EVIDENCE-ALIGNED MULTIMODAL ON-POLICY SELF-DISTILLATION FOR FINE-GRAINED VISUAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained visual understanding requires models to recognize small details within complex images. Multimodal on-policy self-distillation (OPSD) addresses this challenge by using a teacher conditioned on evidence-centered crops to supervise a student conditioned on original images along student-generated trajectories. Ideally, teacher corrections, the distributional changes from the student toward the privileged teacher, should be driven by task-relevant visual evidence. However, the designs that make the teacher effective also introduce other interference. Using a lagged or frozen teacher improves training stability but introduces a model-state gap from the evolving student, while cropping enhances task-relevant evidence but also loses the visual context. These two sources of interference make the teacher corrections not purely rely on the visual evidence. We introduce Evidence-Aligned multimodal on-policy self-Distillation (EAD), which retains the crop-conditioned teacher as the target but constructs a separate evidence reference for weighting the corrections. To exclude the effect of lagged model-state from this reference, EAD measures prediction changes using the current student. To avoid crop-induced context changes, EAD masks the evidence region in the original image while preserving the other visual context. The change from the student's masked-image prediction to its original-image prediction provides a controlled reference for the direction in which the visual evidence shifts the student's prediction. EAD weights each teacher correction by its cosine alignment with the reference, i.e., retaining aligned corrections and downweighting the rest. Retaining only 6% of the supervision mass of dense OPSD, EAD consistently outperforms previous state-of-the-art methods. We also show that with small-scale backbones, EAD achieves competitive performance compared to substantially larger open-weight and closed-source models.

1 INTRODUCTION

Fine-grained visual understanding tasks require multimodal large language models (MLLMs) to recognize small, localized evidence within large or cluttered visual contexts. However, reliably extracting and exploiting such evidence remains challenging, even for models with strong general-purpose visual and reasoning capabilities (Khayatkhoei et al., 2025; Liu et al., 2025; Wang et al., 2026). Recent work (Yuan et al., 2026) addresses this limitation with On-Policy Self-Distillation (OPSD), where the student generates trajectories from the original image and a crop-conditioned teacher, with privileged access to the visual evidence, provides next-token targets conditioned on the same prefixes. This encourages the student to make better use of fine-grained visual evidence from the original image, which has shown obvious gains on fine-grained visual understanding tasks.

The value of the teacher supervision comes from corrections driven by the task-relevant visual evidence exposed through the crop. However, the designs that make the teacher effective also introduce other interferences into these corrections. First, a lagged or frozen teacher is important for stable self-distillation. Vision-OPD (Yuan et al., 2026) shows that directly using the current student as the teacher can destabilize training. The trained model collapses to near-zero accuracy across all benchmarks. However, using a lagged or frozen teacher introduces a model-state gap, so part of the teacher correction can arise from differences in model state. Second, the cropping is essential for

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.

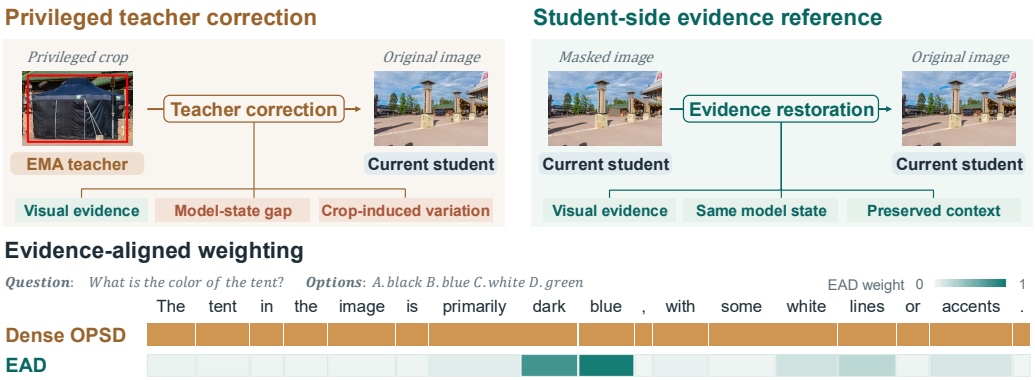

Figure 1: **Motivation and overview of EAD.** The privileged teacher correction in multimodal OPSD contains both the desired evidence-driven signal and interference arising from model-state gap and crop-induced context changes (top left). EAD constructs a separate student-side evidence reference from the current student's prediction shift induced by evidence restoration, which keeps the model state fixed and preserves the surrounding visual context (top right). EAD then weights teacher corrections by the alignment with this evidence reference, concentrating supervision on visually supported corrections (bottom).

exposing fine-grained visual evidence that may be difficult to extract from the original image, but cropping also removes surrounding visual context and changes the visual-token input presented to the model. This will introduce crop-induced variation into the teacher correction. Section 3 quantifies these two effects at the token level, showing that both model-state and crop-induced variations substantially alter the teacher corrections.

Therefore, we introduce Evidence-Aligned multimodal on-policy self-Distillation (EAD), as illustrated in Figure 1. EAD retains the crop-conditioned EMA teacher as the supervision target, while constructing a separate evidence reference to determine how much of each teacher correction to retain. To exclude model-state gap from this reference, EAD computes the prediction change of current student. To avoid crop-induced context changes, EAD masks only the evidence region in the original image while preserving the surrounding visual context. Under the same generation prefix, the change from the student's masked-image prediction to its original-image prediction provides a controlled reference for the direction in which the visual evidence shifts the student's prediction. EAD then compares each privileged teacher correction with this reference. Their directional agreement is measured by cosine similarity and directly determines the token-level supervision weight: downweighting weakly aligned corrections and excluding opposing ones from supervision.

Retaining only 6% of the supervision mass of dense OPSD, EAD consistently outperforms previous state-of-the-art methods. We also show that with small-scale backbones, EAD achieves competitive performance compared to substantially larger open-weight and closed-source models. These results suggest that OPSD is most effective when teacher corrections are selectively retained according to their alignment with privileged evidence.

Our contributions are (1) We identify two sources of interference in multimodal OPSD: model-state gap and crop-induced context loss. Our analysis shows that they substantially affect the teacher corrections. (2) We propose EAD, which retains the crop-conditioned EMA teacher as the supervision target and constructs a separate evidence reference from the shift in the current student's prediction caused by restoring the masked visual evidence. EAD uses the directional alignment between the teacher correction and the evidence reference to weight the distillation supervision. (3) Experiments show that EAD consistently outperforms previous state-of-the-art methods and is competitive with substantially larger models with retaining only about 6% of the supervision mass of dense OPSD.

2 Related Work

Fine-Grained Visual Understanding. Fine-grained visual understanding requires MLLMs to recover small, localized evidence from high-resolution or cluttered images. Existing methods often

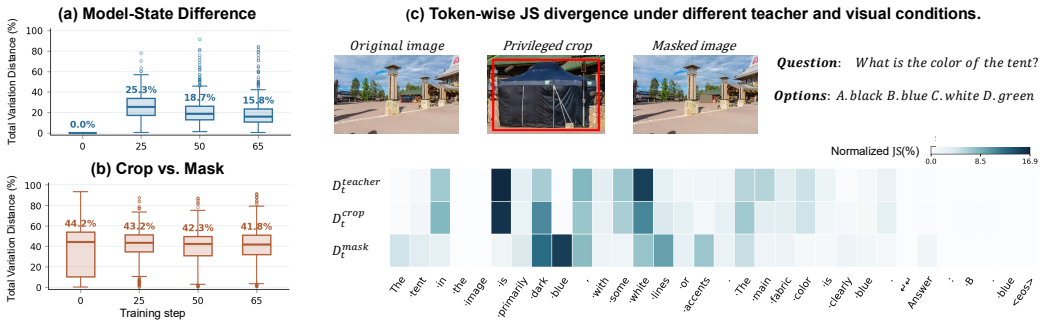

Figure 2: **Diagnosing interference in teacher corrections.** (a) Total Variation (TV) distance between the normalized token-wise JS distributions of D_t^{teacher} and D_t^{crop} over 512 images, showing the effect of model-state gap across training. (b) TV distance between D_t^{crop} and D_t^{mask} under the current student, showing the difference between crop-induced changes and evidence removal. (c) The corresponding token-wise JS distributions for one example. D_t^{mask} places greater emphasis on tokens associated with the task-relevant visual evidence. D_t^{crop} shows large divergence on less vision-relevant tokens.

enhance fine-grained perception at inference time through visual search, localization, or image manipulation (Wu & Xie, 2024; Khayatkhoei et al., 2025; Wang et al., 2025; Zheng et al., 2026; Zhang et al., 2025a). Recently, some approaches improve fine-grained perception through training without additional test-time operations. Zooming without Zooming (Wei et al., 2026) distills the benefits of zooming into full-image training, enabling single-pass fine-grained perception. Vision-OPD (Yuan et al., 2026) uses crop-based teacher supervision to improve recognition of fine details in the original image. Our work also aims to strengthen fine-grained perception without additional inference-time cropping or tool use.

On-Policy Self-Distillation. Knowledge distillation transfers prediction distributions from a teacher to a student (Hinton et al., 2015), and has been extended to autoregressive sequence generation (Kim & Rush, 2016). Recently, on-policy distillation (Gu et al., 2024; Agarwal et al., 2024) applies teacher supervision to prefixes generated by the student's own policy, reducing the mismatch between training and inference prefix distributions. On-policy self-distillation further uses the same model as both teacher and student, with the teacher conditioned on additional privileged information (Zhao et al., 2026). In multimodal on-policy self-distillation, Vision-OPD (Yuan et al., 2026) uses an evidence-centered crop-conditioned teacher to supervise an original-image student. Our work follows this setting but focuses on improving the quality of the teacher corrections.

Evidence-Aware Selective Distillation. Selective distillation selectively retains or reweights teacher supervision at the sequence or token level, rather than treating all teacher signals equally (Wang et al., 2021; Tavor et al., 2026). Visual interventions provide a way to identify how model predictions depend on visual evidence (Leng et al., 2024; Hu et al., 2025; Gao et al., 2026). Building on this idea, multimodal distillation methods use visual contrasts to select or weight supervision. Recent multimodal OPD methods use visual contrasts to identify supervision that is more closely tied to visual evidence (Liu et al., 2026; Qian et al., 2026; Sun et al., 2026). Concurrent works in multimodal OPSD explore evidence-aware selection of teacher supervision. OPD-V (Bi et al., 2026) uses contrastive privileged views to select visually supported supervision, while VAD (Zhang et al., 2026) uses teacher responses to degraded crop to identify visually relevant corrections. Our work adapts visual-evidence extraction to the particular structure of multimodal OPSD, yielding a more controlled signal for weighting teacher corrections.

3 Diagnosing Teacher Corrections

In Vision-OPD (Yuan et al., 2026), the current student parameterized by θ is conditioned on the original image R, while the EMA teacher parameterized by $\bar{\theta}$ is conditioned on the cropped image C. Both are evaluated under the same student-generated prefix $h_t = (x, y_{<t})$, where x denotes the

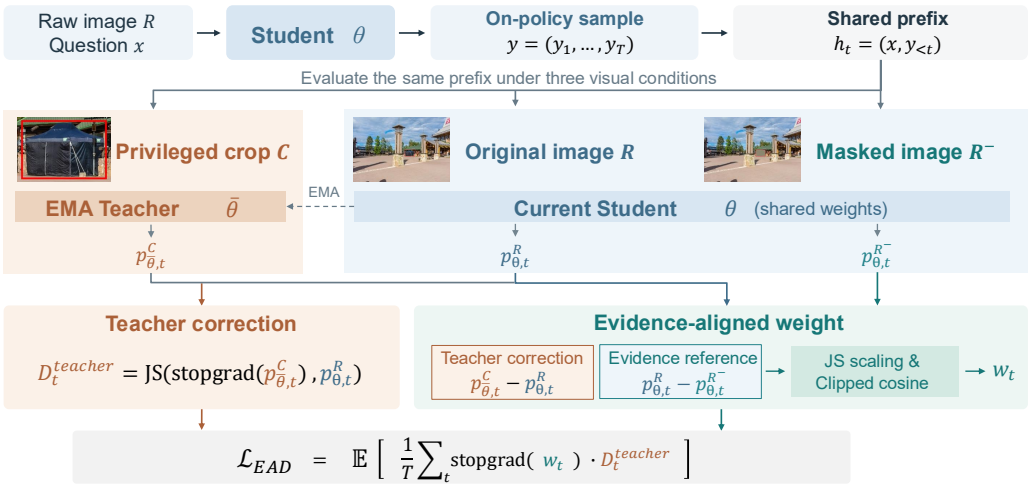

Figure 3: **Overview of EAD.** The crop-conditioned EMA teacher provides the privileged target $p^C_{\bar{\theta},t}$, while the current student produces $p^R_{\theta,t}$ and $p^{R^-}_{\theta,t}$ on the original and masked images, respectively. The Original–Mask prediction change $p^R_{\theta,t} - p^{R^-}_{\theta,t}$ forms the evidence reference. EAD compares the teacher correction and evidence reference after a shared JS-motivated normalization, and uses their clipped cosine similarity to weight the distillation loss.

input prompt and $y_{<t}$ the generated response tokens. For visual input I, we denote the next-token distribution as $p^I_{\theta,t} = p_\theta(\cdot \mid I, h_t)$. The crop-conditioned teacher provides $p^C_{\bar{\theta},t}$ as the target for the original-image student prediction $p^R_{\theta,t}$. Let sg denote the stop-gradient operator. The token-level distillation divergence is

$$D_t^{\text{teacher}} = \mathrm{JS}\left(\mathrm{sg}(p^C_{\bar{\theta},t}), p^R_{\theta,t}\right). \tag{1}$$

So, the teacher distillation divergence involves both a change in model state (θ versus $\bar{\theta}$) and visual input (R versus C), motivating a further diagnosis of how these factors influence the correction.

3.1 Model State Difference Alters Teacher Corrections

To isolate the effect of model state, we compare the teacher correction D_t^{teacher} with a single-state variation that uses the current student for both visual inputs:

$$D_t^{\text{crop}} = \mathrm{JS}\left(p^R_{\theta,t}, p^C_{\theta,t}\right). \tag{2}$$

To quantify how the two divergences differ across response tokens, we normalize their token-level JS values within each response and compute the Total Variation (TV) distance between the resulting distributions, with $\mathrm{TV} \in [0, 1]$:

$$\pi_t^k = \frac{D_t^k}{\sum_{t'} D_{t'}^k}, \qquad \mathrm{TV}\left(\pi^{\text{teacher}}, \pi^{\text{crop}}\right) = \frac{1}{2}\sum_t \left|\pi_t^{\text{teacher}} - \pi_t^{\text{crop}}\right|, \tag{3}$$

where $k \in \{\text{teacher}, \text{crop}\}$. We analyze a Qwen3.5-4B model trained with Vision-OPD (Yuan et al., 2026) using 512 randomly sampled images across training checkpoints. As shown in Figure 2(a), the two token-level distributions coincide at initialization and diverge during training. The median TV distance reaches 25.3% at step 25 and remains substantial throughout training, showing that model-state gap alters the allocation of supervision across response tokens.

3.2 Crop-Induced Change Differs from Evidence Removal

We next examine whether the prediction change induced by the privileged crop faithfully reflects the visual evidence. We decompose the original image as $R = (E, B)$, where E denotes the task-relevant evidence and B the remaining visual context. The Raw–Crop comparison changes both

components: cropping emphasizes E while removing B. To obtain a cleaner evidence reference, we construct a masked image $R^- = (E^-, B)$ which masks the evidence region while preserving the surrounding context. Using the same current student and generation prefix, we compute

$$D_t^{\text{mask}} = \text{JS}\Big(p_{\theta,t}^R, p_{\theta,t}^{R^-}\Big). \tag{4}$$

Following Eq. 3, we normalize D_t^{mask} and D_t^{crop} within each response and compute the TV distance between them. Figure 2(b) shows consistently large median TV distances of approximately 42–44% across training checkpoints. These results reveal a substantial token-level difference between crop-induced change and evidence removal, highlighting the influence of surrounding-context removal in the cropped view.

3.3 D_t^{mask} Emphasizes Visually Relevant Tokens

Figure 2(c) compares the token-wise JS distributions of the three supervision signals in a color-recognition example. Replacing the EMA teacher with the current student redistributes JS mass across response tokens. More importantly, D_t^{mask} places greater emphasis on tokens associated with the task-relevant visual evidence, such as `blue`, whereas D_t^{crop} shows large divergence on tokens less directly tied to the visual evidence, such as `is`. This suggests that the Original–Mask comparison is more closely tied to task-relevant visual evidence, motivating its use as the evidence reference for weighting teacher corrections. More case studies are provided in Appendix A.8.

4 Evidence-Aligned multimodal on-policy self-Distillation

Motivated by the analysis in Section 3, we introduce Evidence-Aligned multimodal on-policy self-Distillation (EAD). As illustrated in Figure 3, EAD retains $p_{\bar{\theta},t}^C$ as the privileged teacher target, while using the current student's Original–Mask prediction change $p_{\theta,t}^R - p_{\theta,t}^{R^-}$ as a separate evidence reference. EAD compares this reference with the teacher correction vector $p_{\bar{\theta},t}^C - p_{\theta,t}^R$ and uses their directional agreement to weight the distillation loss.

4.1 Aligning Teacher Corrections with Visual Evidence

At response step t, the crop-conditioned EMA teacher provides the privileged target distribution $p_{\bar{\theta},t}^C$ for the original-image student distribution $p_{\theta,t}^R$. EAD additionally computes the current student's next-token distribution conditioned on the masked image, denoted by $p_{\theta,t}^{R^-}$. Since $p_{\theta,t}^R$ and $p_{\theta,t}^{R^-}$ are produced by the same model under the same generation prefix and differ only in the evidence region, their distributional change provides a controlled reference for the prediction shift induced by the visual evidence.

For two distributions p and $p+\delta$, a second-order Taylor expansion of the Jensen–Shannon divergence around p gives

$$\text{JS}(p, p+\delta) \approx \frac{1}{8}\sum_j \frac{\delta_j^2}{p_j}. \tag{5}$$

where j indexes vocabulary entries. This motivates representing distributional changes in the locally normalized coordinates $\delta/\sqrt{p}$, so that the directional comparison is consistent with the geometry of Jensen–Shannon divergence. Based on the shared current student distribution $p_{\theta,t}^R$, we define

$$u_t^{\text{corr}} = \frac{p_{\bar{\theta},t}^C - p_{\theta,t}^R}{\sqrt{\max(p_{\theta,t}^R, \epsilon)}}, \qquad u_t^{\text{evid}} = \frac{p_{\theta,t}^R - p_{\theta,t}^{R^-}}{\sqrt{\max(p_{\theta,t}^R, \epsilon)}}. \tag{6}$$

u_t^{corr} represents the direction from the student distribution toward the privileged teacher target, while u_t^{evid} represents the prediction shift induced by the visual evidence. ϵ is a small numerical constant.

We quantify the directional agreement using the clipped cosine similarity:

$$w_t = \max\left(\frac{\langle u_t^{\text{corr}}, u_t^{\text{evid}}\rangle}{\|u_t^{\text{corr}}\|_2 \, \|u_t^{\text{evid}}\|_2}, 0\right), \tag{7}$$

Table 1: **Fine-grained visual understanding.** We report results on six fine-grained visual benchmarks under the single-forward-pass evaluation setting. Results for closed-source and open-weight models are taken from Yuan et al. (2026), while all Qwen3.5-4B/9B-based methods are obtained from our evaluation. All post-training methods use the same backbone and training data, are trained for one epoch, and are evaluated using the final checkpoint. Avg. is the unweighted mean over the six benchmarks. Best results within each Qwen3.5 scale are shown in bold.

Model	Params	V*Bench	ZoomBench	HR-Bench		MME-RealWorld		Avg.
				4K	8K	EN	CN	
Closed-Source Models (Single Forward Pass)								
GPT-5.2 (OpenAI, 2025)	–	79.06	50.89	81.12	78.38	72.60	68.80	71.81
GPT-5.4 (OpenAI, 2026)	–	76.96	52.66	84.00	77.88	74.20	70.93	72.77
Gemini-3.1-Pro (Google, 2026a)	–	87.96	61.18	89.63	86.88	76.53	73.31	79.25
Gemini-3.5-Flash (Google, 2026b)	–	89.01	61.42	89.12	86.62	75.31	73.97	79.24
Open-Weight Models (Single Forward Pass)								
DeepEyes (Zheng et al., 2026)	7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
Thyme (Zhang et al., 2025a)	7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyesV2 (Hong et al., 2026)	7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
SenseNova-MARS (Chng et al., 2025)	8B	92.15	47.81	83.13	78.38	67.90	68.90	73.05
MiMo-VL-RL (Yue et al., 2025)	7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
ZwZ (Wei et al., 2026)	8B	87.96	56.69	83.63	81.75	66.57	68.09	74.12
MiniCPM-V-4.5 (Yu et al., 2026)	9B	70.68	42.60	69.63	61.50	62.65	61.64	61.45
GLM-4.6V (Hong et al., 2025)	106B	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Qwen3-VL-Instruct (Bai et al., 2025)	235B	91.10	56.09	86.13	80.38	71.74	69.04	75.75
Qwen3.5 (Qwen Team, 2026)	397B	87.96	57.16	89.38	85.50	74.82	69.82	77.44
Kimi-K2.6 (Team, Kimi, 2026)	1T	88.48	53.14	81.88	78.00	69.22	66.13	72.81
Qwen3.5-4B Post-Training Strategies								
Vanilla	4B	82.20	48.76	86.12	79.62	58.81	60.44	69.33
GRPO (Shao et al., 2024)	4B	86.91	58.93	84.50	79.38	72.49	70.78	75.50
Vision-OPD (Yuan et al., 2026)	4B	87.43	59.76	80.62	80.25	70.77	69.38	74.70
OPD-V (Bi et al., 2026)	4B	87.43	58.93	83.75	80.12	70.12	68.80	74.86
VAD (Zhang et al., 2026)	4B	91.10	60.71	85.50	81.25	**72.75**	70.07	76.90
EAD (Ours)	4B	**91.62**	**60.83**	**86.38**	**82.88**	72.39	**71.10**	**77.53**
Qwen3.5-9B Post-Training Strategies								
Vanilla	9B	89.01	53.49	85.38	81.88	71.83	67.47	74.84
GRPO (Shao et al., 2024)	9B	89.53	57.51	84.75	81.62	66.47	65.59	74.25
Vision-OPD (Yuan et al., 2026)	9B	89.53	**66.04**	85.62	82.38	71.11	69.61	77.38
OPD-V (Bi et al., 2026)	9B	90.05	62.25	82.75	82.62	71.65	70.63	76.66
VAD (Zhang et al., 2026)	9B	91.10	61.42	87.62	84.12	72.32	70.07	77.78
EAD (Ours)	9B	**93.19**	62.01	**88.62**	**84.62**	**72.46**	**71.07**	**78.66**

Teacher corrections that agree more strongly with the evidence-induced direction receive larger weights, while opposing directions receive zero weight.

4.2 Learning from Evidence-Aligned Corrections

We use $w_{i,t}$ to weight the original distillation loss, yielding the final EAD objective:

$$\mathcal{L}_{\mathrm{EAD}} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{T_i}\sum_{t=1}^{T_i}\mathrm{sg}(w_{i,t})\,\mathrm{JS}\Big(\mathrm{sg}\Big(p^{C}_{\bar{\theta},i,t}\Big), p^{R}_{\theta,i,t}\Big), \tag{8}$$

where B is the number of on-policy responses, T_i is the number of valid response tokens, and sg denotes stop-gradient. The weighted token losses are averaged within each response and then equally across responses. The teacher and masked-image branch are used only during training, and inference only uses the trained student on the original image.

Table 2: **Effect of model state on the evidence reference.** All variants use the same Original–Mask image pair. *Cross-state* uses different model states for the two predictions, whereas *Teacher-side* and *Student-side (EAD)* use the EMA teacher and current student alone, respectively. Avg. denotes the unweighted mean over the six benchmarks. Best results in each column are shown in bold.

Model State	V*Bench	ZoomBench	HR-Bench		MME-RealWorld		Avg.
			4K	8K	EN	CN	
Cross-state	89.53	59.05	85.12	81.38	71.05	69.75	75.98
Teacher-side	89.01	59.76	85.25	**83.62**	**73.18**	**71.35**	77.03
Student-side (EAD)	**91.62**	**60.83**	**86.38**	82.88	72.39	71.10	**77.53**

Table 3: **Effect of visual contrast on the evidence reference.** The first two variants use the current student and differ only in the visual inputs, while the VAD-style variant uses the EMA teacher on the cropped image and its downsampled counterpart. All variants use the same EAD objective. Avg. denotes the unweighted mean over the benchmarks. Best results in each column are shown in bold.

Visual Contrast	V*Bench	ZoomBench	HR-Bench		MME-RealWorld		Avg.
			4K	8K	EN	CN	
Original–Crop	88.48	60.00	83.00	80.25	70.61	68.62	75.16
Original–Mask (EAD)	**91.62**	**60.83**	86.38	82.88	**72.39**	**71.10**	**77.53**
Crop–Downsampled Crop (VAD)	**91.62**	58.82	**86.50**	**83.00**	69.33	68.73	76.33

5 Experiments

5.1 Experimental Setup

Implementation. We closely follow the Vision-OPD (Yuan et al., 2026) training configuration. We train Qwen3.5-4B and Qwen3.5-9B on the same 6.2K fine-grained examples for one epoch and evaluate the final checkpoint. Distillation uses Jensen–Shannon divergence computed from the top-100 logits with partial softmax, and the privileged teacher is updated by EMA with a rate of 0.05. To reduce premature collapse toward short responses that directly output the option letter, we fix the first generated token to *The*, exclude this token from the training loss, and sample all subsequent tokens normally. Further implementation details are provided in Appendix A.1.

Benchmarks. We evaluate fine-grained visual perception on six benchmarks. V*Bench (Wu & Xie, 2024) evaluates the ability to locate and recognize fine visual details in high-resolution and visually crowded images. ZoomBench (Wei et al., 2026) focuses on fine-grained perception where decisive evidence is localized and difficult to recover from the original image. HR-Bench (Wang et al., 2025) evaluates fine-grained single- and cross-instance perception at both 4K and 8K resolutions. MME-RealWorld (Zhang et al., 2025b) further evaluates high-resolution perception and reasoning in challenging real-world scenarios. We use both its English and Chinese subsets. We additionally evaluate held-out generalization in Appendix A.2. For evaluation, we follow the released Vision-OPD pipeline with a calibrated multiple-choice answer-extraction procedure, applied consistently to all methods. Unresolved responses are evaluated by Qwen3.5-122B-A10B (Qwen Team, 2026) as a fallback judge. More details are provided in Appendix A.3.

Baselines. We compare against two groups of baselines. First, we include representative closed-source and open-weight MLLMs under single-forward-pass inference. Their benchmark scores are taken directly from the evaluation reported by Vision-OPD. Second, we compare alternative post-training strategies using the same Qwen3.5 (Qwen Team, 2026) backbones and training data. At both the 4B and 9B scales, we compare EAD against the Vanilla model, GRPO (Shao et al., 2024), Vision-OPD (Yuan et al., 2026), and the concurrent methods OPD-V (Bi et al., 2026) and VAD (Zhang et al., 2026). We retrain each post-training method on the same backbone and training data and evaluate all resulting checkpoints with the same pipeline. All variants are trained for one epoch, and the final checkpoint is used for evaluation.

Table 4: **Effect of local JS scaling and clipped cosine similarity.** *No JS Scaling* removes the local normalization before computing directional alignment, while *Shifted-Cosine* retains negatively aligned teacher corrections.

Variant	V*Bench	ZoomBench	HRBench-4K	HRBench-8K	MME-EN	MME-CN	Avg.
No JS Scaling	89.01	59.29	85.12	82.38	72.37	**71.29**	76.58
Shifted-Cosine	90.58	**61.18**	83.50	81.25	72.22	70.05	76.46
EAD	**91.62**	60.83	**86.38**	**82.88**	**72.39**	71.10	**77.53**

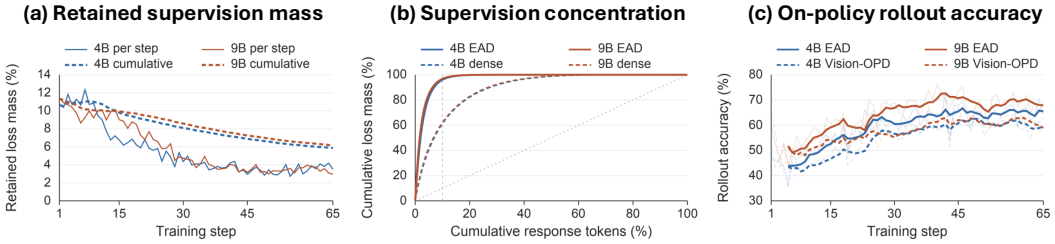

Figure 4: **Sparsity and concentration of EAD supervision.** (a) Fraction of the original distillation-loss mass retained by EAD throughout training. (b) Concentration of the retained loss across response tokens, compared with the original dense distillation loss on the same responses. (c) On-policy rollout accuracy of EAD and Vision-OPD during training. Results are shown for Qwen3.5-4B and Qwen3.5-9B.

5.2 Comparison with State-of-the-Art MLLMs

Table 1 reports the comparison on six fine-grained visual understanding benchmarks. For the Qwen3.5 post-training setting, all methods use the same backbone, training data, and evaluation pipeline. EAD achieves the highest average performance at both the 4B and 9B scales. Compared with Vision-OPD, EAD improves the average by 2.83 and 1.28 points at 4B and 9B, respectively. It also outperforms the strongest concurrent baseline, VAD, by 0.63 and 0.88 points at the two scales. We further compare EAD with representative open-weight and closed-source MLLMs. Despite using substantially smaller backbones, EAD reaches performance comparable to substantially larger models. These consistent gains demonstrate the effectiveness of selectively retaining teacher corrections according to their alignment with visual evidence.

5.3 EAD Retains Sparse and Concentrated Supervision

EAD substantially reduces the amount of teacher supervision used during training, retaining only 5.87% and 6.15% of the original distillation-loss mass for the 4B and 9B models, respectively (Figure 4(a)). Furthermore, this reduction is not uniform: the top 10% highest-loss tokens account for over 95% of the retained loss mass, compared with about 61–62$\%$ under dense distillation (Figure 4(b)). Thus, EAD concentrates supervision on a much smaller subset of corrections rather than simply reducing the overall distillation strength. Despite retaining only a small fraction of the original supervision, EAD maintains higher on-policy rollout accuracy than Vision-OPD throughout training (Figure 4(c)). These results show that EAD produces a sparser and more concentrated distillation objective while maintaining more effective optimization.

5.4 Ablation Studies

We ablate the evidence-reference construction by only varying the prediction difference Δ_t used to form $u_t^{\text{evid}} = \Delta_t / \sqrt{\max(p_{\theta,t}^R, \epsilon)}$, keeping the remaining EAD objective and training procedure fixed. Tables 2 and 3 report the fine-grained results; held-out results are provided in Appendix A.5. All the experiments are conducted on Qwen3.5-4B.

Effect of model state. We first examine which model state should be used to construct the u_t^{evid}. All three variants use the same Original–Mask image pair and differ only in the model states used to compute the two predictions. *Cross-state* uses $\Delta_t = p^R_{\theta,t} - p^{R^-}_{\bar{\theta},t}$, while *Teacher-side* and *Student-side (EAD)* use $\Delta_t = p^R_{\bar{\theta},t} - p^{R^-}_{\bar{\theta},t}$ and $\Delta_t = p^R_{\theta,t} - p^{R^-}_{\theta,t}$, respectively. As shown in Table 2, *Student-side (EAD)* performs best, followed by *Teacher-side*, while *Cross-state* performs worst. Both same-state variants outperform *Cross-state*, showing the benefit of excluding model-state gap when constructing the evidence reference. The further advantage of *Student-side (EAD)* over *Teacher-side* shows that the student's own Original–Mask prediction change provides a more effective evidence reference than the teacher's.

Effect of the visual contrast. We next examine which image pair should be used to construct the evidence reference. For a controlled comparison, both variants use the current student to make predictions and differ only in the visual inputs. *Original–Crop* uses $\Delta_t = p^C_{\theta,t} - p^R_{\theta,t}$, whereas *Original–Mask (EAD)* uses $\Delta_t = p^R_{\theta,t} - p^{R^-}_{\theta,t}$. As shown in Table 3, *Original–Mask (EAD)* consistently outperforms *Original–Crop* across all six benchmarks. This shows that removing the evidence while preserving the surrounding visual context provides a more effective reference than altering the context through cropping.

Comparison with a VAD-style reference. We further consider the image pair used by VAD (Zhang et al., 2026). *Crop–Downsampled Crop (VAD)* uses $\Delta_t = p^C_{\bar{\theta},t} - p^{C_{\text{deg}}}_{\bar{\theta},t}$, where C_{deg} is a downsampled version of the cropped image, while keeping the EAD weighting objective and training procedure unchanged. The *Crop–Downsampled Crop (VAD)* comparison keeps both the model state and cropped region fixed, varying only the visual degradation. However, this distributional change captures the EMA teacher's evidence sensitivity in the cropped view, rather than the current student's evidence sensitivity in the original image. *Original–Mask (EAD)* instead captures the current student's response to visual evidence within the original-image context. As shown in Table 3, *Original–Mask (EAD)* outperforms this VAD-style reference, supporting the use of an evidence reference constructed from the current student in the original-image context.

Effect of local JS scaling. We next examine the effect of the local JS scaling used to compare the teacher correction with the evidence reference. *No JS Scaling* removes the shared $1/\sqrt{\max(p^R_{\theta,t}, \epsilon)}$ scaling and directly computes cosine similarity between the probability differences $p^C_{\bar{\theta},t} - p^R_{\theta,t}$ and $p^R_{\theta,t} - p^{R^-}_{\theta,t}$, while keeping the remaining EAD objective and training procedure unchanged. As shown in Table 4, removing the local JS scaling reduces the average performance from 77.53 to 76.58. This shows that comparing the two prediction changes under the shared local JS scaling provides a more effective alignment signal.

Effect of clipped cosine similarity. We further examine whether negatively aligned teacher corrections should be retained for supervision. *Shifted-Cosine* replaces the EAD weight $\max(\cos(u_t^{\text{corr}}, u_t^{\text{evid}}), 0)$ with $(\cos(u_t^{\text{corr}}, u_t^{\text{evid}})+1)/2$, so that negatively aligned corrections retain non-zero supervision weights. All other settings remain unchanged. As shown in Table 4, *Shifted-Cosine* reduces the average performance from 77.53 to 76.46. This supports excluding teacher corrections that are directionally opposed to the evidence reference.

6 Conclusion

We introduced Evidence-Aligned multimodal on-policy self-Distillation (EAD) for fine-grained visual understanding. EAD retains the privileged teacher as the supervision target while using the current student's Original–Mask prediction change as a separate evidence reference for selectively weighting teacher corrections. By keeping the model state fixed and preserving the surrounding visual context, this reference more cleanly captures the prediction shift induced by task-relevant visual evidence. Across 4B and 9B models, EAD consistently outperforms strong post-training baselines and remains competitive with substantially larger open-weight and closed-source models on fine-grained visual understanding benchmarks.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and to improve the clarity, grammar, and readability of the manuscript. They were not used to formulate research hypotheses, design the methodology or experiments, analyze data, interpret experimental results, or generate scientific claims. All AI-assisted edits were carefully reviewed and revised by the authors. The authors take full responsibility for the final content of the manuscript.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. Opd-v: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, and Yu-Gang Jiang. Thinking with deltas: Incentivizing reinforcement learning via differential visual reasoning policy. *arXiv preprint arXiv:2601.06801*, 2026.

Google. Gemini 3.1 Pro: A smarter model for your most complex tasks, 2026a. URL `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/`.

Google. Gemini 3.5: Frontier intelligence with action, 2026b. URL `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/`.

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.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, volume 2026, pp. 114851–114872, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Nanxing Hu, Xiaoyue Duan, Jinchao Zhang, and Guoliang Kang. Enhancing visual reliance in text generation: A bayesian perspective on mitigating hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4778–4787, 2025.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

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.

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, 2023.

Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jiajia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. *arXiv preprint arXiv:2509.25916*, 2025.

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.

OpenAI. Introducing GPT-5.2, 2025. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Yunhang Qian, Jiaquan Yu, Jiawei Liu, Meng Wang, Hongwei Bran Li, and Xiaobin Hu. Medopd: Improving medical vision-language models via evidence-aware on-policy distillation. *arXiv preprint arXiv:2607.16303*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning. *arXiv preprint arXiv:2606.25319*, 2026.

Almog Tavor, Itay Ebenspanger, Neil Cnaan, and Mor Geva. Rethinking selective knowledge distillation. *arXiv preprint arXiv:2602.01395*, 2026.

Team, Kimi. Kimi k2.6: From code to creation, from one to many, 2026. URL https://www.kimi.com/ai-models/kimi-k2-6/.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024b.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6456–6466, 2021.

Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Ye Tian, Jiahao Meng, Zilong Huang, Guangcan Mai, et al. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 2222–2250, 2026.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

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, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, et al. Vad: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025a.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025b.

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.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

A Appendix

A.1 Additional Implementation Details

Training configuration. We build EAD on the Vision-OPD training pipeline and keep its training configuration unchanged unless otherwise specified. Both Qwen3.5-4B and Qwen3.5-9B are trained for one epoch on the same 6.2K fine-grained examples used by Vision-OPD. We use a training batch size of 96, sample $n = 8$ on-policy responses per example, and optimize the student with a learning rate of 2×10^{-6} and 10 warmup steps. The maximum prompt and response lengths are 8192 and 1024 tokens, respectively. The privileged teacher is updated by exponential moving average (EMA) with an update rate of 0.05.

Table 5: **Preservation of broader visual capabilities.** We evaluate the same 4B and 9B post-training checkpoints from Table 1 on four held-out benchmarks without additional training. Avg. denotes the unweighted mean over the four benchmarks. Best results within each model scale are shown in bold.

Method	MMVP	CV-Bench	MMStar	POPE	Avg.
Qwen3.5-4B Post-Training Strategies					
Vanilla	79.00	86.97	**74.00**	89.33	**82.33**
GRPO (Shao et al., 2024)	78.00	85.78	69.33	87.87	80.25
Vision-OPD (Yuan et al., 2026)	76.67	85.48	70.80	88.99	80.49
OPD-V (Bi et al., 2026)	77.00	84.38	71.40	**89.43**	80.55
VAD (Zhang et al., 2026)	79.67	**87.30**	71.80	88.79	81.89
EAD (Ours)	**80.33**	86.95	73.07	87.70	82.01
Qwen3.5-9B Post-Training Strategies					
Vanilla	82.67	88.41	**77.87**	89.86	84.70
GRPO (Shao et al., 2024)	80.33	87.87	72.80	**89.97**	82.74
Vision-OPD (Yuan et al., 2026)	81.00	86.52	74.33	89.80	82.91
OPD-V (Bi et al., 2026)	80.33	85.36	73.13	89.63	82.11
VAD (Zhang et al., 2026)	**85.00**	**88.82**	76.60	89.54	**84.99**
EAD (Ours)	84.33	88.11	76.40	88.73	84.39

On-policy rollout generation. Training trajectories are generated by the current original-image student. For each sampled response $y = (y_1, \ldots, y_T)$, all subsequent teacher and evidence reference predictions are evaluated under the same student-generated textual prefix $h_t = (x, y_{<t})$. We fix the first generated token to *The* and exclude this token from the training loss; all subsequent response tokens are sampled normally. This construction ensures that the privileged teacher correction and the evidence reference are compared at exactly the same autoregressive states.

Privileged crop and evidence-masked view. We directly use the original-image/crop pairs released by Vision-OPD. Thus, the crop-conditioned EMA teacher receives the same privileged visual input as Vision-OPD. EAD introduces only an additional evidence-masked view of the original image. We construct R^- by masking the GT bounding-box interior with the mean RGB color of its surrounding region, while preserving the red-box annotation and the remaining image context unchanged.

Shared top-K vocabulary support. Following Vision-OPD, all token-level distribution computations use $K = 100$ vocabulary entries. At response position t, we first select the top-100 vocabulary entries according to the current full-image student distribution $p^R_{\theta,t}$. The same vocabulary indices are then used to gather the corresponding logits from the crop-conditioned EMA teacher $p^C_{\bar{\theta},t}$ and the masked-image student $p^{R^-}_{\theta,t}$. Therefore, the three distributions used by EAD are represented in the same vocabulary coordinates, which allows their distributional changes to be compared directly. More explicitly, if

$$\mathcal{S}_t = \mathrm{TopK}\big(p^R_{\theta,t}, K\big), \qquad K = 100,$$

then all quantities in Eqs. 6–7 are computed after restricting $p^R_{\theta,t}$, $p^C_{\bar{\theta},t}$, and $p^{R^-}_{\theta,t}$ to the common support $\mathcal{S}_t$.

A.2 Preservation of broader visual capabilities.

We additionally evaluate the held-out generalization on MMVP (Tong et al., 2024b), CV-Bench (Tong et al., 2024a), MMStar (Chen et al., 2024), and POPE (Li et al., 2023). We use the same 4B and 9B model checkpoints as those evaluated in Table 1. Table 5 shows that EAD largely preserves the base model's average performance across the four held-out benchmarks, reducing the performance degradation relative to dense Vision-OPD by about 83% at both model scales. These results show that EAD improves fine-grained perception while largely preserving broader visual capabilities.

Table 6: **Released Vision-OPD 4B checkpoint under our evaluation pipeline.** All results are obtained using the same revised evaluator applied to the Qwen3.5-based models in this work.

Benchmark	Vision-OPD 4B
V*Bench	90.05
ZoomBench	59.29
HRBench-4K	82.25
HRBench-8K	79.25
MME-RealWorld-EN	70.71
MME-RealWorld-CN	69.43
MMVP	79.00
CV-Bench	86.82
MMStar	70.87
POPE	89.12
Fine Avg.	**75.16**
Holdout Avg.	**81.45**
All-10	**77.68**

A.3 Answer Extraction and Evaluation

Evaluation protocol. All Qwen3.5 post-training variants are evaluated under the same inference and evaluation pipeline. To make multiple-choice grading more robust to free-form model responses, we revise the answer-extraction procedure in the released Vision-OPD evaluator and apply the revised evaluator uniformly to all variants.

Answer extraction. For multiple-choice responses, the released evaluator extracts the predicted option using a sequence of regular-expression patterns. Its final fallback, `([A-Z])`, may extract an uppercase letter from ordinary response text rather than the selected option. For example, for the response `Answer: **D**`, this fallback may match the `A` in `Answer` rather than the stated option `D`.

Therefore, we remove the unrestricted uppercase-letter fallback and only extract option labels from explicit answer expressions or standard option formats. If no label can be resolved, the evaluator matches the response against the option texts and returns a prediction only when a single option is identified. When multiple explicit answer expressions appear, the last one is taken as the final choice. Only unresolved responses are passed to the fallback judge.

Fallback judging. For unresolved responses, we use Qwen3.5-122B-A10B (Qwen Team, 2026) as a text-only fallback judge. The judge receives the question, reference answer, and generated response, but no image, and determines whether the response is equivalent to the reference. We use temperature zero with thinking disabled and constrain the output to `Yes` or `No`.

Evaluation of the released Vision-OPD checkpoint. We evaluate the publicly released Vision-OPD 4B checkpoint with our revised evaluator. This result serves as an additional reference for comparison under a consistent evaluation protocol.

A.4 Details of the Supervision Retention Analysis

Retained supervision mass. Figure 4(a) shows the fraction of distillation loss retained by EAD over the 65 training steps. Let L_s^{dense} and L_s^{EAD} denote the unweighted and EAD-weighted distillation losses at step s, respectively. The per-step and cumulative retained fractions are

$$R_s = \frac{L_s^{\text{EAD}}}{L_s^{\text{dense}}}, \qquad R_{\le S} = \frac{\sum_{s=1}^{S} L_s^{\text{EAD}}}{\sum_{s=1}^{S} L_s^{\text{dense}}}. \tag{9}$$

Solid curves show R_s and dashed curves show $R_{\le s}$. The final cumulative fractions are 5.87% for 4B and 6.15% for 9B. The dense reference is the unweighted loss on the same EAD training trajectories.

Table 7: Summary of supervision retention, concentration, and rollout accuracy. All values are percentages. Top-10% loss mass measures how much of each response's total loss comes from its 10% highest-loss tokens, averaged across responses. Rollout accuracy is computed over all responses from training steps 1–65.

Statistic	4B	9B
Cumulative retained loss mass	5.87	6.15
Top-10% loss mass: EAD	95.98	96.81
Top-10% loss mass: dense	61.10	62.27
Training rollout accuracy: EAD	59.53	64.31
Training rollout accuracy: Vision-OPD	55.06	57.49

Table 8: **Held-out generalization of evidence-reference variants.** The same checkpoints evaluated in Tables 2 and 3 are evaluated on MMVP, CV-Bench, MMStar, and POPE. Holdout Avg. is the unweighted mean over the four held-out benchmarks, Fine Avg. is the mean over the six fine-grained benchmarks, and All-10 averages all ten benchmarks.

Evidence Reference	**MMVP**	**CV-Bench**	**MMStar**	**POPE**	**Holdout Avg.**	**Fine Avg.**	**All-10**
Cross-state masking	77.00	85.49	**73.53**	86.66	80.67	75.98	77.86
Crop-induced change	78.33	86.61	70.87	89.11	81.23	75.16	77.59
Teacher-side masking	78.67	**87.59**	72.67	**89.24**	**82.04**	77.03	79.03
Teacher crop degradation	76.33	84.15	71.87	83.30	78.91	76.33	77.37
Current student (EAD)	**80.33**	86.95	73.07	87.70	82.01	**77.53**	**79.33**

Within-response supervision concentration. For Figure 4(b), we analyze 4B and 9B responses to the same 128 randomly sampled questions from training set. Token-level losses and EAD weights are computed using the step-65 Student and EMA Teacher checkpoints. For each response i, we sort the retained token losses $a_{i,t} = w_{i,t} D_{i,t}^{teacher}$ in descending order. With n_i valid tokens, the fraction of total loss carried by the top k tokens is

$$C_i\left(\frac{k}{n_i}\right) = \frac{\sum_{j=1}^{k} a_{i,(j)}}{\sum_{j=1}^{n_i} a_{i,(j)}}, \qquad k = 0, \ldots, n_i, \tag{10}$$

where $a_{i,(j)}$ is the j-th largest retained token loss. The dense reference follows the same procedure, independently sorting the unweighted losses $D_{i,t}^{teacher}$.

On-policy rollout accuracy. Figure 4(c) compares 4B/9B EAD and Vision-OPD over 65 training steps. Each step uses 96 questions with eight responses per question. All responses are scored using the same answer extractor; unresolved answers count as incorrect. Faint curves show raw per-step accuracy. Bold curves show the mean accuracy over the current step and the preceding four steps.

A.5 Held-Out Generalization of Evidence References

We further evaluate the evidence-reference variants from Section 5.4 on four held-out benchmarks using the same trained Qwen3.5-4B checkpoints. Table 8 reports the held-out results together with the Fine Avg. from the six fine-grained benchmarks and the average over all ten benchmarks.

The held-out results are consistent with the main fine-grained comparison. Both same-state Original–Mask variants outperform the cross-state construction, supporting the benefit of removing teacher–student state differences from the evidence reference. Among the visual comparisons, Original–Mask also generalizes better than the teacher-side crop-degradation reference. Although teacher-side masking achieves a marginally higher Holdout Avg. than EAD, EAD attains the strongest Fine Avg. and All-10 score while maintaining comparable held-out performance.

A.6 Effect of Local JS Scaling

EAD represents both the privileged teacher correction and the learner evidence response in the locally normalized coordinates motivated by the second-order approximation of Jensen–Shannon di-

Table 9: **Effect of local JS scaling.** Vanilla cosine directly compares the unscaled probability differences, whereas EAD applies the shared local normalization $1/\sqrt{\max(p^R_{\theta,t}, \epsilon)}$ before measuring directional alignment. Both variants otherwise use the same evidence reference and training procedure.

Benchmark	Vanilla Cosine	EAD
V*Bench	89.01	**91.62**
ZoomBench	59.29	**60.83**
HRBench-4K	85.12	**86.38**
HRBench-8K	82.38	**82.88**
MME-RealWorld-EN	72.37	**72.39**
MME-RealWorld-CN	**71.29**	71.10
MMVP	79.67	**80.33**
CV-Bench	**87.03**	86.95
MMStar	**73.47**	73.07
POPE	87.57	**87.70**
Fine Avg.	76.58	**77.53**
Holdout Avg.	81.94	**82.01**
All-10	78.72	**79.33**

vergence. We ablate this design by directly computing cosine similarity between the corresponding probability differences without the $1/\sqrt{p^R_{\theta,t}}$ scaling.

Specifically, the unscaled variant uses

$$\tilde{u}^{\text{corr}}_t = p^C_{\theta,t} - p^R_{\theta,t}, \qquad \tilde{u}^{\text{evid}}_t = p^R_{\theta,t} - p^{R^-}_{\theta,t}, \tag{11}$$

and computes

$$\tilde{w}_t = \max\left(\cos\left(\tilde{u}^{\text{corr}}_t, \tilde{u}^{\text{evid}}_t\right), 0\right). \tag{12}$$

All other training settings are kept identical to EAD.

As shown in Table 9, directly comparing the original probability differences already provides a meaningful alignment signal, while local normalization yields stronger performance. These results support the use of the shared local JS geometry.

A.7 Effect of Excluding Negatively Aligned Corrections

EAD uses positive cosine weighting, so teacher corrections with negative alignment to the evidence reference are excluded from supervision. We test this choice with a shifted cosine that retains negatively aligned corrections. Specifically, the shifted-cosine variant uses

$$\tilde{w}_t = \frac{\cos\left(u^{\text{corr}}_t, u^{\text{evid}}_t\right) + 1}{2}, \tag{13}$$

while keeping all other training settings unchanged. As shown in Table 10, Shifted-Cosine yields lower overall performance than EAD. These results support excluding opposing teacher corrections from supervision.

A.8 Additional Token-Level Case Studies

Figure 5 provides additional examples across different fine-grained visual tasks, including attribute recognition, text recognition, shape recognition, and object identification. For each example, we visualize the normalized token-wise JS divergence of D^{teacher}_t, D^{crop}_t, and D^{mask}_t on the same generated response.

Across these cases, D^{mask}_t places greater emphasis on tokens associated with the task-relevant visual evidence, whereas D^{crop}_t and D^{teacher}_t exhibit large divergence on tokens less directly related to that evidence. These examples are consistent with the observation in Figure 2 and further motivate using the Original–Mask comparison as the evidence reference.

Table 10: **Effect of excluding negatively aligned corrections.** Shifted-Cosine retains negatively aligned corrections by mapping cosine similarity from $[-1, 1]$ to $[0, 1]$. All other settings are unchanged.

Benchmark	Shifted-Cosine	EAD
V*Bench	90.58	**91.62**
ZoomBench	**61.18**	60.83
HRBench-4K	83.50	**86.38**
HRBench-8K	81.25	**82.88**
MME-RealWorld-EN	72.22	**72.39**
MME-RealWorld-CN	70.05	**71.10**
MMVP	77.33	**80.33**
CV-Bench	**87.21**	86.95
MMStar	**73.93**	73.07
POPE	**88.88**	87.70
Fine Avg.	76.46	**77.53**
Holdout Avg.	81.84	**82.01**
All-10	78.61	**79.33**

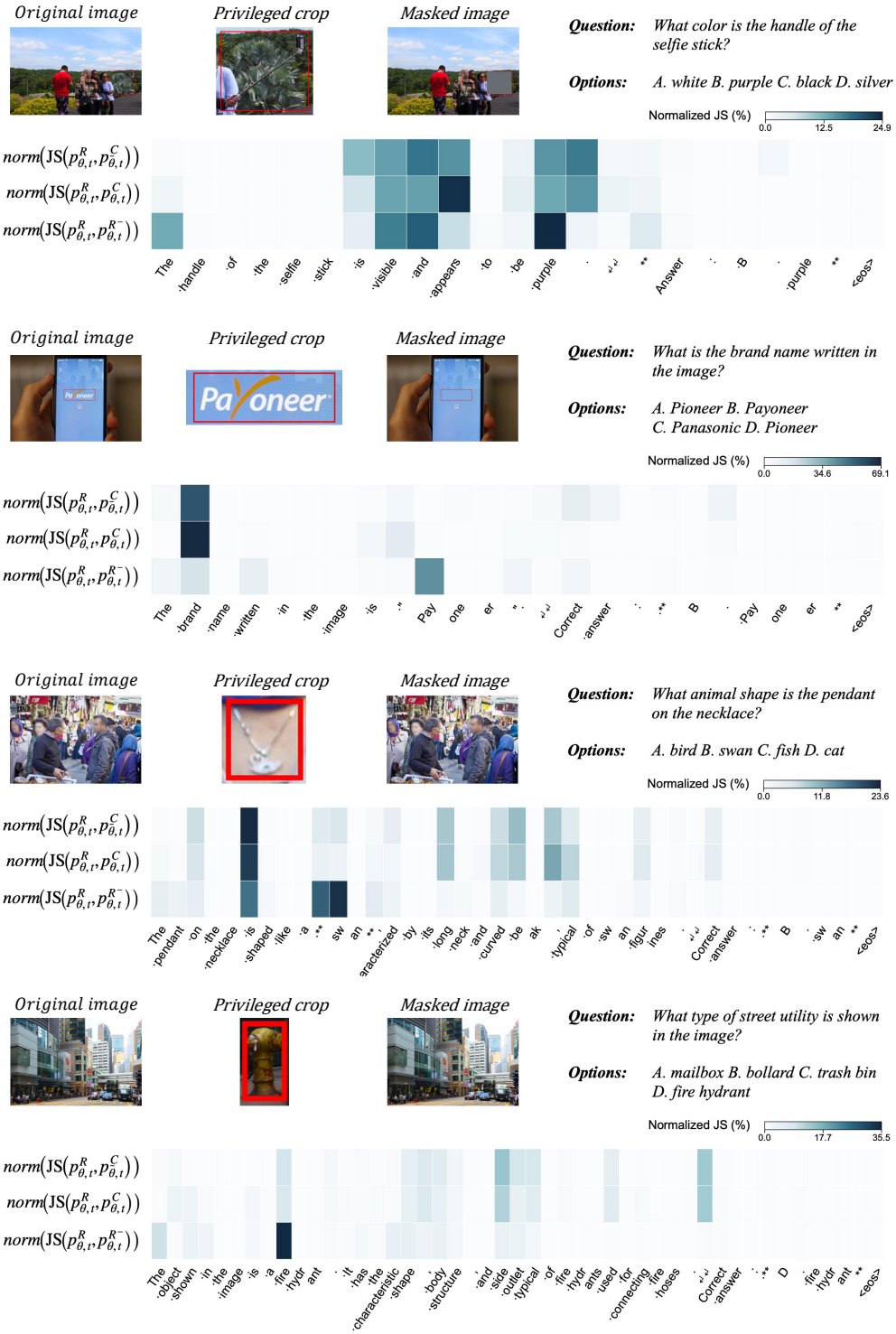

Figure 5: **Additional token-level case studies.** Each example shows the original, cropped, and masked images together with the normalized token-wise JS divergence of D_t^{teacher}, D_t^{crop}, and D_t^{mask}. Across different fine-grained visual tasks, D_t^{mask} tends to emphasize tokens more closely associated with the task-relevant visual evidence, while the other comparisons can assign large divergence to less directly related tokens.