QUESTION-GUIDED SELF-DISTILLATION FOR REDUCING FALSE-POSITIVE HALLUCINATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can affirm nonexistent objects, incorrect attributes, and invalid relations, or answer questions whose visual premises are false. Reducing these false positives requires correcting unsupported content while retaining valid answers. We investigate question-guided self-distillation, which uses factual conditions implicit in the original question as training guidance. Verification subquestions are supplied to a frozen base model without subquestion answers. A student generates responses from the original image and question and learns from the conditioned teacher's token distributions on those same response prefixes. A second distribution-matching term constrains departure from the base model under the original input. Both terms use token-level Jensen–Shannon divergence; deployment requires neither question decomposition nor additional verification calls. With Qwen3-VL-8B and a 60K training pool, *wh* paired accuracy improves over our FINER-DPO baseline from 46.62% to 57.99% on FINER-CompreCap and from 36.42% to 47.01% on FINER-DOCCI. On DOCCI *wh* queries, negative-query accuracy rises from 46.09% to 61.03%, while positive-query accuracy falls from 83.60% to 80.63%. Other query settings and general capability evaluations reveal further trade-offs. These results show improved direct false-premise handling with the complete training recipe, while identifying selective correction and capability retention as remaining challenges.

1 INTRODUCTION

Research on multimodal hallucination has advanced both the evaluation and mitigation of unsupported visual claims. Fine-grained benchmarks now test whether vision-language models distinguish true image content from closely related false descriptions, while verification-based reasoning and preference learning have improved factual reliability across hallucination evaluations (Xiao et al., 2026; Wu et al., 2024; Chen et al., 2026). These advances establish visual factuality as a learnable capability and provide tools for examining errors beyond the mere presence or absence of an object. A central challenge is false-positive hallucination: affirming an absent object, an incorrect attribute, an invalid relation, or a false premise embedded in a question.

Several complementary approaches address this challenge. Wu et al. (2024) introduce bottom-up holistic reasoning that verifies visual and textual information before forming an answer. Xiao et al. (2026) develop FINER and use fine-grained preference pairs for FINER-Tuning, improving hallucination mitigation and general multimodal performance. Chen et al. (2026) use verification-guided search and rewards to construct preference data for iterative self-calibration. Despite this progress, false-premise answering remains difficult: our Qwen3-VL FINER-DPO reproduction improves object, attribute, and relation judgments over Base, but leaves *wh* paired accuracy below Base on both FINER datasets (Table 1). The underlying difficulty is that recognizing plausible visual content does not establish every condition assumed by a question. Evidence must support the correct entity and its associated attributes or relations, and that judgment must constrain the final answer. Strengthening rejection alone also risks discarding valid questions. This motivates learning directly from the factual conditions that determine whether an answer is supported.

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.

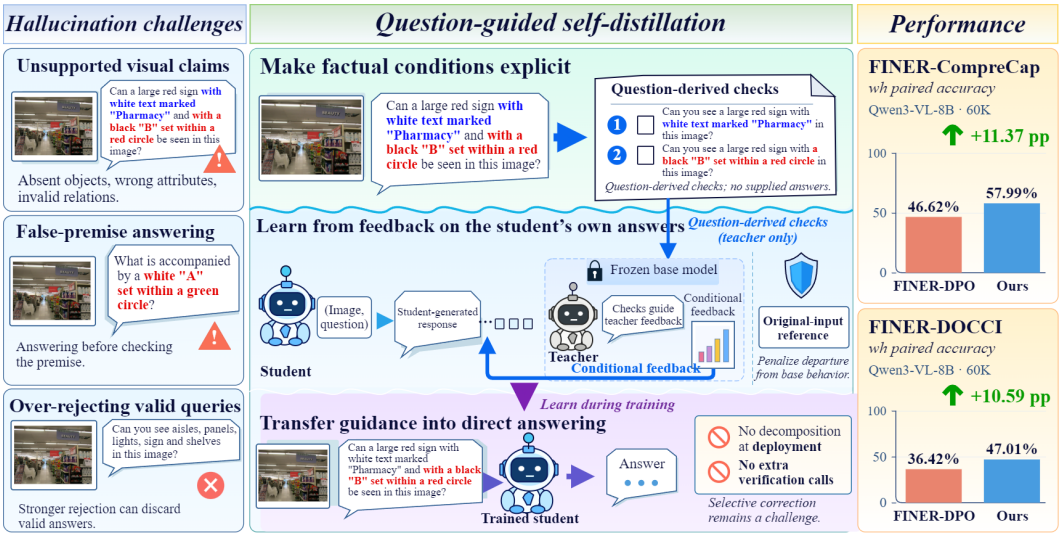

Figure 1: Question-guided self-distillation. Left: hallucination challenges. Center: question-derived checks guide teacher feedback on student answers; inference uses the original image and question. Right: selected *wh* paired-accuracy gains over FINER-DPO.

We introduce question-guided self-distillation, our improved on-policy distillation (OPD) recipe for reducing false-positive hallucination. During training, we reorganize the original question into verification subquestions and supply them, without their answers, to a frozen base model as auxiliary context. The student generates responses from the original image and question, then matches the conditioned teacher's token distributions along its own response prefixes. A second distribution-matching term penalizes departure from the same frozen model under the original input. This makes question-derived verification context a source of feedback for learning direct answering. The resulting student improves false-premise handling while retaining the original image–question interface, with no question decomposition or additional verification calls at deployment (Figure 1).

We evaluate the complete recipe with Qwen3-VL-8B-Instruct and a 60K training pool, comparing against Base and our FINER-DPO reproduction. Experiments cover all four query settings in FINER-CompreCap and FINER-DOCCI, external factuality benchmarks, and five general multimodal capability benchmarks. Paired accuracy requires both the positive and negative query to be answered correctly. Analyses of paired predictions, training dynamics, and qualitative cases further characterize corrections and regressions.

The results demonstrate improved direct answering of questions with false visual premises. For Qwen3-VL, relative to FINER-DPO, *wh* paired accuracy rises from 46.62% to 57.99% on FINER-CompreCap and from 36.42% to 47.01% on FINER-DOCCI, gains of 11.37 and 10.59 percentage points (Table 1). On DOCCI *wh*, negative-query accuracy increases from 46.09% to 61.03% (Section 4.4), and DASH-B accuracy improves from 74.35% to 76.51% (Table 2). These four comparisons show reduced errors on false premises and an additional factuality gain beyond FINER. The accompanying decline on positive DOCCI *wh* queries identifies selective correction as a remaining challenge.

Our contributions are threefold:

❶ ***Question-Guided Self-Distillation.*** We combine question-conditioned feedback on student-generated responses with an original-input reference objective, retaining direct image–question answering at deployment.

❷ ***Improved False-Premise Answering.*** We demonstrate improved false-premise answering over our FINER-DPO baseline on two fine-grained hallucination datasets and evaluate external factuality and general multimodal capabilities.

❸ ***Correction and Retention Analysis.*** We characterize corrections and regressions through paired predictions, query-condition groups, training diagnostics, and verified cases, exposing the distinction between improved false-premise handling and preservation of valid answers.

2 Related Work

2.1 Multimodal Hallucination and Factual Verification

False-positive hallucination includes accepting visual content that is absent or inconsistent with the image. FINER evaluates such behavior using positive–negative query pairs across objects, attributes, relations, and wh questions (Xiao et al., 2026). Judge Before Answer studies recognition of false premises in visual questions (Li et al., 2025). Explicit verification approaches, including bottom-up holistic reasoning, organize checks of visual content before forming a response (Wu et al., 2024). Our work uses these established problem formulations and existing benchmarks. Verification subquestions serve as training context, while the deployed student retains direct answering.

2.2 Learning from Corrective Feedback

Direct preference optimization learns from preferred and rejected responses (Rafailov et al., 2023). FINER-Tuning applies this approach to fine-grained positive and negative content (Xiao et al., 2026). OSCAR uses verification-guided search and rewards to construct preference data for iterative model improvement (Chen et al., 2026). These studies establish that corrective feedback and a model's discriminative behavior can be useful for hallucination mitigation. Our training loss instead matches conditional token distributions on the student's responses. The empirical comparison concerns the complete resulting recipes and does not isolate the objective function from all other training choices.

2.3 Self-Distillation with Privileged Context

Conditional self-distillation transfers behavior elicited by training-time context into a model operating with a less informative input. Vision-OPD uses crop-conditioned guidance for full-image answering (Yuan et al., 2026); Visual-OPSD transfers visual-thought-conditioned behavior (Li et al., 2026). CVPD constructs contrastive guidance from counterfactual visual views and also uses a reference objective (Venkatraman et al., 2026). Our method shares this general learning direction. Its auxiliary context reorganizes the factual conditions in the original question without changing the image or supplying their answers. We evaluate this choice in false-positive hallucination settings and compare question-specific context with a fixed generic reminder in Section 4.3. Independent contributions of current-student prefixes and the reference term remain open.

3 Method

3.1 Overview

We use factual conditions expressed in a visual question as training-time guidance for direct answering. Given an image I and question q, the student S samples a response y using only (I, q). A frozen teacher H receives the same image I, the question q, and auxiliary verification subquestions derived from its text. It then scores the student's sampled prefixes $y_{<t}$.

A reference role R uses the same frozen base model with the original input (I, q). Training matches the student's token distributions to both distributions; inference uses only the adapted student (Figure 2).

The design connects two requirements of false-positive hallucination mitigation. Conditions embedded in q should influence y, while correcting unsupported content should not amount to indiscriminate rejection. Question-derived context makes these conditions explicit to H. The original-input role R retains a distributional comparison with the base model.

Both signals act on the same response y generated under the student's deployment input (I, q). This information flow aims to transfer verification-conditioned behavior into direct answering; its effectiveness and trade-offs are evaluated empirically rather than assumed from the construction.

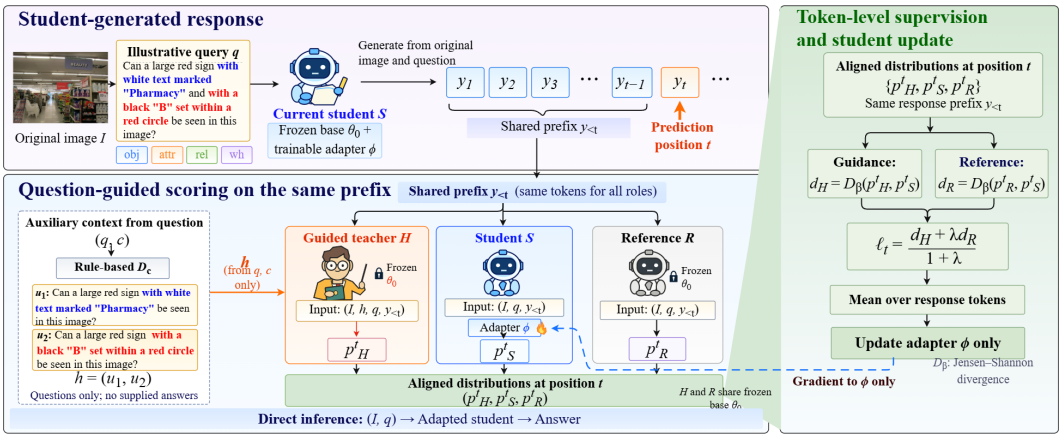

Figure 2: Method overview. A question-guided teacher and an original-input reference score the student's shared response prefixes. Their token-level signals update the student adapter; inference uses only the image and question.

3.2 Question-Derived Teacher Context

Let c denote the training query type and $\mathcal{D}_c$ its rule-based question transformation. We represent the resulting verification context as an ordered sequence of subquestions:

$$h(q,c) \triangleq \mathcal{D}_c(q) = \big(u_k(q,c)\big)_{k=1}^{K(q,c)}, \qquad x_H \triangleq \big(I, h(q,c), q\big). \tag{1}$$

Here $K(q,c)$ is the number of subquestions, and x_H denotes the teacher input, which retains the original query and image. For $c = \text{obj}$, queries are separated into queried objects; for $c \in \{\text{attr}, \text{rel}\}$, the rules retain the entity phrase when separating conditions. For $c = \text{wh}$, $\mathcal{D}_c$ extracts descriptive conditions associated with the queried entity. When a specific rule does not apply, the fallback uses text splitting or a single verification question with $K(q,c) = 1$.

Retaining the entity phrase matters because an attribute or relation is supported only in association with the relevant subject. Figure 2 uses an actual attribute query about a red "Pharmacy" sign and a black "B" within a red circle. Its $K(q,c) = 2$ checks ask separately about the sign with white "Pharmacy" text and the sign with a black "B" in a red circle; both retain the same sign as their subject. The checks reproduce the training-time text transformation and contain no supplied visual answers.

The original q remains in x_H alongside (u_1, u_2), allowing the teacher to consider their joint meaning. The method does not aggregate separate yes/no decisions for u_1 and u_2 or assume that evidence for different entities establishes q.

For a premise-bearing question with $c = \text{wh}$, the requested answer and descriptive premise play different roles. The transformation $h(q,c)$ exposes the description while retaining q. The teacher scores $y_{<t}$ with an explicit reminder of the conditions on which the answer depends.

Each u_k is constructed from (q,c) alone; neither image content nor the target answer determines its wording. Visual information enters when the model evaluates x_H on the unchanged image I.

The context supplies no subquestion answers or atomic truth labels. Let $a \in \{H, R, S\}$ index the conditioned teacher, original-input reference, and student. All three score the same student-generated prefix $y_{<t}$ through a common token-distribution form:

$$p_a^t(v) \triangleq \frac{\exp z_{\vartheta_a}(v; x_a, y_{<t})}{\sum_{w \in \mathcal{V}} \exp z_{\vartheta_a}(w; x_a, y_{<t})}, \qquad a \in \{H, R, S\}, \quad v \in \mathcal{V}. \tag{2}$$

Here z_{ϑ_a} denotes the vocabulary logits. The teacher uses x_H from Eq. (1), whereas $x_R = x_S = x \triangleq (I, q)$. Teacher and reference share frozen parameters $\vartheta_H = \vartheta_R = \theta_0$; the student uses $\vartheta_S = (\theta_0, \phi)$ with trainable adapter ϕ.

Thus p_H^t scores the student's response with question-derived guidance, while p_R^t scores it under x. Neither independently answers and aggregates the subquestions u_k. The question rules can fail to preserve linguistic dependencies, so the conditioned teacher is not assumed to be uniformly more accurate. The type c selects $\mathcal{D}_c$ during training; deployment requires neither c nor $h(q, c)$.

3.3 Learning for Direct Answering

The student generates a fresh response from the original input at each training example. Let $\widetilde{p}_\phi$ denote the current student's rollout distribution under the configured sampling rule. A response of length $T = |y|$ follows the autoregressive factorization

$$y \sim \widetilde{p}_\phi(\cdot \mid x), \qquad \widetilde{p}_\phi(y \mid x) = \prod_{t=1}^{T} \widetilde{p}_\phi(y_t \mid x, y_{<t}). \tag{3}$$

Here $\widetilde{p}_\phi$ includes generation-time sampling settings, whereas Eq. (2) defines the full-vocabulary scoring distributions p_a^t used in the loss. Generation stops at an end token or the response-length cap. The sampled tokens y_t supply the prefixes $y_{<t}$ for learning; they are not treated as correct target answers.

As ϕ changes, newly sampled $y_{<t}$ can change with it. The teacher consequently supplies feedback on continuations the current student actually visits, rather than only on a fixed collection of stored responses.

We append the same response tokens to each input and align predictions by their respective response boundaries. At position t, all roles therefore score the next token after $y_{<t}$, even though the teacher's prompt is longer. Only positions $t \in \{1, \ldots, T\}$ enter the loss. This avoids matching unrelated prompt positions: p_H^t and p_R^t differ in their input conditions at a shared $y_{<t}$.

The targets p_H^t and p_R^t are obtained with the adapter disabled. Their shared θ_0 remains frozen, with no moving-average update.

For a target distribution p_T and student distribution p_S, let $m_\beta = \beta p_T + (1-\beta) p_S$. Their generalized Jensen–Shannon divergence measures distributional disagreement through this shared mixture:

$$D_\beta(p_T, p_S) \triangleq \beta \, \mathrm{KL}(p_T \| m_\beta) + (1 - \beta) \, \mathrm{KL}(p_S \| m_\beta). \tag{4}$$

The mixture m_β is formed at each position t. Matching all $v \in \mathcal{V}$ transfers relative support for alternative tokens, including those not selected as y_t. This differs from treating y_t as a one-hot target: an unsupported continuation can still receive corrective distributional feedback through D_β.

Evaluating $D_\beta(p_H^t, p_S^t)$ requires neither a separate teacher answer nor a preferred–rejected response pair. The conditioned distribution is a training signal, however, not a ground-truth certificate of visual support.

We apply $D_\beta(p_H^t, p_S^t)$ for question-conditioned guidance and $D_\beta(p_R^t, p_S^t)$ for the original-input reference. These terms respectively transfer guidance into direct answering and penalize departure from the base distribution. For training examples drawn from $\mathcal{D}_{\text{train}}$, the objective balances these two terms and averages over generated response tokens:

$$\mathcal{L}(\phi) = \mathbb{E}_{\substack{(I,q,c)\sim\mathcal{D}_{\text{train}} \\ y\sim\widetilde{p}_\phi(\cdot|x)}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \left(\underbrace{\frac{D_\beta(p_H^t, p_S^t)}{1+\lambda}}_{\text{question-conditioned guidance}} + \underbrace{\frac{\lambda D_\beta(p_R^t, p_S^t)}{1+\lambda}}_{\text{original-input reference}} \right) \right]. \tag{5}$$

We use $\beta = 0.5$ and $\lambda = 0.5$, giving guidance and reference weights of $2/3$ and $1/3$, respectively. The factor $1/|y|$ gives each response a length-normalized contribution.

The reference term keeps p_R^t in the objective even when p_H^t favors a different continuation. It imposes a soft penalty rather than a hard constraint on the student's answers, so it does not guarantee capability preservation. The weight λ controls how strongly departure from p_R^t contributes to the training cost.

Table 1: Paired accuracy (%) for direct answers on FINER-CompreCap and FINER-DOCCI.

	FINER-CompreCap				FINER-DOCCI			
Method	Obj ↑	Attr ↑	Rel ↑	wh ↑	Obj ↑	Attr ↑	Rel ↑	wh ↑
Qwen3-VL-8B-Instruct								
Base	74.57$_{\uparrow 0.60}$	73.58$_{\uparrow 3.53}$	64.11$_{\uparrow 2.10}$	46.94$_{\uparrow 11.05}$	59.24$^{\dagger}_{\downarrow 0.12}$	62.55$^{\dagger}_{\uparrow 2.78}$	53.84$^{\dagger}_{\uparrow 1.80}$	37.79$^{\dagger}_{\uparrow 9.22}$
FINER-DPO 60K	**76.48**$_{\downarrow 1.31}$	**77.11**$_{= 0.00}$	**69.07**$_{\downarrow 2.86}$	46.62$_{\uparrow 11.37}$	**63.06**$_{\downarrow 3.94}$	**66.05**$_{\downarrow 0.72}$	**58.92**$_{\downarrow 3.27}$	36.42$_{\uparrow 10.59}$
Ours	75.17	**77.11**	66.21	**57.99**	59.12	65.33	55.64	**47.01**
InternVL3.5-8B-HF								
FINER-DPO 60K	**77.43**$_{\downarrow 3.33}$	**76.33**$_{\downarrow 0.24}$	**55.19**$_{\downarrow 1.87}$	27.35$_{\uparrow 7.45}$	**61.46**$_{\downarrow 4.38}$	57.45$_{\uparrow 0.27}$	**46.77**$_{\downarrow 1.96}$	19.19$_{\uparrow 5.64}$
Ours	74.10	76.09	53.32	**34.81**	57.08	**57.72**	44.81	**24.84**

3.4 Optimization and Deployment

Let $\ell(\phi; x, c, y)$ denote the token-averaged quantity inside the expectation in Eq. (5). For an accumulation batch of B sampled responses, the partial gradient before clipping and the AdamW update is

$$\widehat{g}_\phi = \frac{1}{B}\sum_{i=1}^{B} \nabla_\phi \ell(\phi; x_i, c_i, y_i)\big|_{y_i,\, p_{H,i},\, p_{R,i} \text{ fixed}}, \qquad \Delta\theta_0 = 0. \tag{6}$$

In Eq. (6), $(y_i, p_{H,i}, p_{R,i})$ are fixed during backpropagation, not across examples i. Frozen θ_0 still produces different predictions as $(I_i, q_i, y_{i,<t})$ changes.

Gradients pass through p_S^t, including its contribution to m_β. No gradient passes through discrete sampling of y, and no score-function term differentiates $\widetilde{p}_\phi$ in Eq. (5). Thus the displayed expectation specifies the sampled training loss, while Eq. (6) specifies the actual optimization rule.

Each iteration generates y, evaluates $(p_S^t, p_H^t, p_R^t)_{t=1}^T$, and contributes to $\widehat{g}_\phi$ before the adapter update. The roles share θ_0; enabling or disabling ϕ selects the trainable or frozen path. The implementation uses LoRA and AdamW; numerical settings and generation budgets are recorded in Appendix A.

At deployment, the adapted student $p_{\theta_0,\phi}$ uses only $x = (I, q)$. Constructing $h(q, c)$ and evaluating p_H^t, p_R^t are training-time operations, so direct inference adds neither subquestion generation nor answer aggregation.

No preferred or rejected answer enters $\mathcal{L}(\phi)$. However, $\mathcal{D}_{\text{train}}$ and its type and polarity metadata originate from an existing constructed dataset, so the complete data pipeline is not label-free.

4 Experiments

4.1 Benchmarks

Our primary evaluation uses FINER-CompreCap and FINER-DOCCI (Xiao et al., 2026), with 8,542 and 35,558 released positive–negative question pairs. Each tests four types of queried visual content: objects, attributes, relations, and *wh* premises. Paired accuracy requires correct answers to both members of a pair, capturing false-premise handling together with valid-answer retention. Evaluation is multiple-choice; training responses are free-form.

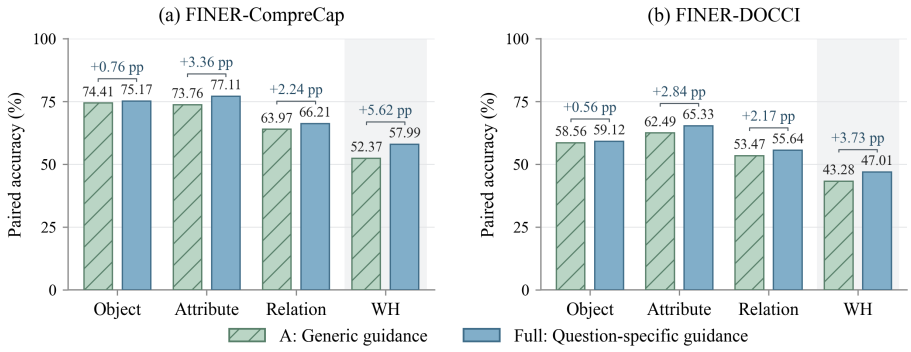

Figure 3: Paired accuracy with question-specific (Full) or generic (A) teacher context on FINER-CompreCap and FINER-DOCCI. Brackets show Full minus A in percentage points.

Table 2: Direct-inference factuality scores (%) on AMBER, CRPE_R, and DASH-B.

Method	AMBER Attr ↑	AMBER Rel ↑	CRPE_R Total ↑	DASH-B Acc ↑
Qwen3-VL-8B-Instruct				
Base	86.84$_{\downarrow 0.02}$	89.00$_{\uparrow 0.06}$	**72.33**$_{\downarrow 0.42}$	74.42$_{\uparrow 2.09}$
FINER-DPO 60K	**86.92**$_{\downarrow 0.10}$	88.52$_{\uparrow 0.54}$	71.54$_{\uparrow 0.37}$	74.35$_{\uparrow 2.16}$
Ours	86.82	**89.06**	71.91	**76.51**
InternVL3.5-8B-HF				
FINER-DPO 60K	80.55$_{\uparrow 0.12}$	79.15$_{\uparrow 3.12}$	**51.90**$_{\downarrow 4.38}$	**72.33**$_{\downarrow 4.51}$
Ours	**80.66**	**82.27**	47.52	67.82

We also evaluate factuality on AMBER, CRPE_R, and DASH-B, and general multimodal capability on MMStar, ChartQA, MMVP, NaturalBench, and V*. The main comparison uses Qwen3-VL-8B-Instruct (Base, FINER-DPO 60K, and Ours); InternVL3.5-8B-HF provides a second-backbone comparison between the two adapted recipes. Training pools and some evaluation settings differ across recipes; DOCCI was consulted during development. Appendix A–B gives the protocols. Tables 1–3 show Ours-minus-comparator changes as colored percentage-point subscripts; comparisons under mismatched protocols are descriptive.

4.2 Main Results

The largest gains over FINER-DPO occur on *wh* pairs (Table 1): Qwen3-VL improves by 11.37 points on CompreCap and 10.59 on DOCCI, while InternVL improves by 7.45 and 5.64 points. Other settings are mixed: Qwen3-VL trails DPO on objects and relations, and InternVL also loses ground on several non-*wh* settings. These are gains in direct answering, without inference-time verification.

4.3 Ablation Studies

We replace the question-derived teacher context with a generic verification reminder (A), keeping the other training components and ordered 60K pool fixed. Both models answer directly at evaluation; full protocols appear in Appendix E.

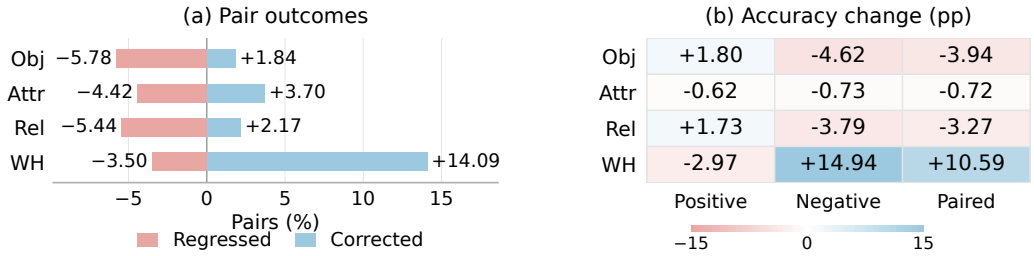

Figure 4: Pair-level corrections and regressions (a) and accuracy differences by query polarity (b) on FINER-DOCCI.

Table 3: General multimodal benchmark scores (%) for Qwen3-VL and InternVL under the adopted evaluation protocols.

Method	MMStar ↑	ChartQA ↑	MMVP ↑	NaturalBench ↑	V* ↑
Qwen3-VL-8B-Instruct					
Base	63.93$_{\uparrow 3.20}$	83.80$_{\downarrow 0.28}$	86.33$_{\downarrow 1.00}$	39.58$_{\downarrow 0.58}$	81.15$_{\uparrow 3.14}$
FINER-DPO 60K	65.27$_{\uparrow 1.86}$	83.44$_{\uparrow 0.08}$	86.00$_{\downarrow 0.67}$	40.47$_{\downarrow 1.47}$	81.15$_{\uparrow 3.14}$
Ours	67.13†	83.52†	85.33	39.00	84.29‡
InternVL3.5-8B-HF					
FINER-DPO 60K	41.73$_{\downarrow 0.73}$	86.84$_{\downarrow 1.52}$	76.67$_{\downarrow 1.67}$	16.68$_{\downarrow 1.53}$	70.16$_{\downarrow 0.52}$
Ours	41.00	85.32	75.00	15.16	69.63

Full exceeds A in all eight settings, most strongly on *wh* (+5.62 points on CompreCap; +3.73 on DOCCI; Figure 3). This single-seed comparison favors question-specific guidance over the tested reminder but does not isolate prompt wording, sampling, or reference effects. Complete scores are in Table 7.

4.4 Further Evaluation and Analysis

External factuality results are also mixed (Table 2). Qwen3-VL gains on DASH-B and AMBER relations, while InternVL gains on AMBER relations but loses on CRPE_R and DASH-B. Thus, the improvement on FINER *wh* does not imply a uniform factuality gain.

For Qwen3-VL on DOCCI, saved-prediction analysis (Figure 4) finds 1,475 corrected and 366 regressed *wh* pairs relative to FINER-DPO, a net gain of 1,109 correct pairs. Regressions outweigh corrections for the other three types. Here and in Figure 5, Full denotes Ours.

On *wh*, negative-query accuracy rises by 14.94 points while positive-query accuracy falls by 2.97 points (Figure 4b). The gain therefore coexists with over-rejection of valid questions; condition-count results appear in Appendix C.

On general capability benchmarks (Table 3), Qwen3-VL's displayed mean is 71.85%, versus 71.27% for DPO; its direct mean before evaluation-stage processing was 68.44%. Normalization affects two scores and additional visual verification affects V*. Accordingly, these scores describe the adopted evaluations, not a controlled demonstration of capability retention; Appendix B reports the direct results and processing details.

During training, divergence to the guided teacher decreases, while divergence to the original-input reference increases (Figure 5). For *wh*, endpoint JSD changes from 0.0741 to 0.0378 and from 0.0383 to 0.0942, respectively.

These curves describe evolving training trajectories, not independent evidence of reduced hallucination or capability preservation.

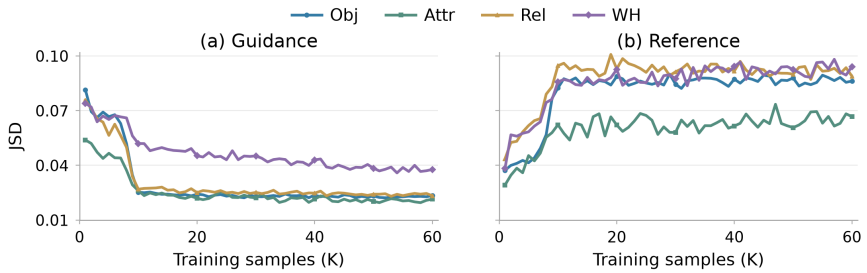

Figure 5: Training-time Jensen–Shannon divergence to the question-guided teacher (a) and original-input reference (b).

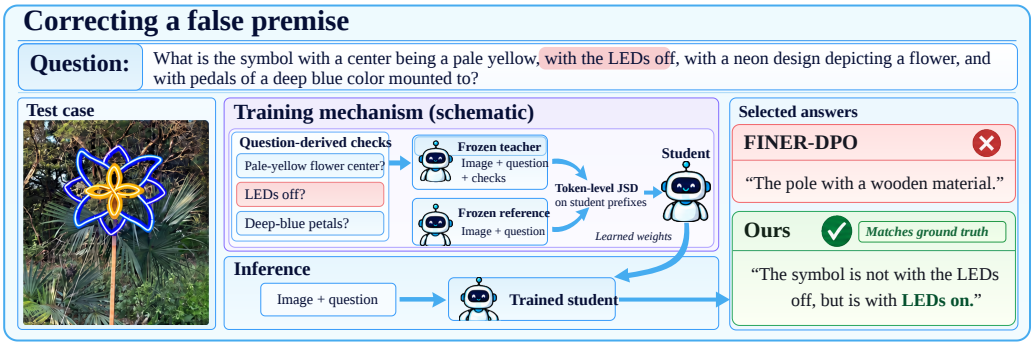

Figure 6: Correcting a false premise about an illuminated sign. FINER-DPO selects the wooden-pole answer, while Ours identifies that the LEDs are on. The schematic illustrates question-guided training and direct student inference.

4.5 Qualitative Case Study

Figure 6 shows Ours correcting the false premise that an illuminated sign has its LEDs off, alongside a schematic of question-guided training and direct inference.

Two verified DOCCI cases illustrate the same trade-off (Figure **??** in Appendix D): Ours corrects a false premise about unlit LEDs but wrongly rejects a true raised-arm premise. They illustrate selected multiple-choice decisions, not response explanations or prevalence estimates.

5 Conclusion and Limitations

Question-guided self-distillation improves wh paired accuracy relative to our FINER-DPO baseline on two existing fine-grained hallucination datasets, using only the original image and question at deployment. Qwen3-VL pair-level analysis locates the benefit in improved responses to negative queries and reveals a simultaneous loss on valid questions. This makes selective correction, rather than rejection alone, a concrete target for further improvement.

The current evidence covers two 8B backbones and the completed training runs. The teacher-context ablation favors question-specific guidance over one fixed generic reminder, but does not separate prompt structure, length, or wording effects. Independent contributions of current-student sampling and the reference objective remain unresolved. Training recipes are not identical across methods, and DOCCI was used during development. Improvements are uneven across query types. General-capability results include the evaluation-stage updates documented in Table 3; the original direct scores are lower, and the updated comparison uses different processing conditions across methods. The debugging procedures used benchmark feedback. Rule-based question transformation is also tied to the available query forms. These limits bound both the empirical result and claims about its mechanism; multi-seed uncertainty and transfer to broader natural question distributions remain untested.

AI USE STATEMENT

Generative tools were used only for language editing and LaTeX formatting during manuscript preparation. The authors reviewed the problem formulation, question transformation rules, training procedure, experimental execution, saved predictions, analyses, and conclusions; no generative tool was used as a substitute for those scientific decisions or for verification of the reported results.

REPRODUCIBILITY STATEMENT

The manuscript specifies the question-guided transformation rules, model configurations, optimization settings, data splits, evaluation prompts, paired-accuracy protocol, and prediction-audit procedures used in the experiments. The study uses the released FINER-CompreCap and FINER-DOCCI benchmarks together with the documented external evaluation sets. Appendix A records the Qwen3-VL and InternVL3.5 training settings, while the supplementary tables and audit records identify the retained samples, comparison conditions, and known limitations of the available artifacts.

ETHICS STATEMENT

This work studies multimodal hallucination mitigation using publicly released benchmark images, questions, and annotations, together with publicly available vision-language models. It does not collect human-subject data, personal information, or new human annotations. The question-derived checks are used during training to expose factual conditions in existing queries; they are not presented as independent human judgments. We report both improvements and regressions, retain the valid-answer trade-off in the analysis, and state the limits of the benchmark coverage and single-seed evidence.

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. *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. OPD-V: Visual On-Policy Self-Distillation with Modality Balance. *arXiv preprint arXiv:2608.05131*, 2026. https://arxiv.org/abs/2608.05131.

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. DASH: Detection and Assessment of Systematic Hallucinations of VLMs. *arXiv preprint arXiv:2503.23573*, 2025. https://arxiv.org/abs/2503.23573.

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

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning Transfer from LLMs to Vision-Language Models via On-Policy Distillation. *arXiv preprint arXiv:2510.23497*, 2025. https://arxiv.org/abs/2510.23497.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*, 2024. https://arxiv.org/abs/2403.20330.

Zihan Chen, Hengguang Zhou, Yuan Kang, Yiming Zhang, Wenhui Fang, Zenghui Ding, Yining Sun, and Cho-Jui Hsieh. BAS-OPD: Budget-Aware Selective On-Policy Self-Distillation for Fine-Grained Multimodal Perception. *arXiv preprint arXiv:2609.25891*, 2026. https://arxiv.org/abs/2609.25891.

Minghui Chen, Chenxu Yang, Hengjie Zhu, Dayan Wu, Zheng Lin, and Qingyi Si. Online Self-Calibration Against Hallucination in Vision-Language Models. *arXiv preprint arXiv:2605.00323*, 2026. https://arxiv.org/abs/2605.00323.

Zhisheng Chen, Jinhan Li, Yuxuan Li, Yuan Gao, Hao Wu, Zheng Lu, Jinlong Du, Kun Wang, and Bo An. VeinCast: Physics-Guided Dynamic Field Graphs with Graph-Conditioned Fusion for Global Medium-Range Weather Forecasting. *arXiv preprint arXiv:2608.09286*, 2026. https://arxiv.org/abs/2608.09286.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or Vision: Do Vision-Language Models Have Blind Faith in Text? *arXiv preprint arXiv:2503.02199*, 2025. https://arxiv.org/abs/2503.02199.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs. *arXiv preprint arXiv:2501.16629*, 2025. https://arxiv.org/abs/2501.16629.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking On-Policy Distillation of Large Language Models II: One Training Example. *arXiv preprint arXiv:2609.04172*, 2026. https://arxiv.org/abs/2609.04172.

Zhiyu Gui, Kexin Huang, Jia Guo, Junkang Wu, Zihao Wang, Zhiqiang Zhang, Jun Zhou, Jiancan Wu, and Xiang Wang. Know When to Stop, Where to Restart: Accelerating Multi-Turn Agentic On-Policy Distillation. *arXiv preprint arXiv:2609.14636*, 2026. https://arxiv.org/abs/2609.14636.

Jiawei Kong, Hao Fang, Shunxiang Liao, Jinyu Li, Bin Chen, Hao Wu, Shu-Tao Xia, and Min Zhang. Reasoning Matters: Mitigate Hallucination in Multimodal Large Reasoning Models via Reasoning-Conditioned Preference Optimization. *arXiv preprint arXiv:2605.27906*, 2026. https://arxiv.org/abs/2605.27906.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*, 2023. https://arxiv.org/abs/2305.10355.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. *arXiv preprint arXiv:2410.14669*, 2024. https://arxiv.org/abs/2410.14669.

Jidong Li, Lingyong Fang, Haodong Zhao, Sufeng Duan, and Gongshen Liu. Judge Before Answer: Can MLLM Discern the False Premise in Question? *arXiv preprint arXiv:2510.10965*, 2025. https://arxiv.org/abs/2510.10965.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-Policy Self-Distillation without Any Supervision. *arXiv preprint arXiv:2608.06296*, 2026. https://arxiv.org/abs/2608.06296.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe. *arXiv preprint arXiv:2604.13016*, 2026. https://arxiv.org/abs/2604.13016.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Zesheng Yang, Fangzhi Xu, and Jun Liu. Visual-OPSD: Cross-Modal On-Policy Self-Distillation for Efficient Unified Multimodal Reasoning. *arXiv preprint arXiv:2606.18974*, 2026. https://arxiv.org/abs/2606.18974.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. Unveiling the Ignorance of MLLMs: Seeing Clearly, Answering Incorrectly. *arXiv preprint arXiv:2406.10638*, 2024. https://arxiv.org/abs/2406.10638.

Jiayuan Ma, Yuqi Lu, Weiyang Guo, Chenrui Wang, Junyi Shu, Xuebo Liu, Min Zhang, and Jing Li. FPCO-Dialog: A Multi-Turn False-Premise Benchmark for Correction and Cooperation in Vision-Language Models. *arXiv preprint arXiv:2609.03331*, 2026. https://arxiv.org/abs/2609.03331.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. *arXiv preprint arXiv:2203.10244*, 2022. https://arxiv.org/abs/2203.10244.

Tengjun Ni, Xin Yuan, Shenghong Li, Kai Wu, Ren Liu, Wei Ni, and Wenjie Zhang. BOLT: Decision-Aligned Distillation and Budget-Aware Routing for Constrained Multimodal QA on Robots. *International Conference on Learning Representations*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/054a9c9f6249eee0093a85ccee5b3313-Abstract-Conference.html.

Chenhao Qiu, Dawei Li, Yechao Zhang, Lei Gong, and Zhen Tan. OPD-Aha: From Linguistic Momentum to Visual Reflection in Multimodal On-Policy Distillation. *arXiv preprint arXiv:2609.16459*, 2026. https://arxiv.org/abs/2609.16459.

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. Senses Wide Shut: A Representation-Action Gap in Omnimodal LLMs. *arXiv preprint arXiv:2605.13737*, 2026. https://arxiv.org/abs/2605.13737.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*, 2023. https://arxiv.org/abs/2305.18290.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting Hallucination in Vision-LLMs via Claim Decomposition and Program of Thought Verification. *arXiv preprint arXiv:2407.02352*, 2024. https://arxiv.org/abs/2407.02352.

Xinrui Shi, Kai Liu, Ziqing Zhang, Jianze Li, Anqi Li, and Yulun Zhang. DRScaffold: Boosting Dense-Scene Reasoning in Lightweight Vision Language Models. *arXiv preprint arXiv:2605.26038*, 2026. https://arxiv.org/abs/2605.26038.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. *arXiv preprint arXiv:1904.08920*, 2019. https://arxiv.org/abs/1904.08920.

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*, 2023. https://arxiv.org/abs/2309.14525.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. *arXiv preprint arXiv:2401.06209*, 2024. https://arxiv.org/abs/2401.06209.

Shravan Venkatraman, Omkar Thawakar, Ritesh Thawkar, Abdelrahman Shaker, and Rao Muhammad Anwer. Perception Before Supervision: Self-Contained Visual Distillation from Counterfactual Blind Spots. *arXiv preprint arXiv:2608.09931*, 2026. https://arxiv.org/abs/2608.09931.

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. https://arxiv.org/abs/2311.07397.

Zhecan Wang, Garrett Bingham, Adams Yu, Quoc Le, Thang Luong, and Golnaz Ghiasi. HaloQuest: A Visual Hallucination Dataset for Advancing Multimodal Reasoning. *arXiv preprint arXiv:2407.15680*, 2024. https://arxiv.org/abs/2407.15680.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The All-Seeing Project V2: Towards General Relation Comprehension of the Open World. *arXiv preprint arXiv:2402.19474*, 2024. https://arxiv.org/abs/2402.19474.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. https://arxiv.org/abs/2508.18265.

Penghao Wu, and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023. https://arxiv.org/abs/2312.14135.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. *arXiv preprint arXiv:2412.11124*, 2024. https://arxiv.org/abs/2412.11124.

Yike Wu, Yiwei Wang, and Yujun Cai. ChainMPQ: Interleaved Text-Image Reasoning Chains for Mitigating Relation Hallucinations. *arXiv preprint arXiv:2510.06292*, 2025. https://arxiv.org/abs/2510.06292.

Hao Wu, Yongheng Zhang, Yuan Gao, Fan Xu, Fan Zhang, Ruobing Xie, Ruijian Gou, Yuxuan Liang, Xiaomeng Huang, and Xian Wu. OMNIFLOW: A Physics-Grounded Multimodal Agent for Generalized Scientific Reasoning. *arXiv preprint arXiv:2603.15797*, 2026. https://arxiv.org/abs/2603.15797.

Hao Wu, Yuqi Li, Yuan Gao, Fan Xu, Fan Zhang, Kun Wang, Penghao Zhao, Qiufeng Wang, Yizhou Zhao, Weiyan Wang, Yingli Tian, Xian Wu, and Xiaomeng Huang. RoboAlign-R1: Distilled Multimodal Reward Alignment for Robot Video World Models. *arXiv preprint arXiv:2605.03821*, 2026. https://arxiv.org/abs/2605.03821.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. *arXiv preprint arXiv:2309.17453*, 2023. https://arxiv.org/abs/2309.17453.

Rui Xiao, Sanghwan Kim, Yongqin Xian, Zeynep Akata, and Stephan Alaniz. FINER: MLLMs Hallucinate under Fine-grained Negative Queries. *arXiv preprint arXiv:2603.17662*, 2026. https://arxiv.org/abs/2603.17662.

Feng Xiong, Leyan Xue, and Hongyu Lin. Correcting What You Cannot See: Credit Assignment for Perception Distillation in Multimodal Reasoners. *arXiv preprint arXiv:2607.28336*, 2026. https://arxiv.org/abs/2607.28336.

Qinwu Xu. Grounded Post-Training with Hard Examples for Reducing Hallucination in Multimodal Large Language Models. *arXiv preprint arXiv:2605.16411*, 2026. https://arxiv.org/abs/2605.16411.

Binhan Xu, Ligang Chang, Chenrui Yang, Jiaqi Xu, Kaiqi Qian, Jingze Chen, and Yanhui Du. Premise blindness in large vision-language models: Understanding and mitigating false-premise hallucination. *Information Sciences, article 124045*, 2027. https://www.sciencedirect.com/science/article/pii/S002002552600976X.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key. *arXiv preprint arXiv:2501.09695*, 2025. https://arxiv.org/abs/2501.09695.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-Policy Context Distillation for Language Models. *arXiv preprint arXiv:2602.12275*, 2026. https://arxiv.org/abs/2602.12275.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence Aware Heterogeneous Multi-Teacher Multimodal On-policy Distillation. *arXiv preprint arXiv:2607.02592*, 2026. https://arxiv.org/abs/2607.02592.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed On-Policy Distillation for Vision-Language Reasoning: Steering Gradients for Visual Grounding. *arXiv preprint arXiv:2606.00564*, 2026. https://arxiv.org/abs/2606.00564.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. DOPD: Dual On-policy Distillation. *arXiv preprint arXiv:2606.30626*, 2026. https://arxiv.org/abs/2606.30626.

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. https://arxiv.org/abs/2605.18740.

Yu Zhang, Chuyang Sun, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. Mitigating Multimodal Hallucination via Phase-wise Self-reward. *arXiv preprint arXiv:2604.17982*, 2026. https://arxiv.org/abs/2604.17982.

Fan Zhang, Vireo Zhang, Shengju Qian, Haoxuan Li, Zheng Lian, Hao Wu, Yuan Gao, Xinyu Geng, Xin Wang, and Pheng-Ann Heng. Struct-Searcher: Agentic Structural Thinking Advances Multimodal Deep Information Seeking. *arXiv preprint arXiv:2606.07689*, 2026. https://arxiv.org/abs/2606.07689.

Yongheng Zhang, Guang Yang, Ruihan Hou, Qiguang Chen, Ziang Liu, Xiaolong Liu, Manman Zhang, Yanchao Hao, Zheng Wei, Hao Wu, Libo Qin, Peishan Dai, Yinghui Li, Di Yin, and Xing Sun. Thinking in Video: Can Video Generators Really Reason About the Real World? *arXiv preprint arXiv:2607.17523*, 2026. https://arxiv.org/abs/2607.17523.

Huatian Zhang, Zhendong Mao, Lei Zhang, and Yongdong Zhang. Uncertainty-Aware Exploratory Direct Preference Optimization for Multimodal Large Language Models. *arXiv preprint arXiv:2605.04874*, 2026. https://arxiv.org/abs/2605.04874.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: Attributing Visual Evidence for Target Reconstruction in Multimodal On-Policy Distillation. *arXiv preprint arXiv:2607.28590*, 2026. https://arxiv.org/abs/2607.28590.

Fan Zhang, Guangming Yao, Jinyang Wu, Hao Wu, Zheng Lian, Xinyu Geng, Jingdong Chen, Yi Yuan, and Pheng-Ann Heng. VideoGAIA: A Benchmark for General AI Assistants on Agentic Video Understanding. *arXiv preprint arXiv:2608.14718*, 2026. https://arxiv.org/abs/2608.14718.

Tong Zhao, Yuyang Hu, Yutao Zhu, Reed Li, Haijin Liang, Haibo Shi, Yu Lu, and Zhicheng Dou. OPOD: On-Policy Omni Distillation. *arXiv preprint arXiv:2607.20918*, 2026. https://arxiv.org/abs/2607.20918.

Zhixiao Zheng, Zheren Fu, Zhiyuan Yao, Dongming Zhang, and Zhendong Mao. CatPO: Cross-modal Adaptive Token-rewards for Preference Optimization in Truthful Multimodal LLMs. *International Conference on Learning Representations*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/b8bad5b0cc42e1ca0d47760e0d9f9449-Abstract-Conference.html.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. *arXiv preprint arXiv:2407.00569*, 2024. https://arxiv.org/abs/2407.00569.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: Resolution-Privileged On-Policy Self-Distillation for Multimodal Large Language Models. *arXiv preprint arXiv:2607.24447*, 2026. https://arxiv.org/abs/2607.24447.

A Implementation Details

The complete model uses Qwen3-VL-8B-Instruct with LoRA rank 32, scaling parameter 64, zero dropout, and adapters on query and value projections. The implementation uses bfloat16, AdamW with learning rate 10^{-4}, gradient accumulation over eight records, and seed 0. The 60K records are shuffled before a single traversal. Student rollouts use temperature 1.0 and a 32-token generation cap; the image processor has a maximum pixel budget of 401,408. Teacher and reference distributions are produced with adapters disabled. Modules remain in evaluation mode during training while adapter parameters receive gradients. The main loss uses the full vocabulary; top-16 statistics are diagnostic only.

The guidance prompt asks for verification of a numbered list of subquestions before the original query. Specific rules handle objects, attributes, relations, and wh descriptions; fallback rules split text or retain a single verification unit. This transformation is not an additional visual model or an independently supervised fact extractor. The reference loss weight is 0.5 and the JSD mixture coefficient is 0.5. For each record, gradients pass through the student distribution conditioned on its sampled response prefix, not through sequence sampling or the frozen guidance distributions.

For InternVL3.5-8B-HF, Ours retains the same loss weights, LoRA rank/scaling, learning rate, accumulation, seed, rollout temperature, and response cap. Adapters are restricted to language-model query and value projections; the vision tower and projector remain frozen. Training images use the FINER InternVL area cap of 262,144 pixels before native processing. The 60,000 records contain 7,500 examples per query-type/polarity combination. InternVL FINER-DPO uses rank 32, learning rate 5×10^{-6}, $\beta = 0.1$, microbatch size 4 and accumulation 16 (global batch 64), completing one epoch in 933 updates on 59,700 training records. These settings and the data pools differ across recipes.

B Evaluation Details and Additional Results

The main FINER evaluation follows the released task definitions and multiple-choice scoring (Xiao et al., 2026). The released DOCCI attribute split contains 14,315 pairs; we use the released data rather than adding samples to reproduce the paper's reported count of 14,324. No historical internal model variant or inference-time answer correction is substituted in the FINER comparison. The general-capability processing reported in Table 3 is described separately below.

For Qwen3-VL, Ours processes 60,000 training records; FINER-DPO uses 59,880 training and 120 validation records from its 60K input pool. Their hyperparameters differ, and identical training pools have not been established. FINER-CompreCap evaluation uses a two-token generation cap. On DOCCI, Ours and DPO use three tokens, while the historical Base run uses two († in Table 1).

Table 4: Additional hallucination benchmark scores (%) for Qwen3-VL and InternVL under the adopted protocols.

Method	POPE ↑	RePOPE ↑	AMBER ↑	HallusionBench ↑
Qwen3-VL-8B-Instruct				
Base	88.73	94.12	89.09	$56.15^{\dagger}$
FINER-DPO 60K	88.81	94.22	89.31	75.60
Ours	88.57	93.96	89.58	74.87
InternVL3.5-8B-HF				
FINER-DPO 60K	85.92	88.46	83.06	68.87
Ours	86.36^{*}	87.54^{*}	82.67	69.72

For InternVL, Ours processes 60,000 records and DPO uses 59,700 training and 300 validation records from the official first-60K pool; these pools are not sample-matched and differ from FINER's reported 160K-scale training. Internal evaluation uses four generated tokens for Ours and three for DPO. These comparisons measure complete recipes under their recorded settings, not isolated objective effects.

The external evaluations in Table 2 contain 7,628 AMBER attribute queries, 1,664 AMBER relation queries, 1,894 CRPE_R groups (7,576 queries), and 2,682 DASH-B queries. CRPE_R counts a group as correct only if all four answers are correct. It is not the arithmetic mean of its subject, predicate, and object category scores.

Table 3 uses MMStar overall accuracy (1,500 queries), ChartQA test relaxed accuracy (2,500), MMVP accuracy (300), NaturalBench group accuracy (1,900 groups, 7,600 queries), and V* accuracy (191). In that table, † marks answer normalization and ‡ marks additional visual verification. For Qwen3-VL, Base and DPO retain their original direct results. For Ours, MMStar uses final-option normalization of frozen answers (1,007/1,500 correct); and ChartQA uses format post-processing v2 with the original relaxed-accuracy scorer (2,088/2,500 correct). These two updates require no retraining or fresh answer generation. MMVP retains 256/300 correct. NaturalBench explicit-conclusion normalization changes no answers and retains 741/1,900 correct groups. V* uses an additional visual-localization and verification module with frozen weights, obtaining 161/191 correct; it adds 868 model calls and is not a format-only correction. The original FINER scoring functions remain in use, including exact-matching fallbacks where applicable. The normalizers do not receive ground-truth answers, and visual verification does not use them to select predictions; however, the processing designs used benchmark-error feedback and are exploratory rather than independent held-out evaluations.

For Qwen3-VL before these evaluation-stage updates, the direct Ours scores in the same column order were 63.40, 76.48, 85.33, 39.00, and 78.01%, with a descriptive mean of 68.44%. Thus the updated mean of 71.85% includes processing effects and cannot be attributed solely to training. Base and DPO are not evaluated with the same complete processing configuration in Table 3, and there is no Base comparison with the V* verification module. These protocol differences preclude interpreting the displayed means as a controlled demonstration of capability retention.

For Table 4, POPE and RePOPE are means across three accuracy subsets, AMBER is a three-category accuracy mean, and HallusionBench uses aAcc; all are reported as percentages. The $*$ on InternVL POPE/RePOPE denotes the regeneration protocol described below.

For Qwen3-VL, † denotes rule-only parsing; DPO and Ours use assisted parsing for HallusionBench, so the historical Base score is not a like-for-like comparator. The DPO and current-model HallusionBench results contain 719 and 712 correct answers out of 951, respectively. POPE and RePOPE cover 8,910 and 8,185 queries; AMBER covers 14,216 queries, including the two subsets in Table 2.

For Qwen3-VL, HaloQuest scores are 83.17% for DPO and 83.33% for our model, with false-premise subset scores of 89.11% and 89.44%. Both use the recorded Gemini-2.5-Flash judge, differing from the original FINER judge version. Generation budgets also differ: DPO uses a 4,096-token cap throughout; our run initially uses 512 tokens, with two truncated responses regenerated at 4,096. These results are reported as protocol-qualified observations, not evidence of a controlled gain. DPO

Table 5: DOCCI query-pair outcome counts for FINER-DPO and Ours.

Setting	Pairs	Both correct	Corrected	Regressed	Both wrong
Obj	5,000	2,864	92	289	1,755
Attr	14,315	8,822	530	633	4,330
Rel	5,771	3,086	125	314	2,246
wh	10,472	3,448	1,475	366	5,183

has a recorded MMHal score of 5.56/6 and hallucination rate of 7.29%; no verified current-model score is available, so no MMHal comparison is made.

InternVL evaluation protocols and remaining metrics. The adopted InternVL results use the original 60K adapters. For Table 3, ChartQA, MMVP, and V* use the same native image-first input for both methods, with the redundant literal image marker removed, original benchmark instructions retained, and official scoring without custom output postprocessing. MMVP reports per-question accuracy, not paired accuracy. MMStar and NaturalBench retain their historical configurations, including final-option and explicit-conclusion normalization for Ours. Thus the five-benchmark means describe mixed evaluation protocols. A separately completed unified-native comparison and an additionally trained recovery variant are not substituted into these adopted rows.

In Table 4, InternVL Ours regenerates only historically unparsed POPE/RePOPE responses with a 128-token cap, retaining all other responses; DPO keeps its original three-token outputs. The previous Ours scores were 84.07% and 84.87%. These changes do not alter model weights and cannot be attributed to training. InternVL HallusionBench uses rule-recognized outputs for both models, with no auxiliary judge calls.

InternVL CRPE_R subject/predicate/object group accuracies are 63.02/48.34/47.24% for DPO and 54.96/44.93/45.09% for Ours. HaloQuest overall accuracy is 70.63% versus 71.45%, respectively; false-premise accuracy is 73.93% for both. Both HaloQuest runs use Gemini-2.5-Flash rather than the original FINER judge version. MMHal scores are 4.45 versus 4.75 out of six, with hallucination rates of 19.79% versus 15.62%, respectively. MMHal uses GPT-4.1-mini, consistent with FINER Appendix D, on 96 questions; the reference-based judge receives text rather than the image, and scores below three count as hallucinations. No multi-seed significance claim is made.

C Prediction and Training-Diagnostic Audit

The counts and training diagnostics in this section concern Qwen3-VL.

Number of queried conditions. Condition-count grouping does not support a growing advantage as more conditions are included. On DOCCI object queries, the two models tie at one condition, while our model trails DPO for counts two through six. At six conditions, DPO achieves 58.35% paired accuracy and our model achieves 53.34%. Attribute and relation results also lack a consistent increasing advantage. Counts come from the released query types; the groups differ in composition and size. This is an observational comparison, not a controlled measure of reasoning difficulty or training-data scaling.

The following table partitions each DOCCI query type into pairs jointly correct under both models, corrected only by Ours, regressed from FINER-DPO, and jointly incorrect under both models.

The paired wh gain is $(1475-366)/10472 = 10.59$ percentage points. Positive-query correct counts are 8,755 for DPO and 8,444 for Ours; negative-query counts are 4,827 and 6,391. Both wh prediction files contain no invalid choice indices. The comparison checks 142,232 saved predictions across both models for matching question, image, type, options, and target index; each paired comparison uses the same released samples.

The next table reports unweighted component JSD means from the first and last recorded bins, ending at 1K and 60K processed training records, respectively.

Table 6: Mean guidance and reference JSD in the first and last recorded training bins.

	Guidance JSD		Reference JSD	
Setting	First	Last	First	Last
Obj	0.0815	0.0234	0.0373	0.0863
Attr	0.0539	0.0215	0.0292	0.0667
Rel	0.0757	0.0234	0.0435	0.0889
wh	0.0741	0.0378	0.0383	0.0942

Case Study: Faded Street Markings



Question: Can you see the words with the text 'FIRE LANE', with a heavily faded appearance, and with white paint in this image?

Our work:

1. Can words with the text "FIRE LANE" be seen in the image?
2. Do the words have a heavily faded appearance?
3. Are the words rendered with white paint?

Ours / GT: Yes, I can see the words with the text 'FIRE LANE', with a heavily faded appearance, and with white paint in this image.

Figure 7: Question-derived guidance for a faded street marking. The answer matches the ground truth.

Mean response lengths in these endpoint bins are close to the 32-token cap. This is relevant to interpretation: optimization statistics alone do not establish successful completion of visual verification. The 240 setting-by-bin rows satisfy the recorded total-loss formula $(D_H + 0.5D_R)/1.5$.

D Qualitative Case Provenance

The selected examples use official DOCCI visualization previews. Images are credited to Google LLC under CC BY 4.0. The local source manifest records the image URL, dimensions, and SHA256 for each downloaded preview. Pair indices in Figure **??** are zero-based. Selected options are taken from the saved multiple-choice outputs and are not presented as generated reasoning.

For test_00014, the negative question describes the illuminated sign as having its LEDs off. DPO selects "The pole with a wooden material." Our model selects "The symbol is not with the LEDs off, but is with LEDs on." Both models choose the wooden-pole answer for the positive question.

For test_00002, the positive question asks what is spread in front of the woman with her left arm raised. DPO selects the benchmark's correct option, "The seahorse with its mouth wide open." Our model selects "The woman is not with her left arm raised up, but is with her left leg raised up." Visual inspection supports the raised-arm premise; the failure interpretation concerns this premise, not an independent taxonomy of the surrounding statues. Both models correct the raised-leg premise in the negative question.

E Teacher-Context Ablation Details

Table 7 reports the completed Qwen3-VL teacher-context comparison separately from the main baseline results. Full is the current complete model, not a historical T-Hint variant. A changes the teacher's additional context to the following fixed instruction, followed by two newlines and the original question:

Case Study: Weathered Stone Walkway



Question: Can you see the walkway with white stones, with broken steps, with some green and pink growth patches, with a dated stone material, and with large, flat, layered stones in this image?

Our work:

1. Can a walkway with white stones be seen in the image?
2. Does the walkway have broken steps?
3. Are some green and pink growth patches visible?
4. Is the walkway made of a dated stone material?
5. Does it contain large, flat, layered stones?

Ours: Yes, I can see the walkway with white stones, with broken steps, with some green and pink growth patches, with a dated stone material, and with large, flat, layered stones in this image.

Figure 8: Question-derived guidance for the attributes of a weathered stone walkway.

Case Study: Spatial Relation Between Boats



Question: Can you see the boat that is to the left of the boat in this image?

Our work:

1. Can a boat be seen in the image?
2. Is there another boat serving as the reference object?
3. Is the queried boat located to the left of the reference boat?
4. Does the image support "to the left of" rather than another spatial relation?

Ours / GT: Yes, I can see the boat that is to the left of the boat in this image.

Figure 9: Question-derived guidance for the spatial relation between two boats.

> Inspect the image carefully. Verify whether the facts and premises in the question are supported by the image before answering. Do not assume that a mentioned object, attribute, or relation is present or valid. Answer the original question based on the image.

Full instead prepends *First, verify each of the following sub-questions about the image:*, followed by the numbered outputs of the rule-based question transformation. Neither variant supplies answers to these checks. The full-vocabulary guidance and reference JSD terms retain coefficients $2/3$ and $1/3$, respectively. A is trained from the original initialization rather than continued from Full. It completes 60,000 records, 7,500 updates, and zero skipped records. The pool contains 7,500 records per query-type/polarity combination and 32,672 unique images. Initialization, data-pool and order identities, loss settings, and completion records were checked against the archived Full configuration. In Table 7, the Full $-$ A row reports percentage-point differences computed before rounding, and boldface marks the higher score in each matched setting.

Evaluation follows the frozen FINER prompts, option parser, and paired-accuracy computation, without training hints or output postprocessing. Generation caps are two tokens for CompreCap and three for DOCCI. DOCCI retains all 14,315 released attribute pairs and has prior development exposure. A's archived completion and evaluation records include adapter and prediction-file hashes, but its per-question outputs remain unavailable for independent rescoring. No student-sampling or reference-removal results are included.

Table 7: Paired accuracy (%) for question-specific (Full) and generic (A) teacher guidance on FINER-CompreCap and FINER-DOCCI.

	FINER-CompreCap				FINER-DOCCI			
Variant	Obj	Attr	Rel	wh	Obj	Attr	Rel	wh
Full	**75.17**	**77.11**	**66.21**	**57.99**	**59.12**	**65.33**	**55.64**	**47.01**
A: Generic guidance	74.41	73.76	63.97	52.37	58.56	62.49	53.47	43.28
Full − A (pp)	+0.76	+3.36	+2.24	+5.62	+0.56	+2.84	+2.17	+3.73
Evaluation pairs	3,150	1,669	2,140	1,583	5,000	14,315	5,771	10,472