100 steps on one NVIDIA RTX 4090. VS-OPSD uses Gaussian blur with $\sigma = 10$, forward KL for visual sensitivity, a top-30% token budget, and equal-weight Jensen–Shannon divergence for policy distillation.

Evaluation. We evaluate open-ended object hallucination on AMBER (Wang et al., 2023) and CHAIR (Rohrbach et al., 2018), discriminative object hallucination on POPE (Li et al., 2023b), and general multimodal capability on MME (Fu et al., 2026) and ScienceQA (Lu et al., 2022). Object cues are never available during evaluation, and all matched methods use the same decoding configuration. CHAIR generation uses at most 128 new tokens by default. Additional preprocessing, external-reference sources, and evaluation details are provided in Appendix A.4.

4.2 MAIN RESULTS

Table 1 presents the main results. Key findings are as follows.

(1) Visual selection consistently improves Vanilla OPSD. On LLaVA-1.5-7B, VS-OPSD reduces AMBER Hal from 25.5 to 22.4 over Vanilla OPSD and reduces CHAIRs/CHAIRi from 49.6/15.5 to 43.2/12.6 while retaining the same recall of 75.3. On Qwen, it reduces Hal from 26.9 to 25.0 and Cog from 1.8 to 1.5. On LLaVA-1.5-13B, it further reduces Hal from 23.8 to 21.1 and CHAIRs/CHAIRi from 43.2/13.1 to 41.4/12.4 while preserving object recall. The consistent improvements across model families and scales indicate that concentrating distillation on visually sensitive positions is more effective than uniformly correcting the entire response.

(2) On-policy soft supervision is more effective than fixed pseudo-captions. On the two backbones with completed SFT controls, Vanilla OPSD yields substantially lower AMBER Hal than SFT, decreasing it from 31.4 to 25.5 on LLaVA-1.5-7B and from 32.2 to 26.9 on Qwen2.5-VL-3B. Notably, although SFT increases LLaVA Cover from 46.6 to 49.6, it also raises Hal from 29.8 to 31.4. Under open-ended response ambiguity, object cues are therefore more effectively transferred through soft distributions along student-generated trajectories than converted into fixed token-level targets.

(3) Hallucination reduction does not require substantial general-capability degradation. LLaVA VS-OPSD maintains a POPE score of 84.83 and retains 98.7% of the base model's MME score, while ScienceQA changes only from 63.66 to 63.38. MME and ScienceQA also do not decrease on Qwen. These results require only training-time COCO object cues, without external preference pairs or complete target responses. As a non-directly-comparable reference, DPO trained for 300 steps obtains lower Hal but exhibits pronounced drops on POPE and MME, illustrating the trade-off between stronger hallucination suppression and general-capability preservation.

Table 1: Main results on hallucination mitigation and general multimodal capability. We report AMBER, CHAIR, POPE, MME, and ScienceQA, where MME is the sum of its perception and reasoning scores. Each method is reported with its best checkpoint. The best result in each panel is shown in bold, and the second-best result is underlined. For DPO trained for 300 steps, muted red highlights the degradation in general capability relative to the referenced base model. External-reference results use different training and evaluation settings and are not directly comparable.

	AMBER (Generative)				CHAIR		POPE	General Capability			
Backbone / method	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	CHAIRs ↓	CHAIRi ↓	F1 ↑	MME-P ↑	MME-R ↑	MME ↑	SQA ↑
Referenced Results (Not Directly Comparable)											
LLaVA-1.5-7B (Liu et al., 2023)	7.1	46.6	29.8	3.3	52.6	16.4	84.82	1476.03	281.43	1757.46	63.66
+ EOS (Yue et al., 2024)	5.1	49.1	22.7	2.0	40.3	17.8	–	–	–	–	–
+ DPO (100 steps)	5.5	51.8	26.7	2.3	50.2	14.2	82.41	1386.81	259.64	1646.45	63.81
+ DPO (300 steps)	3.4	55.0	21.0	1.3	35.3	8.8	74.29	1114.17	234.29	1348.46	62.61
+ HA-DPO (Zhao et al., 2025)	6.7	49.8	30.9	3.3	38.2	11.0	86.90	1502.58	313.93	1816.51	–
+ HALVA (Sarkar et al., 2025)	6.6	53.0	32.2	3.4	41.4	11.7	84.80	1527.0	–	–	–
3B Multimodal LLMs											
Qwen2.5-VL-3B (Team, 2025)	6.4	59.2	38.9	2.7	40.4	8.5	86.34	1540.02	611.43	2151.44	79.45
+ SFT	5.7	**62.0**	32.2	2.1	32.8	8.0	**86.75**	1528.43	619.64	2148.07	79.45
+ Vanilla OPSD	4.8	61.7	26.9	1.8	**31.2**	8.2	86.21	**1557.37**	627.86	**2185.22**	79.59
+ VS-OPSD	**4.7**	61.0	**25.0**	**1.5**	32.4	**7.4**	86.04	1547.50	**630.71**	2178.21	**79.92**
7B Multimodal LLMs											
LLaVA-1.5-7B (Liu et al., 2023)	7.1	46.6	29.8	3.3	52.6	16.4	84.82	**1476.03**	281.43	**1757.46**	63.66
+ SFT	7.3	**49.6**	31.4	3.7	57.8	18.0	**85.04**	1469.56	282.50	1752.06	**64.28**
+ Vanilla OPSD	5.7	48.2	25.5	2.6	49.6	15.5	84.84	1443.24	**283.57**	1726.81	63.95
+ VS-OPSD	**5.0**	47.7	**22.4**	**2.3**	**43.2**	**12.6**	84.83	1466.06	269.29	1735.34	63.38
13B Multimodal LLMs											
LLaVA-1.5-13B (Liu et al., 2023)	6.2	**50.8**	29.8	3.1	50.2	14.8	85.53	1090.52	247.86	1338.38	65.47
+ SFT	5.0	48.1	23.5	2.4	**39.0**	**11.5**	**85.89**	1097.28	247.86	1345.14	**66.19**
+ Vanilla OPSD	4.9	49.1	23.8	2.2	43.2	13.1	85.80	**1109.07**	248.57	**1357.64**	65.62
+ VS-OPSD	**4.4**	49.0	**21.1**	**2.0**	41.4	12.4	85.88	1087.70	**251.79**	1339.49	65.86

Table 2: **Core design ablations on COCO-500.** Random-30 selects 30% of response tokens at random. Top-30 (uniform) selects the 30% highest-Visual-KL tokens with equal weights. T–S KL denotes ordinary teacher–student disagreement, and top/bottom select the highest-/lowest-scoring positions. Panel (a) isolates selection and weighting, panel (b) compares ranking signals, and panel (c) varies the token budget.

Method	Selection score	Selected tokens	Weighting	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
(a) Selection and weighting							
Vanilla OPSD	–	All	Uniform	6.0	**48.8**	25.4	2.7
Random-30	–	Random 30%	Uniform	6.3	48.4	27.3	3.0
Top-30 (uniform)	Visual-KL	Top 30%	Uniform	5.8	48.3	25.1	2.6
VS-OPSD	**Visual-KL**	**Top 30%**	**Visual-KL score**	**5.0**	47.7	**22.4**	**2.3**
(b) Selection signal: visual dependence vs. teacher–student KL							
T–S KL bottom-30	T–S KL	Bottom 30%	Uniform	6.9	48.2	29.1	3.2
T–S KL top-30	T–S KL	Top 30%	Uniform	5.9	48.7	26.9	3.0
Visual-KL bottom-30	Visual-KL	Bottom 30%	Uniform	6.2	**48.8**	27.7	2.9
Visual-KL top-30	Visual-KL	Top 30%	Uniform	**5.8**	48.7	**25.4**	**2.8**
(c) Token budget							
VS-OPSD (10%)	Visual-KL	Top 10%	Visual-KL score	5.4	**47.8**	23.3	2.6
VS-OPSD (50%)	Visual-KL	Top 50%	Visual-KL score	5.3	47.7	23.0	**2.3**
VS-OPSD (30%)	**Visual-KL**	**Top 30%**	**Visual-KL score**	**5.0**	47.7	**22.4**	**2.3**

4.3 Ablation Studies

Q1. Source of gains

Does VS-OPSD gain from visual dependence rather than sparsity?

Table 2(a): Visual-KL selection and weighting both contribute. Random-30 increases Hal from 25.4 with Vanilla OPSD to 27.3, showing that the gain does not arise from sparsification alone. On the same Visual-KL top-30% token set, sensitivity weighting reduces Hal from 25.1 to 22.4, indicating that Visual-KL both locates correction positions and effectively allocates supervision strength.

Table 3: **Robustness to visual intervention, sampling, and cue noise.** Panel (a) changes only the weak-image construction. None denotes Vanilla OPSD without weak-image selection. Panel (b) reports mean ± standard deviation over three paired seeds. Panel (c) replaces COCO ground-truth cues with Grounding DINO detections (Liu et al., 2024b). Panels (a, b) report step 50 and panel (c) reports step 100 under the original black-image condition. Bold entries in panel (c) mark the better result for each cue source.

Condition / method	CHAIR↓	Cover↑	Hal↓	Cog↓
(a) Weak visual condition				
None (Vanilla OPSD)	5.9	**48.5**	26.4	3.0
Black image	5.2	47.9	22.9	**2.2**
Gaussian-noise image	5.6	48.1	23.8	2.4
Gaussian blur (default, $\sigma = 10$)	**5.0**	47.7	**22.4**	2.3
Random patch masking (75%)	5.6	48.0	23.7	2.4
(b) Multi-seed stability				
Vanilla OPSD	5.97±0.06	**48.63±0.21**	25.63±0.68	2.77±0.12
VS-OPSD	**5.43±0.15**	48.03±0.25	**23.73±0.85**	**2.47±0.12**
(c) Robustness to automatically detected cues				
GT cues / Vanilla OPSD	5.7	48.2	25.5	2.6
GT cues / **VS-OPSD**	**5.5**	**48.5**	**23.9**	**2.4**
Grounding DINO / Vanilla OPSD	6.4	**48.9**	27.9	3.1
Grounding DINO / **VS-OPSD**	**5.6**	48.6	**24.7**	**2.5**

Table 2(b): Visual-KL outperforms generic teacher–student disagreement. This panel reports step 100. With a fixed token budget and uniform weighting, Visual-KL top-30 obtains 25.4 Hal, outperforming teacher–student-KL top-30 at 26.9 and Visual-KL bottom-30 at 27.7. Visual dependence is therefore a more effective selection signal for hallucination mitigation than generic teacher–student disagreement.

Table 2(c): Top-30% provides the best balance. Top-10%, top-30%, and top-50% obtain 23.3, 22.4, and 23.0 Hal, respectively, with top-30% also achieving the lowest CHAIR. A smaller budget may omit visual signals, whereas a larger one reintroduces low-sensitivity tokens.

Q2. Robustness

Are the gains robust to interventions, random seeds, and noisy cues?

Table 3(a): Robustness to weak-image construction. On a fixed 200-image probe, all intervention pairs have Spearman rank correlations above 0.92 and retain 83–93% of the tokens selected under the black-image condition. End-to-end training with black images, Gaussian noise, Gaussian blur, and patch masking yields 22.9, 23.8, 22.4, and 23.7 Hal, respectively, all below the 26.4 of Vanilla OPSD. Visual selection therefore does not depend on artifacts from a particular degradation operator.

Table 3(b): Stability across random seeds. Across three matched seeds, VS-OPSD reduces CHAIR, Hal, and Cog in every pair and lowers mean Hal from 25.63 ± 0.68 to 23.73 ± 0.85, while Cover decreases by 0.60 on average.

Table 3(c): Robustness to automatically detected cues. Grounding DINO cues increase Vanilla OPSD Hal from 25.5 to 27.9. Under the same detector cues, VS-OPSD reduces Hal to 24.7, only 0.8 above its ground-truth-cue counterpart. Thus, VS-OPSD does not require perfect manual cues and filters part of the noise introduced by automatic detection.

4.4 Mechanism Analysis

Visual sensitivity is not generic distillation difficulty. Using a fixed probe set, Figure 4 compares Visual-KL with teacher-student KL, which measures the disagreement between teacher and student. Their correlation is low ($r = 0.14$), while selected tokens have much larger Visual-KL than unselected tokens (0.695 vs. 0.030 after clipping at the 99th percentile). The selected set overlaps with the largest teacher-student KL tokens by 0.584 on average. The two signals are related but far from identical. Visual selection therefore identifies positions where the image changes the privi-

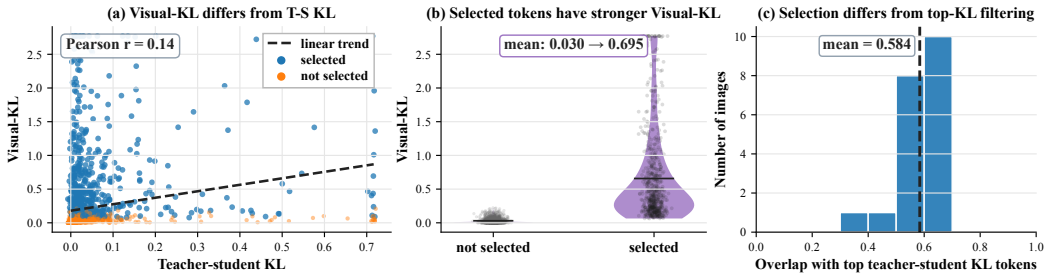

Figure 4: **Visual-KL isolates image-sensitive positions rather than generic teacher–student disagreement.** (a) Visual-KL has weak correlation with ordinary teacher–student KL. (b) Top-30% selected tokens carry a substantially stronger visual signal. (c) Their overlap with tokens selected by ordinary teacher–student KL is only partial. Statistics use a fixed probe set.

leged teacher, rather than merely choosing tokens that the student already finds difficult to imitate. Additional fixed-rollout alignment and training-dynamics analyses are provided in Appendix D.2.

5 Related Work

MLLM hallucination mitigation. Existing work mitigates multimodal hallucination through training-time alignment and inference-time correction. Training-based methods improve visual faithfulness using hallucination-aware supervision, preference optimization, or fine-grained grounding signals (Jiang et al., 2024; Yu et al., 2024; Wang et al., 2024; Sarkar et al., 2025), whereas inference-time methods use visual contrast, decoding-time intervention, or post-hoc response correction (Leng et al., 2024; Huang et al., 2024; Yin et al., 2024). In contrast, VS-OPSD requires neither target captions nor rejected responses and does not modify test-time decoding. It uses visual intervention only during training to identify which positions should receive privileged supervision.

On-policy and multimodal distillation. On-policy distillation matches token-level teacher distributions along trajectories generated by the student, mitigating the mismatch between fixed training sequences and the autoregressive states visited at inference (Agarwal et al., 2024). MiniLLM (Gu et al., 2024), GKD (Agarwal et al., 2024), and DistiLLM (Ko et al., 2024) further develop generative distillation through divergence design, student sampling, and training-data construction, while OPSD (Zhao et al., 2026) introduces a self-teacher conditioned on privileged context. Recent work has begun to extend on-policy distillation to multimodal models. Vision-OPD (Yuan et al., 2026) constructs a regional teacher from an evidence-centered crop and distills its fine-grained local perception into a full-image student. VA-OPD (Liu et al., 2026) compares an external teacher's probabilities for generated tokens under the original and pixelated images, using the resulting visual advantage for token grouping and multi-rollout weighting. In the matched setting, VS-OPSD obtains a lower AMBER hallucination rate than our reproduction of VA-OPD (Appendix E.1). While these works primarily target fine-grained visual understanding and reasoning, we study how to allocate non verifying visual cues to visually dependent positions for open-ended object hallucination.

6 Conclusion

We extend OPSD to open-ended VLM hallucination mitigation, where no unique reference answer exists, by using non-verifying object cues as training-time privileged context. We further observe that response tokens depend on visual evidence to markedly different degrees. Based on this observation, we propose VS-OPSD, which estimates token-level visual dependence with Visual-KL and concentrates sensitivity-weighted distillation on visually sensitive positions. Extensive experiments show that VS-OPSD reduces hallucination while preserving general multimodal capability. Ablations and robustness evaluations further validate the visual selection mechanism and its consistency across visual interventions, on-policy sampling, and cue quality. Promising directions for future work are discussed in Appendix F.

Ethics Statement

This work uses publicly available models and datasets and does not involve human participants, private data, or personally identifiable information. Although the proposed method aims to improve visual faithfulness, it does not eliminate hallucinations completely, and model outputs should not be relied upon without verification in safety-critical applications.

Reproducibility Statement

We provide the model configurations, training-data construction, prompts, optimization settings, random seeds, evaluation protocol, and evaluation procedures used in our experiments. The main paper and supplementary material further report implementation details, controlled ablations, and multi-seed results. All datasets and pretrained model checkpoints used in this work are publicly available. Experiments were conducted on NVIDIA RTX 4090 GPUs.

AI Use Statement

Generative AI tools were used solely to improve the clarity, grammar, and readability of author-written manuscript text. They were not used for research ideation, experimental design or execution, literature retrieval, data generation, result interpretation, or mathematical reasoning. All suggested edits were reviewed and revised by the authors, who take full responsibility for the final content of this work.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036. IEEE, 2024.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, volume 2025, pp. 73796–73846, 2025.

Qwen Team. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 5(6), 2025.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11766–11781, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 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.

Appendix

The organization of the appendix is as follows:

Appendix A: Implementation and Reproducibility Details

Appendix B: Statistical Robustness and Complete Results

Appendix C: Extended Ablations

Appendix D: Additional Mechanism and Qualitative Analyses

Appendix E: Additional Comparisons and Scope

Appendix F: Future Work

A Implementation and Reproducibility Details

A.1 Data Construction and Privileged Cues

Our matched experiments use 500 images sampled from COCO train2014 (Lin et al., 2014). We remove crowd instances and objects whose annotated area is below 1,024 pixels. We then deduplicate category names, rank the remaining categories by object area, and retain at most five categories per image. This produces a set of coarse object cues without attributes, relations, or a complete target caption.

The student receives the image and the ordinary captioning instruction. The privileged teacher receives the same image and an instruction augmented with the object cues. As illustrated in Figure 5, for an image containing the retained object categories *laptop*, *book*, and *mouse*, the textual instructions before applying the backbone-specific conversation template are:

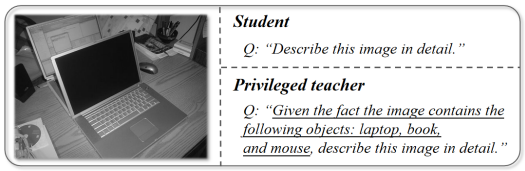

Figure 5: Examples of Teacher and Student Instructions

The weak-image teacher uses exactly the same privileged instruction and student prefix as the privileged teacher. Its only changed input is the image. The student rollout is therefore generated without access to object cues, and all teacher distributions are evaluated on that same rollout.

A.2 Weak-Image Construction

The default weak image is obtained with strong Gaussian blur. For an image of width W and height H, the blur radius is

$$r = 10 \frac{\min(W, H)}{336}. \tag{9}$$

This preserves coarse layout and low-frequency image statistics while weakening edges, textures, and fine-grained object evidence. The weak-image ablation also uses an all-zero black image, Gaussian noise with standard deviation 0.25, and random masking of 75% of a 24×24 grid. All variants keep the student rollout, privileged cues, token budget, and optimization settings fixed.

A.3 Training Configuration

Table 4 reports the default configuration. We use the same frozen base model for the privileged and weak-image teachers. LoRA adapters are enabled only for the student update and are disabled while computing teacher targets. We also provide additional details on the SFT and DPO baselines. For SFT, the corresponding base VLM generates one caption per image using object cues, after which

the cues are removed and the fixed caption is used as the cross-entropy target. For DPO, we train on 1,000 preference pairs randomly sampled from RLHF-V for 100 or 300 optimizer steps. Since DPO uses different training data, it is included only as an external reference.

Table 4: Default training configuration for matched COCO-500 experiments.

Component	Configuration
Train images	500 COCO train2014 images
Training steps	100 optimizer steps
Per-device batch size	1
Gradient accumulation	8
Learning rate	5×10^{-6}
Maximum gradient norm	0.5
Precision	Bfloat16 with 4-bit base weights
LoRA rank / scaling / dropout	64 / 128 / 0.05
LoRA targets	All linear layers
Rollout sampling	Temperature 0.7, top-p 0.9
Maximum new tokens	128
Sequence length	2,048
Policy divergence	Equal-weight JSD ($\beta = 0.5$)
Distillation temperature	1.0
Vocabulary approximation	Teacher top-128 logits
Visual sensitivity	Forward KL
Weak image	Gaussian blur, $\sigma = 10$ at 336 pixels
Selected-token budget	Top 30%
Hardware	One NVIDIA GeForce RTX 4090

A.4 Evaluation Protocol

We evaluate open-ended hallucination on the 1,004-image AMBER generative split (Wang et al., 2023) and on a fixed 500-image COCO val2014 CHAIR subset (Rohrbach et al., 2018). POPE measures discriminative object hallucination (Li et al., 2023b), while MME and ScienceQA measure general multimodal capability (Fu et al., 2026; Lu et al., 2022). Object cues are never provided during evaluation. AMBER generation uses greedy decoding with at most 128 new tokens, and all methods in a matched comparison use the same prompt and decoding configuration.CHAIR generation uses at most 128 new tokens by default. For Qwen CHAIR, we allow up to 512 new tokens for all variants to reduce truncation of its longer responses.

A.5 End-to-End Training Procedure

For completeness, one optimization step consists of the following operations:

1. Sample a response $\hat{y}$ from the current student using only the image I and the prompt q.
2. Evaluate the frozen privileged teacher on $(I, q, h, \hat{y}_{<t})$ for every response position.
3. Replace only the teacher image with the weak image I^{-} and evaluate the weak-image teacher on the same cues and prefixes.
4. Compute token sensitivity $a_t = \mathrm{KL}(p_T^t \,\|\, p_W^t)$, retain the top-$\rho$ positions, and normalize their sensitivity weights.
5. Minimize sensitivity-weighted JSD between the student and privileged teacher distributions at the selected positions. Gradients flow only through the student logits.

Algorithm 1 summarizes the same procedure for a minibatch. The two teacher branches are frozen and evaluated on the identical student-generated trajectory, and only the final student loss is differentiated.

Algorithm 1 One optimization step of VS-OPSD

Require: Minibatch $\mathcal{B} = \{(I, q, h)\}$, student p_θ, frozen teacher $p_{\bar{\theta}}$, weak-image transform g, token budget ρ

for all $(I, q, h) \in \mathcal{B}$ **do**
  Sample $\hat{y} \sim p_\theta(\cdot \mid I, q)$ and set $I^- \leftarrow g(I)$
  for $t = 1, \ldots, |\hat{y}|$ **do**
    $p_S^t \leftarrow p_\theta(\cdot \mid I, q, \hat{y}_{<t})$
    $p_T^t \leftarrow p_{\bar{\theta}}(\cdot \mid I, q, h, \hat{y}_{<t})$
    $p_W^t \leftarrow p_{\bar{\theta}}(\cdot \mid I^-, q, h, \hat{y}_{<t})$
    $a_t \leftarrow \mathrm{KL}(p_T^t \,\|\, p_W^t)$
  end for
  $\mathcal{S} \leftarrow \mathrm{Top}_\rho(\{a_t\}_{t=1}^{|\hat{y}|})$
  $\omega_t \leftarrow a_t \, \mathbf{1}[t \in \mathcal{S}]$
  $\mathcal{L}(I, q, h) \leftarrow \dfrac{\sum_t \omega_t \, \mathrm{JSD}\big(p_S^t \,\|\, \mathrm{sg}(p_T^t)\big)}{\sum_t \omega_t + \epsilon}$
end for
Update θ using $|\mathcal{B}|^{-1} \sum_{(I,q,h)\in\mathcal{B}} \mathcal{L}(I, q, h)$

B Statistical Robustness and Complete Results

B.1 Per-Seed Results

Table 5 expands the mean and standard deviation reported in the main paper. Across three paired runs sharing the same data manifest, initialization rule, sampling configuration, and training budget, VS-OPSD consistently reduces CHAIR, Hal, and Cog, with average improvements of 0.54, 1.90, and 0.30, respectively. The consistent direction across seeds indicates that the gains are not driven by a single favorable on-policy sample and supports the robustness of visual selection to sampling randomness. The magnitude of the Hal reduction nevertheless varies across seeds, reflecting the inherent variance of on-policy rollouts.

Table 5: Complete paired-seed results under Gaussian blur at step 50. Standard deviations are sample standard deviations over three independent runs.

Method	Seed	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
Vanilla OPSD	42	6.0	48.8	25.4	2.7
Vanilla OPSD	43	6.0	48.4	26.4	2.9
Vanilla OPSD	44	5.9	48.7	25.1	2.7
Vanilla OPSD	Mean ± std	5.97 ± 0.06	$\mathbf{48.63 \pm 0.21}$	25.63 ± 0.68	2.77 ± 0.12
VS-OPSD	42	5.4	48.0	22.9	2.4
VS-OPSD	43	5.3	47.8	23.7	2.4
VS-OPSD	44	5.6	48.3	24.6	2.6
VS-OPSD	Mean ± std	$\mathbf{5.43 \pm 0.15}$	48.03 ± 0.25	$\mathbf{23.73 \pm 0.85}$	$\mathbf{2.47 \pm 0.12}$
Paired difference	Mean ± std	-0.53 ± 0.21	-0.60 ± 0.20	-1.90 ± 1.22	-0.30 ± 0.20

B.2 Response Length and Conservativeness

Shorter or more conservative responses may naturally obtain lower hallucination scores. We therefore measure response length, object mentions, and length-normalized hallucination density on the same AMBER generations in Table 6. Relative to Vanilla OPSD, VS-OPSD reduces average length and object mentions by only 3.3% and 3.7%, while reducing hallucinated objects per response and per 100 tokens by 15.2% and 12.3%, respectively. Thus, mild conservativeness does not explain the overall gain. VS-OPSD remains more effective after controlling for response length while reducing Cover by only 0.5 points, indicating that the improvement is not obtained merely by shortening the outputs.

Table 6: Response-length and hallucinated-object statistics on AMBER. Object mentions follow the AMBER noun extraction and synonym-matching protocol.

Method	Avg. tokens	Objects / response	Hall. objects / response	Hall. objects / 100 tokens	Cover ↑	Hal ↓
Base	82.84	7.04	0.496	0.599	46.6	29.8
SFT	85.73	7.40	0.537	0.626	**49.6**	31.4
Vanilla OPSD	81.06	7.05	0.400	0.494	48.2	25.5
VS-OPSD	78.38	6.79	**0.340**	**0.433**	47.7	**22.4**

B.3 Training Trajectories Across Data Regimes

We additionally train both methods for 300 steps on COCO-200 and COCO-2.4k to examine their full training trajectories across data regimes. In Table 7 , VS-OPSD is better than Vanilla OPSD at 10 of 12 COCO-200 checkpoints, reducing Hal by 0.57 on average, and at all six COCO-2.4k checkpoints, with a larger average reduction of 1.60. On COCO-200 (Figure 6(a)), the advantage is concentrated earlier in training and narrows after repeated exposure to the small dataset, which may reflect overfitting across multiple epochs. On COCO-2.4k (Figure 6(b)), the improvement is larger and more consistent, suggesting promise at greater training-data scale. Moreover, under the COCO-500 setting used in the main experiments, model performance approaches a stable regime within 100 steps. We therefore use 100 steps as the default training budget in the main paper.

Table 7: Checkpoint summaries across two data regimes. Checkpoint averages describe correlated trajectories and are not treated as independent samples.

Data regime	Method	Checkpoints	Avg. CHAIR ↓	Avg. Hal ↓	Final Hal ↓
COCO-200	Vanilla OPSD	12	5.600	24.325	24.0
COCO-200	VS-OPSD	12	**5.450**	**23.750**	**23.4**
COCO-2.4k	Vanilla OPSD	6	5.700	25.333	26.0
COCO-2.4k	VS-OPSD	6	**5.517**	**23.733**	**24.5**

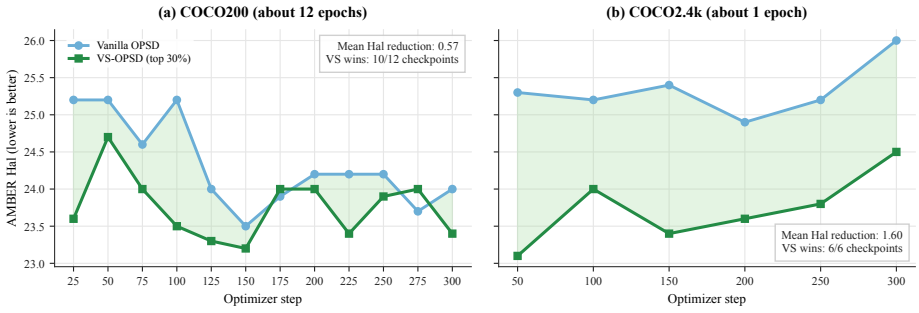

Figure 6: **Hallucination trajectories across two data regimes.** AMBER Hal over checkpoints for COCO-200 and COCO-2.4k. Shading marks checkpoints where VS-OPSD is better than Vanilla OPSD. Checkpoints from one run are correlated and are shown to characterize training dynamics.

C Extended Ablations

C.1 Blur Strength and Sensitivity Divergence

Table 8 varies only the weak-image blur strength or the divergence used to compute visual sensitivity. All variants use COCO-500, a top-30% token budget, and the same optimization setup. The default forward KL with $\sigma = 10$ provides the best step-50 result (Table 9). Across all configurations, Hal ranges from 22.4 to 24.2 and remains below the matched Vanilla OPSD result of 26.4, showing that the gain is robust to the exact blur strength and sensitivity divergence. Forward KL with $\sigma = 10$ achieves both the lowest CHAIR (5.0) and Hal (22.4), making it an empirical sweet

spot between visual degradation strength and sensitivity discrimination. We therefore adopt it as the default configuration.

Table 8: Sensitivity-estimation ablation at step 50.

Blur σ	Sensitivity divergence	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
5	Forward KL	5.5	47.8	23.0	**2.3**
10	**Forward KL**	**5.0**	47.7	**22.4**	**2.3**
20	Forward KL	5.4	**48.4**	24.2	2.4
10	Reverse KL	5.6	48.0	24.1	2.5
10	JSD	5.4	48.0	23.4	2.4

Table 9: Complete AMBER Hal trajectories for the sensitivity-estimation ablation. Bold values mark the best result in each checkpoint column.

Configuration	Step 25	Step 50	Step 75	Step 100	Best Hal (step)
Blur $\sigma = 5$, forward KL	25.4	23.0	**24.0**	24.6	23.0 (50)
Blur $\sigma = 10$, forward KL	23.7	**22.4**	24.1	24.3	**22.4 (50)**
Blur $\sigma = 20$, forward KL	24.9	24.2	24.1	24.3	24.1 (75)
Blur $\sigma = 10$, reverse KL	24.8	24.1	25.1	25.0	24.1 (50)
Blur $\sigma = 10$, JSD	**22.8**	23.4	24.7	**24.2**	22.8 (25)

C.2 Granularity of Privileged Context

To examine how privileged-context granularity affects token selection, we compare VS-OPSD using complete COCO reference captions with its default variant using object-category cues. We include the LLaVA-1.5-7B base model and Vanilla OPSD with object cues as references. All trained variants use COCO-500 and the same training budget. Table 10 retains the results reported in the main paper for the other methods and reports the step-100 result for reference-caption VS-OPSD. Notably, reference captions provide substantially more detailed image annotations than object categories and typically require greater resources to obtain.

Table 10: Effect of privileged-context granularity under the matched COCO-500 setting.

Method	Privileged context	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
LLaVA-1.5-7B Base	None	7.1	46.6	29.8	3.3
Vanilla OPSD	Object categories	5.7	**48.2**	25.5	2.6
VS-OPSD	Reference caption	5.2	47.8	23.2	**2.2**
VS-OPSD (Ours)	**Object categories**	**5.0**	47.7	**22.4**	2.3

Using object categories as privileged information does not weaken effective supervision. Compared with a complete reference caption, the default VS-OPSD obtains lower CHAIR (5.0 vs. 5.2) and Hal (22.4 vs. 23.2), while maintaining comparable Cover and Cog. Category-level object cues therefore provide sufficient privileged visual information for object-hallucination mitigation and visual alignment without requiring high-quality complete captions. Together with the robustness to automatically detected cues reported in the main paper, this result suggests a practical route to scaling the training data with inexpensive detector-derived supervision.

D Additional Mechanism and Qualitative Analyses

D.1 Selected Tokens Extend Beyond the Provided Cues

On the fixed 200-image probe set, only 7.35% of selected words exactly match a provided object cue, while 55.03% are non-cue content words. Figure 7 provides four representative cases spanning



Figure 7: **Visual-KL identifies grounded concepts beyond coarse object cues.** From top to bottom, the examples highlight entity/scene/tool, action/interaction, attribute/scene, and count/entity/action information. Gray dotted underlines mark cue-matched words, while blue solid underlines mark additional visually sensitive concepts selected by VS-OPSD. Bars report the token-level sensitivity of representative selected tokens. All cases use the Gaussian-blur weak-image condition.

entities, scenes, tools, attributes, quantities, actions, and interactions. Visual-KL identifies visually grounded concepts more specific than the coarse object categories, showing that it supervises visually dependent tokens rather than merely copying cue tokens.

D.2 Where Teacher Alignment Improves During Training

Figure 8 compares the training objectives and alignment dynamics of Vanilla OPSD and VS-OPSD. Because the two methods optimize different token supports, each objective is normalized by its own initial-window mean. Their normalized objectives decrease at similar rates, showing that VS-OPSD does not derive its gains from faster convergence. We then measure teacher-student alignment on a fixed probe using three metrics. Top-k overlap is the fraction shared by the teacher's and student's top-k candidate sets and measures support agreement. Overlap KL measures the distributional discrepancy after renormalizing both distributions over their shared candidates, while entropy gap measures the difference in predictive uncertainty. Higher overlap and lower overlap KL and entropy gap indicate better alignment. We compute the improvement over the base model separately for selected and unselected tokens and report their ratio. VS-OPSD yields a larger selected-to-unselected improvement ratio than Vanilla OPSD for all three metrics. Thus, it concentrates teacher-alignment gains at visually sensitive positions rather than simply strengthening global teacher imitation. Fig-

ure 8(c) further shows that VS-OPSD maintains lower AMBER Hal along the extended training trajectory, indicating that its gains are not confined to a particular early checkpoint.

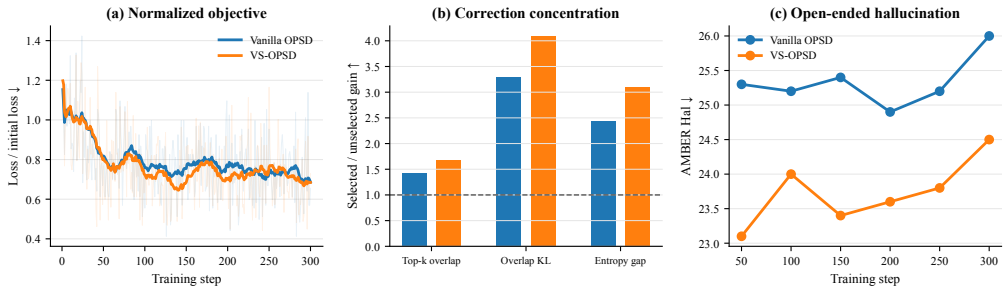

Figure 8: **VS-OPSD reallocates correction without faster global convergence.** (a) Objective normalized by the first 20-step mean, with a 20-step moving average. (b) Alignment improvement on selected tokens relative to unselected tokens on a fixed 200-image probe. (c) AMBER Hal over checkpoints. Checkpoints are correlated trajectory observations.

E Additional Comparisons and Scope

E.1 Matched VA-OPD-Style Comparison

As shown in Table 11, we implement a matched VA-OPD-style baseline (Liu et al., 2026) with the same LLaVA backbone, COCO-500 images, object cues, and 100-step budget. It samples four sibling rollouts per input, estimates visual advantage from the probability of generated tokens under the original and a 10%-resolution pixelated image, performs grouped distillation over high- and low-advantage tokens, and weights sibling rollouts by their mean visual advantage. This is a controlled adaptation rather than a strict reproduction of the original VA-OPD setting, which uses an external teacher and visual reasoning data.

Table 11: Matched AMBER comparison with our VA-OPD reproduction. VA-OPD is reported at its best checkpoint (step 25), while the remaining results match the main table.

Method	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
LLaVA-1.5-7B	7.1	46.6	29.8	3.3
+ SFT	7.3	**49.6**	31.4	3.7
+ Vanilla OPSD	<u>5.7</u>	48.2	<u>25.5</u>	<u>2.6</u>
+ VA-OPD (reproduced)	6.3	<u>48.7</u>	28.0	2.9
+ VS-OPSD	**5.0**	47.7	**22.4**	**2.3**

Compared with SFT, VA-OPD reproduction reduces CHAIR, Hal, and Cog from 7.3, 31.4, and 3.7 to 6.3, 28.0, and 2.9, respectively, showing that visual advantage provides a useful signal for open-ended generation. It nevertheless does not outperform Vanilla OPSD on these metrics, whereas VS-OPSD achieves the lowest CHAIR (5.0), Hal (22.4), and Cog (2.3).

VA-OPD considers only the decrease in an external teacher's log-probability of the realized token y_t between the original and pixelated images, with negative values clipped to zero. It therefore measures whether removing fine-grained visual information reduces the teacher's confidence in the sampled token. Visual-KL instead computes forward KL between the top-k next-token distributions of an in-model privileged teacher under real and blurred images, while holding the object cues and generated prefix fixed. It therefore captures the redistribution of probability mass over the candidate set rather than only the probability change of a single sampled token. This difference in signal leads to different optimization strategies. VA-OPD uses visual advantage for sibling-rollout weighting and high-/low-advantage token-group distillation. VS-OPSD selects the highest-Visual-KL positions and weights them by their sensitivity, requiring neither a separate external teacher nor sibling-rollout

weighting. Our work is therefore positioned as allocating privileged visual supervision for open-ended hallucination mitigation, whereas VA-OPD primarily transfers fine-grained visual reliance in visual understanding and reasoning.

E.2 Computational Efficiency

We profile Vanilla OPSD, VS-OPSD, and our VA-OPD reproduction on a single NVIDIA RTX 4090. All methods use the same LLaVA-1.5-7B backbone, COCO-500 data, per-device batch size of 1, eight gradient-accumulation steps, and 4-bit LoRA. We discard five warm-up optimizer steps and measure the next ten steps, excluding checkpoint I/O.

Table 12: Training cost on one NVIDIA RTX 4090. Step time is the mean and standard deviation over ten optimizer steps, and relative time is normalized to Vanilla OPSD. VA-OPD uses gradient checkpointing, so its memory value is not directly comparable to the other rows.

Method	Step time (s)	Samples/s	Peak reserved (GiB)	Relative time
Vanilla OPSD	108.76 ± 5.30	0.074	17.30	1.00×
VA-OPD (reproduced)	158.59 ± 7.03	0.050	11.60	1.46×
VS-OPSD	109.77 ± 7.17	0.073	18.12	1.01×

Although VS-OPSD additionally computes Visual-KL and introduces a weak-image teacher forward path, its mean step time and peak reserved memory increase by only 0.9% and 4.7%, respectively, relative to Vanilla OPSD. This indicates that visual sensitivity estimation introduces no substantial additional training overhead in our implementation. VS-OPSD is also more computationally efficient than our VA-OPD reproduction, in part because it does not require multiple sibling rollouts to compute within-group visual advantage. The current implementation still performs dense forward computation over the complete rollout and selects the top-30% tokens only when computing the loss. This token budget therefore reflects more effective supervision allocation rather than a reduction in FLOPs. All additional branches are used only during training, so VS-OPSD adds no inference parameters, latency, or decoding operations.

F Future Work

Lightweight visual-sensitivity estimation. Although the measurements in Appendix E.2 show only modest overhead over Vanilla OPSD, Visual-KL still requires an additional weak-image teacher forward pass. Future work could explore lower-cost approximations of token-level visual dependence and systematically study the trade-off between efficiency and performance.

Scaling without ground-truth cues. The main experiments use COCO object annotations, but the Grounding DINO experiment (Liu et al., 2024b) shows that automatically detected cues remain effective for VS-OPSD without ground-truth object labels. This result opens a path toward constructing privileged cues automatically from large-scale unlabeled image collections. Future work could explore this strategy at larger data scales, across different domains, and with different detectors.