PARTIAL-OBSERVATION-AWARE SELF-DISTILLATION FOR FINE-GRAINED VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models often recognize fine-grained details more accurately in a relevant image crop than in the full image. Local-view on-policy self-distillation (OPSD) uses a crop-conditioned teacher to transfer this advantage to a full-image student. However, we uncover an unexpected localization–observation tradeoff: while local-view OPSD strengthens fine-grained observation, it can degrade the ability to localize relevant evidence under full-image inference. Through theoretical analysis and token-level empirical studies, we trace this failure to the partial-observation nature of local-view supervision: it assigns systematically negative optimization signals to tokens that depend on the omitted image context, disproportionately suppressing localization-related tokens. Based on this diagnosis, we propose Partial-Observation-Aware On-Policy Self-Distillation (PA-OPSD), which selectively attenuates unreliable strongly negative optimization signals and introduces a complementary box-guided teacher that restores localization guidance while preserving full-image context. Across fine-grained visual reasoning and localization-oriented benchmarks with 4B and 9B models, PA-OPSD consistently outperforms local-view OPSD and other on-policy baselines, recovering localization while preserving observation benefits. It requires no additional annotations beyond local-view OPSD and no inference-time visual operations.

1 INTRODUCTION

Vision-language models (VLMs) have achieved remarkable progress in general visual understanding and multimodal reasoning (Qwen Team, 2026; Google, 2025; OpenAI, 2026). However, they still struggle on fine-grained visual reasoning tasks, where the decisive evidence is often small, visually cluttered, or embedded in high-resolution images (Wu & Xie, 2024; Wang et al., 2025c; Zhang et al., 2024). Recent studies suggest that success on such tasks requires two tightly coupled capabilities: *localizing* the relevant evidence and *observing* it accurately once localized (Zhang et al., 2025a; Li et al., 2026a).

Recent "Thinking-with-Images" methods (Zhang et al., 2025b; Zheng et al., 2025) improve fine-grained visual reasoning by cropping or zooming into relevant regions during inference, making local evidence easier to observe. However, these operations introduce additional inference overhead. Motivated by the observation that the same model often performs better on a relevant local crop than on the full image, recent local-view on-policy self-distillation (OPSD) methods, such as Vision-OPD (Yuan et al., 2026) and related approaches (Cai et al., 2026; Aniri et al., 2026), seek to internalize this observation benefit. They use a privileged teacher conditioned on the relevant local crop to provide token-level supervision to a student conditioned on the full image, enabling inference from the full image alone without explicit cropping operations.

Despite this intuitive motivation, our systematic empirical analysis reveals an unexpected limitation of local-view OPSD. As shown in Figure 1, Vision-OPD outperforms GRPO (Shao et al., 2024), a widely adopted on-policy learning baseline, on 37 of the 39 subtasks in MME-RealWorld (Zhang et al., 2024). On the remaining two subtasks, however, Vision-OPD substantially underperforms GRPO. A closer examination reveals that these two subtasks primarily evaluate *object localization*, whereas the remaining subtasks mainly require fine-grained observation and reasoning over already localized evidence. This contrast reveals an unexpected *localization–observation tradeoff*: local-view supervision substantially improves observation-related reasoning while weakening the

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.

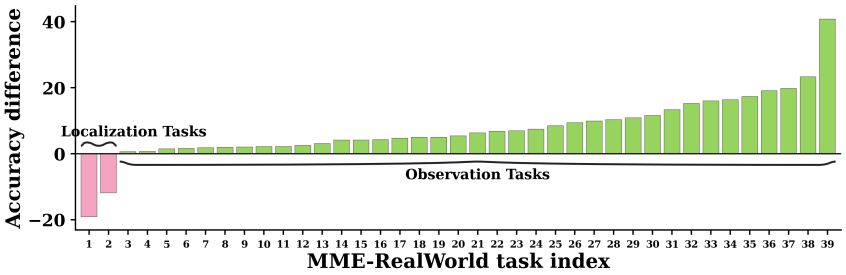


Figure 1: Vision-OPD vs. GRPO on MME-RealWorld using Qwen3.5-4B. Vision-OPD outperforms GRPO on 37 of 39 subtasks, while the two exceptions are both localization-oriented tasks.

model's ability to localize relevant evidence under full-image inference. This tradeoff suggests that supervision from a privileged local view is not uniformly beneficial, raising a natural question: *why does local-view supervision benefit observation while degrading localization?*

These observations motivate us to revisit local-view OPSD from the perspective of optimization dynamics. We first show that partial observation can induce negative optimization signals for tokens whose prediction benefits from context omitted from the teacher's local view. We then examine which reasoning behaviors exhibit this dependence. Two complementary token-level analyses reveal a consistent pattern: on verified correct trajectories, correct localization behavior receives negative supervision while observation behavior receives positive supervision; under the on-policy token distribution, localization tokens become increasingly concentrated in the strongly negative tail. Together, these results identify a localization-specific optimization bias induced by partial observation.

Motivated by this diagnosis, we propose **P**artial-Observation-**A**ware **O**n-**P**olicy **S**elf-**D**istillation (**PA-OPSD**). PA-OPSD preserves the useful observation supervision provided by the local-view teacher while correcting its localization-specific optimization bias. It combines partial-observation-aware calibration, which selectively attenuates strongly negative optimization signals, with a box-guided teacher that provides complementary localization guidance under the full-image observation.

Experiments across fine-grained visual reasoning and localization-oriented benchmarks show that PA-OPSD consistently improves over the local-view OPSD baseline and other on-policy learning baselines. Further ablation and optimization analyses demonstrate that our design effectively corrects the identified localization-specific optimization bias.

Our contributions are summarized as follows:

- We uncover an *unexpected localization–observation tradeoff* in local-view on-policy self-distillation, and trace it to a localization-specific optimization bias through theoretical and token-level empirical analyses.
- We propose **PA-OPSD**, which selectively attenuates strongly negative optimization signals induced by local-view supervision and provides complementary localization guidance through a box-guided teacher.
- PA-OPSD consistently improves over local-view OPSD and other on-policy baselines across fine-grained reasoning and localization-oriented benchmarks, demonstrating that it effectively recovers localization capability while preserving the observation benefits of local-view distillation.

2 Related Work

2.1 On-Policy Distillation and Self-Distillation

On-policy distillation (OPD) trains a student on its self-generated trajectories while a teacher provides dense token-level supervision on the states actually visited by the student (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025; Gu et al., 2024). On-policy self-distillation (OPSD) further constructs the teacher from the same model using privileged training context, enabling self-

improvement without a stronger external teacher (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026; Li et al., 2026d). Recent analyses, however, show that privileged context can also introduce non-transferable or misleading token-level supervision, motivating methods that selectively refine the resulting distillation signal (Shen et al., 2026a;b). Multimodal extensions instantiate privilege through localized visual evidence, higher-resolution inputs, visual reasoning traces, or other enriched visual contexts (Yuan et al., 2026; Cai et al., 2026; Zhu et al., 2026; Li et al., 2026b; Aniri et al., 2026; Sun et al., 2026). Our work studies a complementary failure mode: when the privileged teacher observes only a partial view of the image, its supervision can become systematically biased for behaviors that depend on the omitted visual context.

2.2 Fine-Grained Visual Understanding for VLMs

Fine-grained visual reasoning remains challenging for VLMs when task-relevant evidence is small, cluttered, or embedded in high-resolution images (Wu & Xie, 2024; Wang et al., 2025c; Zhang et al., 2024). Recent "Thinking-with-Images" methods address this challenge by explicitly manipulating visual inputs during inference, e.g., through searching, cropping, zooming, or visual editing (Hu et al., 2024; Fu et al., 2025; Zheng et al., 2025; Zhang et al., 2025b; Wang et al., 2025a). Other approaches strengthen fine-grained visual capabilities during training by leveraging localized visual evidence (Wang et al., 2025b; Li et al., 2026a; Wu et al., 2026; Wei et al., 2026). Complementary work on visual grounding shows that localizing referred targets remains challenging for VLMs, particularly under fine-grained distractors and reasoning-intensive referring expressions (Li et al., 2026c; Dong et al., 2026). Local-view OPSD offers another route by using privileged local-view supervision during training to improve a full-image policy, avoiding explicit visual operations at inference (Yuan et al., 2026; Aniri et al., 2026; Cai et al., 2026; Sun et al., 2026).

3 Preliminaries

On-Policy Self-Distillation. On-policy self-distillation (OPSD) (He et al., 2026; Zhao et al., 2026; Cui et al., 2026) derives a student and a privileged teacher from the same model under different conditioning contexts. Given an input x, the student samples an autoregressive rollout $Y = (Y_1, \ldots, Y_T) \sim \pi_\theta(\cdot \mid x)$, with prefix $H_t = Y_{<t}$. Their token distributions are

$$P_S^t(v) = \pi_\theta(v \mid x, H_t), \qquad P_T^t(v) = \pi_{\bar{\theta}}(v \mid x, Z, H_t), \tag{1}$$

where Z denotes privileged training context unavailable at inference time, such as a ground-truth answer or reasoning trajectory, and $\bar{\theta}$ is a frozen or slowly updated teacher-side copy of the student parameters. OPSD minimizes a token-level divergence along the student rollout:

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{(x,Z)\sim\mathcal{D}}\mathbb{E}_{Y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{T}\sum_{t=1}^{T} D\big(\mathrm{sg}(P_T^t)\big\|P_S^t\big)\right], \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. This preserves the student's on-policy state distribution while providing dense token-level supervision from the privileged teacher.

Local-View OPSD. For fine-grained visual reasoning, local-view OPSD (Yuan et al., 2026; Cai et al., 2026; Aniri et al., 2026) uses a full-image student and a privileged teacher receiving a question-relevant local view. We write the full image as $I = (C, O)$, where C denotes the local crop and O the remaining full-image context. Given the same prefix H_t, their token distributions are

$$P_S^t(v) = \pi_\theta(v \mid I, q, H_t), \qquad P_{\mathrm{LV}}^t(v) = \pi_{\bar{\theta}}(v \mid C, q, H_t). \tag{3}$$

Local-view OPSD instantiates Eq. (2) with $P_T^t = P_{\mathrm{LV}}^t$ and student rollouts generated from the full image. This formulation serves as the baseline throughout the paper.

4 Diagnosing Partial-Observation Mismatch

4.1 An Unexpected Localization–Observation Tradeoff

We begin by examining the behavior of local-view OPSD at the subtask level. As illustrated in Figure 1, Vision-OPD outperforms GRPO under identical training data and settings on 37 of the 39 MME-RealWorld subtasks, yet substantially underperforms it on the remaining two.

Notably, both exceptional subtasks specifically evaluate *object localization*, whereas the other subtasks predominantly require fine-grained observation or reasoning over localized visual evidence. The degradation is not limited to comparison with GRPO: Vision-OPD also falls below the initialization model in overall localization performance (Table 4). The same pattern is further observed on the localization-oriented benchmarks in Table 2, where Vision-OPD underperforms the base model on both GroundingME (Li et al., 2026c) and Ref-Adv (Dong et al., 2026).

Together, these results reveal an unexpected *localization–observation tradeoff*: local-view distillation provides benefits for fine-grained observation and reasoning, but can degrade the ability to localize relevant evidence under full-image inference. This raises the question underlying our analysis: *what property of local-view distillation gives rise to this asymmetric optimization behavior?*

4.2 Optimization Dynamics Under Partial Observation

To understand the origin of the capability tradeoff identified above, we next analyze the optimization dynamics induced by local-view distillation.

Using the notation introduced in Section 3, the student and local-view teacher token distributions are $P_S^t(v) = \pi_\theta(v \mid C, O, q, H_t)$ and $P_{\mathrm{LV}}^t(v) = \pi_{\bar{\theta}}(v \mid C, q, H_t)$, where v denotes an arbitrary candidate token in the vocabulary. Following previous analyses of on-policy self-distillation (Shen et al., 2026a; Pan et al., 2026), we characterize their discrepancy using the local-view teacher–student log-ratio $u_{\mathrm{LV}}^t(v) = \log P_{\mathrm{LV}}^t(v) - \log P_S^t(v)$.

We first show that the sign of the optimization signal is determined by the teacher–student log-ratio. Under the regularity conditions specified in Appendix B, the descent direction of an f-divergence admits the following policy-gradient form:

$$-\nabla_\theta D_f\big(P_{\mathrm{LV}}^t \big\| P_S^t\big) = \mathbb{E}_{v\sim P_S^t}\left[A_f\big(u_{\mathrm{LV}}^t(v)\big)\,\nabla_\theta \log P_S^t(v)\right], \tag{4}$$

where $A_f(\cdot)$ is a strictly increasing function satisfying $A_f(0) = 0$. Therefore, the policy-gradient coefficient has the same sign as $u_{\mathrm{LV}}^t(v)$: tokens with positive log-ratios are encouraged during optimization, whereas tokens with negative log-ratios are suppressed. Forward KL, reverse KL and the Jensen–Shannon divergence adopted by Vision-OPD all satisfy this property. The proof of Eq. (4), together with the derivations for the three divergences, is provided in Appendix B.

Since the teacher parameters $\bar{\theta}$ are introduced as a stabilized counterpart of the student, and the local-view condition is not directly optimized during student training, their predictions are expected to remain close when conditioned on the same local-view observation. We empirically validate this approximation in Appendix C, where replacing the frozen local-view teacher with the trained student produces only a small change in the resulting teacher–student log-ratios. Under this approximation,

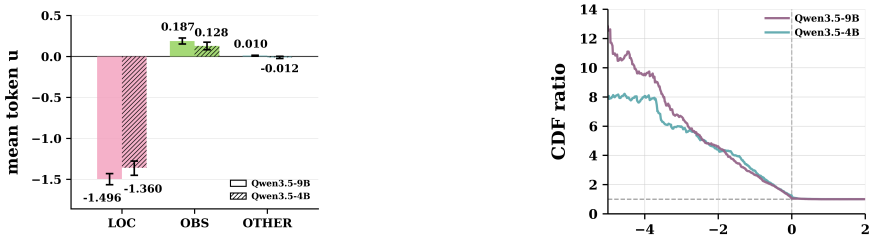
$$u_{\mathrm{LV}}^t(v) = \log \frac{\pi_{\bar{\theta}}(v \mid C, q, H_t)}{\pi_\theta(v \mid C, O, q, H_t)} \approx \log \frac{\pi_\theta(v \mid C, q, H_t)}{\pi_\theta(v \mid C, O, q, H_t)} = -i_\theta(v; O \mid C, q, H_t), \tag{5}$$

where $i_\theta(Y_t = v; O \mid C, q, H_t)$ denotes the conditional pointwise mutual information between the next token and the omitted image context. Eq. (5) therefore indicates that tokens whose prediction benefits from the omitted image context tend to receive more negative optimization signals under local-view distillation. Which reasoning behaviors exhibit this dependence, however, remains an empirical question. We examine this question in the next subsection.

4.3 Which Reasoning Behaviors Depend on the Omitted Context?

The analysis above suggests that reasoning behaviors that rely more strongly on positive evidence from the omitted image context should exhibit distinct optimization patterns under local-view distillation. To identify these behaviors, we conduct two complementary analyses on samples from the Vision-OPD-6K training set (Yuan et al., 2026) using Qwen3.5-4B and Qwen3.5-9B, characterizing the optimization signal from different perspectives:

- **Reference-trajectory analysis.** We construct 200 independently verified reasoning trajectories from sampled examples. Since the evaluated token sequence is fixed, this analysis decouples token selection from the student policy and evaluates the teacher–student log-ratio on correct reasoning paths. It therefore reveals whether each reasoning behavior is encouraged ($u > 0$) or suppressed ($u < 0$) under local-view distillation.



(a) Mean teacher–student log-ratio on verified reasoning trajectories.

(b) Lower-tail ratio under the on-policy token distribution.

Figure 2: Optimization patterns under local-view distillation.

- **On-policy analysis.** We independently sample 200 question–answer pairs and generate reasoning trajectories with each model. This analysis preserves the on-policy token distribution encountered during training and characterizes where different reasoning behaviors are distributed along the teacher–student log-ratio spectrum.

For both analyses, reasoning spans are annotated as LOC (localization), OBS (observation), or OTHER, and mapped to individual tokens. Details of trajectory construction and annotation are provided in Appendix D.

Both diagnostics exhibit consistent patterns for Qwen3.5-4B and Qwen3.5-9B. We first examine the reference-trajectory analysis. Figure 2a reports the mean log-ratio on the verified reasoning trajectories. Because these trajectories represent correct reasoning paths, the sign of the log-ratio directly reflects whether correct behavior is encouraged or suppressed by the local-view teacher. Localization tokens receive substantially negative mean log-ratios, while observation tokens receive positive supervision and OTHER tokens remain approximately neutral. These results show that **correct localization behavior is systematically suppressed** under local-view distillation, even when the underlying reasoning trajectory is correct.

We next study how this suppression is distributed under the student's on-policy token distribution. For a threshold x, we define $R(x) = \Pr(u^t_{\mathrm{LV}}(Y_t) < x \mid \mathrm{LOC}) / \Pr(u^t_{\mathrm{LV}}(Y_t) < x \mid \mathrm{OBS})$, where $R(x) > 1$ indicates that localization tokens are more likely than observation tokens to receive optimization signals below threshold x. Figure 2b shows that localization tokens become increasingly over-represented as the threshold moves deeper into the negative tail. This suggests that strongly negative optimization signals are increasingly concentrated on localization-related behavior, while log-ratios near zero contain a larger mixture of ordinary observation-related error correction. A token-level qualitative example illustrating this pattern is provided in Appendix F.2.

Taken together, the reference analysis shows that correct localization behavior receives systematic negative supervision, while the on-policy analysis shows that this bias becomes increasingly concentrated toward the strongly negative tail. Consequently, **increasingly negative optimization signals are increasingly likely to reflect omitted-context-dependent localization rather than ordinary error correction.** This localization-specific optimization bias provides a mechanistic explanation for the localization degradation observed in Section 4.1.

5 Partial-Observation-Aware OPSD

Figure 3 summarizes the motivation and overall framework of PA-OPSD. Under partial observation, localization-related behavior becomes increasingly concentrated in the strongly negative tail of the local-view teacher–student log-ratio, while observation-related reasoning continues to receive reliable supervision. PA-OPSD addresses this bias through two complementary corrections: calibrating strongly negative local-view supervision and introducing a box-guided teacher that restores localization guidance under full-image observation. These signals are combined into a fused distillation target for the full-image student.

PA-OPSD builds on the same local-view OPSD setup without introducing additional training examples or annotations. For each training example $(I, q, B^\star)$, where $B^\star$ denotes the annotated evidence

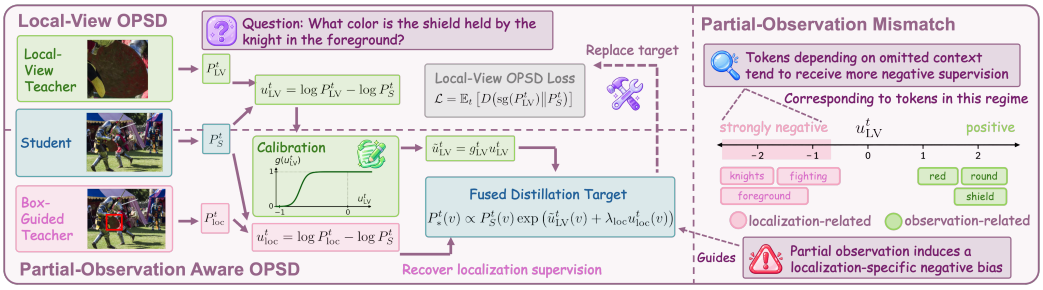


Figure 3: Overview of PA-OPSD. Under local-view distillation, tokens depending on omitted image context tend to receive strongly negative supervision, disproportionately affecting localization-related behavior. PA-OPSD calibrates the local-view signal to attenuate such strongly negative supervision and introduces a box-guided teacher that provides complementary localization guidance while preserving full-image context. The two signals are combined into a fused distillation target for the full-image student.

box, the student generates on-policy trajectories from the full image I, while the local crop C is constructed from $B^\star$ following the baseline setup. PA-OPSD modifies only the distillation target through two complementary components: partial-observation-aware calibration and box-guided localization guidance constructed from the same evidence-box annotation.

Partial-observation-aware calibration. To preserve useful error-correction supervision, we retain positive and mildly negative log-ratios while selectively attenuating the strongly negative tail. Specifically, given the local-view teacher–student log-ratio $u^t_{\mathrm{LV}}(v)$, we define a calibration weight

$$g^t_{\mathrm{LV}}(v) = \begin{cases} 1, & u^t_{\mathrm{LV}}(v) \geq 0, \\ \sigma\left(\gamma\left(u^t_{\mathrm{LV}}(v) + \tau\right)\right), & u^t_{\mathrm{LV}}(v) < 0, \end{cases} \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The parameter τ controls the location of the transition in the negative log-ratio regime, with $g^t_{\mathrm{LV}}(v) = 1/2$ at $u^t_{\mathrm{LV}}(v) = -\tau$, while γ controls its sharpness. Accordingly, mildly negative log-ratios above $-\tau$ receive relatively large weights, whereas the weight decreases smoothly as the log-ratio moves deeper into the negative tail.

The calibrated local-view log-ratio is then defined as $\widetilde{u}^t_{\mathrm{LV}}(v) = g^t_{\mathrm{LV}}(v)\, u^t_{\mathrm{LV}}(v)$. Positive log-ratios are preserved exactly, while negative signals are attenuated according to their magnitude. This calibration avoids discarding negative supervision: ordinary error-correction signals near zero remain largely preserved, whereas strongly negative signals, which are increasingly likely to arise from omitted-context-dependent localization, receive progressively smaller influence.

Recovering localization guidance. Calibration prevents the local-view teacher from excessively suppressing localization-related behavior, but attenuation alone cannot provide affirmative guidance about where the relevant evidence lies. We therefore seek a complementary privileged teacher specifically for localization.

Such a teacher should satisfy three properties. First, it should preserve the full-image context available to the student, avoiding the partial-observation mismatch responsible for the bias identified above. Second, it should make the location of the relevant evidence explicit, providing privileged *where-to-look* information rather than only a local observation after localization has already been performed. Third, it should require no additional supervision beyond the evidence-box annotation already used by local-view OPSD.

These requirements naturally motivate a box-guided localization teacher. Using the same annotated evidence box $B^\star$, we construct $I^{\square} = \mathrm{draw}(I, B^\star)$, which preserves the complete image while explicitly marking the question-relevant region. The corresponding box-guided teacher distribution is $P^t_{\mathrm{loc}}(v) = \pi_{\bar{\theta}}(v \mid I^{\square}, q, H_t)$, and its teacher–student log-ratio is $u^t_{\mathrm{loc}}(v) = \log P^t_{\mathrm{loc}}(v) - \log P^t_S(v)$. Unlike the local-view teacher, the box-guided teacher provides explicit localization guidance while

preserving the full-image observation available to the student. Together, the two teachers provide complementary supervision: the local-view teacher supports fine-grained observation, while the box-guided teacher recovers localization guidance.

Unified distillation objective. We combine the calibrated local-view supervision and the box-guided localization guidance by expressing both as log-ratio corrections relative to the student distribution. The resulting fused target distribution is

$$P_\star^t(v) \propto P_S^t(v) \exp\big(\widetilde{u}_{\mathrm{LV}}^t(v) + \lambda_{\mathrm{loc}} u_{\mathrm{loc}}^t(v)\big)\,, \tag{7}$$

where λ_{loc} controls the contribution of the box-guided teacher. The first correction preserves reliable local-view supervision while reducing the influence of its strongly negative tail, and the second provides complementary localization guidance. The fused target is used in the same on-policy distillation objective as Eq. (2), replacing P_T^t with $P_\star^t$. Thus, PA-OPSD modifies only the distillation target, while the overall OPSD training pipeline and inference procedure remain unchanged.

6 Experiments

6.1 Experimental Setup

Training setup. We conduct experiments using Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026) as backbone models. All methods are trained on the Vision-OPD-6K dataset (Yuan et al., 2026), which contains 6.2K image–question pairs annotated with evidence regions. Following Vision-OPD, we construct the local-view teacher input by enlarging the annotated evidence region by $2\times$. For the box-guided teacher, the same evidence box is rendered on the original image.

We use a frozen copy of the initial checkpoint as the teacher throughout training. All models are trained for one epoch, corresponding to 65 optimization steps, with a learning rate of 2×10^{-6} and linear warmup over the first 10 steps. The maximum generation length is 1024. Following Vision-OPD, we adopt the Jensen–Shannon divergence with $\beta = 0.5$ as the distillation objective, implemented using top-K distillation ($K = 100$) together with a tail-probability approximation for the remaining vocabulary. Unless otherwise specified, we fix $\lambda_{\mathrm{loc}} = 0.2$, $\tau = 0.7$, and $\gamma = 20$ throughout all experiments. Training is conducted on $4\times$ RTX PRO 6000 GPUs.

Benchmarks. We evaluate our method on two groups of benchmarks. The first group measures general fine-grained visual reasoning, including V* Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench 4K, HR-Bench 8K (Wang et al., 2025c), and MME-RealWorld-Lite (Zhang et al., 2024). The second group consists of localization-oriented benchmarks, including GroundingME (Li et al., 2026c) and Ref-Adv (Dong et al., 2026). We follow the official input format for each benchmark. Responses are evaluated using rule-based parsing whenever possible; otherwise, they are assessed by Qwen3.5-9B as an LLM judge with its default decoding configuration.

Baselines. We compare PA-OPSD with representative on-policy learning methods, including GRPO, DAPO, OPSD, and Vision-OPD. GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025) optimize the model using reinforcement learning with verifiable rewards. OPSD (Zhao et al., 2026) performs on-policy self-distillation by conditioning the teacher on the ground-truth answer as privileged information. Vision-OPD (Yuan et al., 2026), as a representative local-view OPSD method, instead provides the teacher with a question-relevant local crop as privileged visual information while the student operates on the original full image. Unless otherwise specified, all methods are trained on the same dataset for one epoch using identical model initialization and evaluation settings. All reported baseline results are reproduced using our unified training and evaluation pipeline.

6.2 Main Results

Table 1 reports results on five fine-grained visual reasoning benchmarks. PA-OPSD achieves the best average performance at both model scales, improving over Vision-OPD from 75.16 to 77.45 on Qwen3.5-4B and from 76.79 to 78.43 on Qwen3.5-9B. The gains are also broadly consistent across benchmarks: PA-OPSD improves over Vision-OPD on all five benchmarks with the 4B model and

Table 1: Main results on fine-grained visual reasoning benchmarks. "w/o C." removes partial-observation-aware calibration, and "w/o B." removes the box-guided teacher. The best result within each model scale is shown in bold.

Method	V^* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW-Lite	Average
Qwen3.5-4B						
Base model	84.29	50.65	83.25	80.13	47.06	69.08
GRPO	83.77	55.38	81.88	78.25	61.49	72.15
DAPO	84.82	55.74	84.00	78.00	62.27	72.97
OPSD	85.34	53.85	82.25	79.38	59.93	72.15
Vision-OPD	90.05	59.76	83.50	80.38	62.12	75.16
PA-OPSD w/o C.	92.15	**62.72**	83.13	80.50	63.16	76.33
PA-OPSD w/o B.	92.15	61.42	83.75	80.25	64.62	76.44
PA-OPSD	**93.19**	61.54	**85.75**	**81.13**	**65.66**	**77.45**
Qwen3.5-9B						
Base model	84.82	52.31	84.75	81.50	59.09	72.49
GRPO	88.48	57.51	84.50	81.50	60.55	74.51
DAPO	88.48	55.62	86.25	84.13	61.07	75.11
OPSD	91.10	57.51	86.88	84.25	58.88	75.72
Vision-OPD	86.91	**63.55**	86.75	84.88	61.86	76.79
PA-OPSD w/o C.	90.05	62.49	87.13	84.50	63.47	77.53
PA-OPSD w/o B.	90.58	61.42	87.38	84.13	63.11	77.32
PA-OPSD	**91.62**	62.25	**88.25**	**85.50**	**64.52**	**78.43**

Table 2: Evaluation on localization-oriented benchmarks.

	Qwen3.5-4B		Qwen3.5-9B	
Method	**GroundingME**	**Ref-Adv**	**GroundingME**	**Ref-Adv**
Base model	49.88	58.23	49.75	62.87
Vision-OPD	44.15	55.34	34.08	60.51
PA-OPSD w/o C.	51.24	58.93	51.87	64.71
PA-OPSD w/o B.	51.62	59.37	51.24	64.48
PA-OPSD	**52.11**	**61.82**	**52.36**	**65.32**

on four of five with the 9B model. These results show that the overall improvement is not driven by a single benchmark and demonstrate the effectiveness of PA-OPSD for fine-grained visual reasoning.

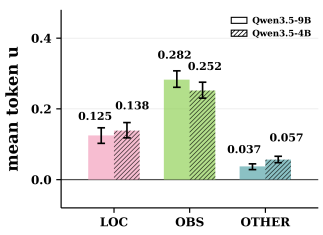
Table 2 reports results on localization-oriented benchmarks. At both model scales, Vision-OPD underperforms the base model on GroundingME and Ref-Adv, consistent with the localization degradation identified in Section 4.1. In contrast, PA-OPSD not only reverses this degradation but also surpasses the base model on both benchmarks. Relative to Vision-OPD, PA-OPSD improves GroundingME from 44.15 to 52.11 on Qwen3.5-4B and from 34.08 to 52.36 on Qwen3.5-9B, while improving Ref-Adv from 55.34 to 61.82 and from 60.51 to 65.32, respectively. Together with the gains on fine-grained reasoning benchmarks, these results show that PA-OPSD effectively recovers localization capability while preserving the observation benefits of local-view distillation.

6.3 Ablation Studies

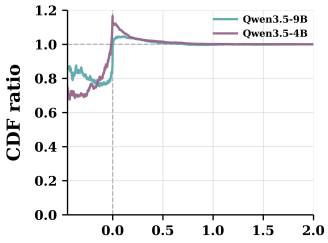
We first examine the contributions of the two components using the variants reported in Tables 1 and 2. Starting from Vision-OPD, introducing either partial-observation-aware calibration (w/o B.) or the box-guided localization teacher (w/o C.) consistently improves performance, showing that each component is effective on its own. Combining the two yields the best overall results, indicating that they provide complementary benefits for fine-grained reasoning and localization. We further compare alternative implementations of the two PA-OPSD design principles on Qwen3.5-4B in Table 3. For calibration, negative scaling, hard thresholding, and sigmoid gating all improve over PA-OPSD without calibration, supporting the benefit of attenuating unreliable negative supervision. Sigmoid gating performs best, consistent with our motivation to smoothly suppress strongly neg-

Table 3: Ablations of the calibration strategy and localization teacher design on Qwen3.5-4B.

Calibration Strategy			Localization Teacher Design		
Method	**MME-RW-Lite**	**Ref-Adv**	**Method**	**MME-RW-Lite**	**Ref-Adv**
PA-OPSD w/o C.	63.16	58.93	PA-OPSD w/o B.	64.62	59.37
Negative Scaling	64.12	59.63	Coordinate Prompt	64.30	59.81
Hard Threshold	65.14	60.42	Context Blur	64.56	59.11
Sigmoid Gating	**65.66**	**61.82**	**Bounding Box Overlay**	**65.66**	**61.82**



(a) Mean teacher–student log-ratio on verified reasoning trajectories.



(b) Lower-tail ratio under the on-policy token distribution.

Figure 4: Optimization patterns induced by the PA-OPSD target.

ative signals while preserving mildly negative supervision. For localization guidance, coordinate prompting and context blurring are less effective than bounding-box overlay, which improves both benchmarks and supports the use of a direct visual cue for localization guidance. Moreover, we report the hyperparameter sensitivity results for τ, λ_{loc}, and γ in Appendix E; PA-OPSD remains stable over a broad range of settings, with the default values providing a favorable balance between fine-grained visual reasoning and localization performance.

6.4 Does PA-OPSD Correct the Optimization Bias?

To test whether PA-OPSD corrects the optimization bias identified in Section 4.3, we recompute the same two diagnostics for Qwen3.5-4B and Qwen3.5-9B using the fused-target log-ratio, $u^t_\star(v) = \log P^t_\star(v) - \log P^t_S(v)$.

As shown in Figure 4, both model scales exhibit a consistent correction. On the verified reasoning trajectories, localization tokens shift from strongly negative to positive supervision under PA-OPSD, while observation tokens continue to receive positive supervision (Figure 4a). Under the on-policy distribution, the lower-tail ratio remains mostly below or close to 1 and no longer increases toward the strongly negative tail (Figure 4b), indicating that localization-related tokens are no longer disproportionately concentrated among strongly negative optimization signals. Together, these results show that PA-OPSD substantially mitigates the localization-specific optimization bias while preserving positive supervision for observation-related reasoning.

7 Conclusion

Our results reveal an unexpected localization–observation tradeoff in local-view on-policy self-distillation: local-view supervision improves fine-grained observation while degrading localization capability under full-image inference. Our theoretical and token-level empirical analyses show that partial observation induces a localization-specific negative optimization bias under local-view distillation. PA-OPSD mitigates this localization-specific optimization bias through partial-observation-aware calibration and a complementary box-guided teacher. Experiments across fine-grained visual reasoning and localization-oriented benchmarks show that PA-OPSD consistently improves over local-view OPSD and other on-policy baselines, recovering localization capability while preserving the observation benefits of local-view distillation.

AI USE STATEMENT

Generative AI tools were used to translate author-written text into English. They were not used to formulate research ideas or hypotheses, design the proposed method or experiments, or derive theoretical claims or proofs. GPT-5.6 Luna was used to generate candidate reference trajectories and annotate reasoning spans for the diagnostic analyses. Retained reference trajectories were manually verified, and annotation reliability was assessed through a human audit of 120 trajectories (Appendix D). Qwen3.5-9B was used as an LLM judge when rule-based evaluation was insufficient. Additionally, generative AI tools were used to assist with code modification, literature summarization, English language polishing, consistency checks between text, tables, and figures, and LaTeX formatting. All AI-assisted content was reviewed by the author. AI-assisted code modifications were further verified by running the complete experimental pipeline and cross-checking the resulting behavior and results against the original implementation. The author takes full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The experimental setup, including the training dataset and evaluation protocol, is described in Section 6.1. The training dataset and evaluation benchmarks are publicly available. The supplementary material includes training and evaluation code, experiment scripts, and data-processing scripts. Randomness is controlled using fixed random seeds. Complete proofs of the theoretical results are provided in Appendix B. The diagnostic trajectory-generation and annotation procedures, including prompts and human verification, are described in Appendix D. These procedures use GPT-5.6 Luna; reproducing the original generation and automatic-annotation steps requires access to that model.

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.

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.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, et al. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *arXiv preprint arXiv:2606.08719*, 2026.

Fangming Cui, Sunan Li, and Jiahong Li. A brief overview: On-policy self-distillation in large language models. *arXiv preprint arXiv:2605.18141*, 2026.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-adv: Exploring mllm visual reasoning in referring expression tasks. *arXiv preprint arXiv: 2602.23898*, 2026.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

Google. Gemini 3. `https://blog.google/products-and-platforms/products/gemini/gemini-3/`, 2025.

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

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, 2024.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Hongxing Li, Xiufeng Huang, Dingming Li, Wenjing Jiang, Zixuan Wang, Haolei Xu, Hanrong Zhang, Haiwen Hong, Longtao Huang, Hui Xue, et al. Perceive-to-reason: Decoupling perception and reasoning for fine-grained visual reasoning. *arXiv preprint arXiv:2607.01191*, 2026a.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026b.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. Groundingme: Exposing the visual grounding gap in mllms through multi-dimensional evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2412–2422, June 2026c.

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*, 2026d.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 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, Y. K. Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guobin Shen, Xiang Cheng, Chenxiao Zhao, Lei Huang, Jindong Li, Dongcheng Zhao, and Xing Yu. Anti-self-distillation for reasoning rl via pointwise mutual information. *arXiv preprint arXiv:2605.11609*, 2026a.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026b.

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.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b.

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, 2025c.

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.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*, 2026.

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

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

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.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, 2025a.

Yi-Fan 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? *arXiv preprint arXiv:2408.13257*, 2024.

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*, 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, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

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.

A PER-SUBTASK RESULTS ON MME-REALWORLD

Table 4 reports the performance of all compared methods on each of the 39 subtasks in MME-RealWorld. Following the analysis in Section 4.1, we group the subtasks into three categories according to their primary capability requirements: *Localization*, *Observation*, and *Other Reasoning*. Category-wise overall accuracies are also reported to facilitate comparison.

Overall, Vision-OPD yields substantial gains on observation-oriented subtasks, while its performance degrades on localization subtasks. In contrast, PA-OPSD consistently improves localization performance while preserving the gains on observation-oriented subtasks.

B f-DIVERGENCE POLICY-GRADIENT COEFFICIENTS

This section derives the optimization coefficients induced by a broad family of f-divergence objectives. We show that, under mild regularity conditions, the sign of the scalar coefficient in the resulting policy-gradient representation is completely determined by the teacher–student log-ratio. We then instantiate the general result for Reverse KL, Forward KL, and the Jensen–Shannon divergence.

B.1 GENERAL RESULT

Let $\mathcal{V}$ denote the vocabulary. Let the teacher distribution P_T be fixed with respect to θ (i.e., stop-gradient), and let the student distribution P_S be parameterized by θ. We consider the f-divergence

$$D_f(P_T \| P_S) = \sum_{v \in \mathcal{V}} P_S(v) f\left(\frac{P_T(v)}{P_S(v)}\right). \tag{8}$$

Assume that $P_T(v) > 0$ and $P_S(v) > 0$ for all $v \in \mathcal{V}$. Define the teacher–student log-ratio

$$u(v) = \log P_T(v) - \log P_S(v), \tag{9}$$

and the corresponding probability ratio

$$r(v) = \frac{P_T(v)}{P_S(v)} = e^{u(v)}. \tag{10}$$

Proposition B.1. *Suppose that $f\colon (0,\infty) \to \mathbb{R}$ is twice continuously differentiable and satisfies*

$$f(1) = 0, \qquad f''(r) > 0 \quad \textit{for all } r > 0.$$

Then the negative gradient of D_f admits the representation

$$-\nabla_\theta D_f = \mathbb{E}_{v \sim P_S}\left[A_f(u(v)) \nabla_\theta \log P_S(v)\right], \tag{11}$$

where

$$A_f(u) = e^u f'(e^u) - f(e^u) - f'(1). \tag{12}$$

Moreover,

$$A_f(0) = 0, \tag{13}$$

and

$$\frac{dA_f(u)}{du} = e^{2u} f''(e^u) > 0. \tag{14}$$

Consequently,

$$\operatorname{sign}(A_f(u)) = \operatorname{sign}(u). \tag{15}$$

Thus, for every $v \in \mathcal{V}$, the scalar coefficient $A_f(u(v))$ has the same sign as the corresponding teacher–student log-ratio $u(v)$.

Proof. From Eq. (10), and since P_T is fixed with respect to θ,

$$\nabla_\theta r(v) = -r(v) \nabla_\theta \log P_S(v).$$

Table 4: Per-task accuracy (%) on the 39 MME-RealWorld subtasks using Qwen3.5-4B, grouped by primary capability. Overall accuracies are computed over all examples within the corresponding group. Best results are shown in **bold**, and second-best results are underlined.

Category	Subtask	Base	GRPO	Vision-OPD	PA-OPSD
Localization					
Remote Sensing	Object Position	70.33	79.79	67.94	**82.33**
Monitoring	Vehicle Location	16.91	**55.15**	36.03	47.79
	Overall	65.11	77.38	64.82	**78.96**
Observation					
Autonomous Driving	Multi-Pedestrian Motion	20.49	36.71	37.32	**39.35**
Autonomous Driving	Multi-Vehicle Motion	34.75	41.19	**48.60**	47.27
Autonomous Driving	Pedestrian Motion	36.59	34.15	**49.39**	48.78
Autonomous Driving	Vehicle Motion	50.63	54.43	61.39	**63.29**
Autonomous Driving	Traffic Signal Attribute	51.74	76.12	78.11	**79.60**
Autonomous Driving	Object Counting	35.14	41.81	44.31	**51.25**
Autonomous Driving	Object Identification	28.88	63.76	64.58	**70.21**
Monitoring	Person Counting	49.09	38.71	39.42	**49.60**
Monitoring	Vehicle Counting	63.82	45.23	50.66	**65.13**
Monitoring	Person Color	25.84	30.34	49.44	**59.55**
Monitoring	Person Orientation	0.00	26.32	**31.58**	**31.58**
Monitoring	Vehicle Color	32.49	35.03	54.82	**65.48**
Monitoring	Vehicle Orientation	4.52	23.87	35.48	**41.29**
Remote Sensing	Object Color	60.56	65.67	72.43	**73.78**
Remote Sensing	Object Counting	37.77	35.56	37.36	**41.68**
OCR with Complex Context	Advertisement and Product OCR	81.07	87.74	92.04	**92.68**
OCR with Complex Context	Book, Map, and Poster OCR	85.92	90.35	91.96	**93.12**
OCR with Complex Context	License Plate OCR	89.08	88.97	93.08	**93.78**
OCR with Complex Context	Phone and Address OCR	78.86	87.18	**91.85**	91.68
OCR with Complex Context	Text Recognition	77.05	86.14	91.15	**91.74**
Diagram and Table	Diagram Reading	91.31	89.61	91.80	**92.30**
Diagram and Table	Table Reading	88.83	85.02	88.13	**89.72**
	Overall	67.06	70.32	74.05	**76.76**
Other Reasoning					
Autonomous Driving	Traffic Signal Attention	61.75	60.83	64.98	**71.89**
Autonomous Driving	Ego Intention Prediction	8.88	29.28	**39.14**	35.20
Autonomous Driving	Pedestrian Intention Prediction	42.72	22.33	63.11	**78.64**
Autonomous Driving	Vehicle Intention Prediction	36.23	47.83	49.28	**53.14**
Autonomous Driving	Ego–Pedestrian Interaction	21.70	29.25	**38.68**	32.08
Autonomous Driving	Ego–Traffic Signal Interaction	28.57	24.76	38.10	**44.76**
Autonomous Driving	Ego–Vehicle Interaction	28.71	30.69	41.58	**42.57**
Autonomous Driving	Object–Object Interaction	18.41	19.40	27.86	**31.84**
Monitoring	Counting Calculation	36.00	27.33	43.33	**46.67**
Monitoring	Intention Reasoning	25.51	37.76	55.10	**63.27**
Monitoring	Object Property Reasoning	31.00	60.00	65.00	**74.00**
OCR with Complex Context	Character Identification	63.60	64.40	66.40	**72.00**
OCR with Complex Context	Scene Understanding	66.40	69.60	76.40	**82.80**
Diagram and Table	Diagram Reasoning	82.18	72.99	83.33	**83.91**
Diagram and Table	Table Reasoning	85.89	61.35	84.66	**89.57**
	Overall	46.13	46.13	57.46	**61.33**
Overall		64.42	67.83	71.51	**75.03**

Applying the product rule to Eq. (8) gives

$$\begin{aligned}\nabla_\theta D_f &= \sum_{v\in\mathcal{V}} \Big[\nabla_\theta P_S(v)\, f(r(v)) + P_S(v) f'(r(v))\nabla_\theta r(v)\Big] \\ &= \sum_{v\in\mathcal{V}} P_S(v)\left[f(r(v)) - r(v)f'(r(v))\right]\nabla_\theta \log P_S(v). \end{aligned} \tag{16}$$

Equivalently,

$$-\nabla_\theta D_f = \mathbb{E}_{v\sim P_S}\left[(r(v)f'(r(v)) - f(r(v)))\,\nabla_\theta \log P_S(v)\right].$$

The score function has zero expectation:

$$\mathbb{E}_{v\sim P_S}\left[\nabla_\theta \log P_S(v)\right] = 0.$$

Therefore, subtracting the constant $f'(1)$ from the scalar coefficient does not change the gradient. Using $r(v) = e^{u(v)}$, we obtain

$$-\nabla_\theta D_f = \mathbb{E}_{v\sim P_S}\left[\Big(e^{u(v)}f'(e^{u(v)}) - f(e^{u(v)}) - f'(1)\Big)\,\nabla_\theta \log P_S(v)\right],$$

which is exactly Eq. (11) with A_f defined in Eq. (12).

Since $f(1) = 0$,

$$A_f(0) = f'(1) - f(1) - f'(1) = 0.$$

Differentiating Eq. (12) yields

$$\begin{aligned}\frac{dA_f(u)}{du} &= \frac{d}{du}\left(e^u f'(e^u) - f(e^u)\right) \\ &= e^u f'(e^u) + e^{2u} f''(e^u) - e^u f'(e^u) \\ &= e^{2u} f''(e^u).\end{aligned}$$

Because $e^{2u} > 0$ and $f''(e^u) > 0$ for all $u \in \mathbb{R}$, it follows that $dA_f(u)/du > 0$. Hence A_f is strictly increasing. Together with $A_f(0) = 0$, this implies

$$\operatorname{sign}(A_f(u)) = \operatorname{sign}(u),$$

which proves the result. □

B.2 Special Cases

We next instantiate Proposition B.1 for three commonly used divergences.

Reverse KL. Consider

$$D_{\mathrm{RKL}} = D_{\mathrm{KL}}(P_S\|P_T). \tag{17}$$

A corresponding f-divergence generator is

$$f(r) = -\log r.$$

Substituting this generator into Eq. (12) gives

$$A_{\mathrm{RKL}}(u) = u. \tag{18}$$

Therefore,

$$-\nabla_\theta D_{\mathrm{RKL}} = \mathbb{E}_{v\sim P_S}\left[u(v)\nabla_\theta \log P_S(v)\right]. \tag{19}$$

Forward KL. Consider

$$D_{\mathrm{FKL}} = D_{\mathrm{KL}}(P_T\|P_S). \tag{20}$$

A corresponding generator is

$$f(r) = r\log r.$$

Substituting into Eq. (12) gives

$$A_{\mathrm{FKL}}(u) = e^u - 1. \tag{21}$$

Equivalently, for a token v,

$$A_{\mathrm{FKL}}(u(v)) = \frac{P_T(v)}{P_S(v)} - 1.$$

Hence,

$$-\nabla_\theta D_{\mathrm{FKL}} = \mathbb{E}_{v\sim P_S}\left[\Big(e^{u(v)} - 1\Big)\,\nabla_\theta \log P_S(v)\right]. \tag{22}$$

Jensen–Shannon Divergence. The Jensen–Shannon divergence is

$$D_{\mathrm{JS}} = \frac{1}{2} D_{\mathrm{KL}}(P_T \| M) + \frac{1}{2} D_{\mathrm{KL}}(P_S \| M), \tag{23}$$

where

$$M = \frac{1}{2}(P_T + P_S).$$

A corresponding f-divergence generator is

$$f(r) = \frac{1}{2}\left[r \log \frac{2r}{1+r} + \log \frac{2}{1+r} \right]. \tag{24}$$

Substituting into Eq. (12) yields

$$A_{\mathrm{JS}}(u) = \frac{1}{2} \log \frac{1+e^u}{2}. \tag{25}$$

Therefore,

$$-\nabla_\theta D_{\mathrm{JS}} = \mathbb{E}_{v \sim P_S}\left[\frac{1}{2} \log \frac{1+e^{u(v)}}{2} \nabla_\theta \log P_S(v) \right]. \tag{26}$$

Under the policy-gradient representation in Proposition B.1, these divergences differ through the functional form of their scalar coefficient A_f. Nevertheless, all three coefficients satisfy

$$\mathrm{sign}(A_f(u)) = \mathrm{sign}(u).$$

Hence, the teacher–student log-ratio alone determines the sign of a token's scalar coefficient in the policy-gradient representation. In this sense, tokens with positive teacher–student log-ratios are encouraged by the distillation objective, whereas tokens with negative log-ratios are suppressed.

C Empirical Validation of the Local-View Teacher Approximation

The approximation in Eq. (5) replaces the frozen local-view teacher with the corresponding student under the same local-view observation. To quantify the error introduced by this approximation, we decompose the original teacher–student log-ratio as

$$u^t_{\mathrm{LV}}(v) = u^t_{\mathrm{PMI}}(v) + \delta^t(v), \tag{27}$$

where

$$u^t_{\mathrm{PMI}}(v) = \log \frac{\pi_\theta(v \mid C, q, H_t)}{\pi_\theta(v \mid C, O, q, H_t)} = -i_\theta(v; O \mid C, q, H_t), \qquad \delta^t(v) = \log \frac{\pi_{\bar{\theta}}(v \mid C, q, H_t)}{\pi_\theta(v \mid C, q, H_t)}. \tag{28}$$

Here, $u^t_{\mathrm{PMI}}(v)$ is the approximation used in Eq. (5), while $\delta^t(v)$ isolates the model discrepancy between the frozen teacher and the trained student under an identical local-view observation.

We evaluate this residual using a Vision-OPD model initialized from Qwen3.5-4B as the student. We randomly sample 200 examples from the Vision-OPD-6K training set and generate full-image student rollouts using greedy decoding. For each sampled token along these rollouts, we compute the log-ratio twice while keeping the full-image student distribution fixed: once using the frozen initialization model as the local-view teacher, and once using the trained Vision-OPD model itself under the same local-view input. The difference between the two log-ratios is exactly $\delta^t(Y_t)$.

We find that the residual is small, with a mean absolute value of $\mathbb{E}[|\delta^t(Y_t)|] = 0.0369$ over sampled tokens. This corresponds to only 8.11% of the average magnitude of the original log-ratio, $\mathbb{E}[|u^t_{\mathrm{LV}}(Y_t)|]$. These results indicate that the teacher–student model discrepancy under the same local-view observation is relatively small in our setting, providing empirical support for the approximation used in Eq. (5).

D Details of the Diagnosis Experiments

This appendix describes the implementation details of the two diagnosis analyses presented in Section 4.3, including the construction of the on-policy and reference reasoning trajectories, as well as the annotation procedure for reasoning behaviors.

```
<image>
{question}

Before providing your final answer, provide a concise reasoning process grounded in the
visual evidence from the image. Keep your reasoning within 300 words.

Answer in exactly this format:

<your reasoning, up to 300 words>
Final: <one option letter only>
```

Figure 5: Prompt used to generate candidate reference trajectories with GPT-5.6 Luna.

```
<image>

Task:
Label non-overlapping character spans in the exact response text. Use only LOC and OBS. Leave all other text unlabeled.

Labels:
- LOC: where the relevant region/object is in the full image, including spatial positions, landmarks, crop/box references, or
relations used to find it.
- OBS: direct visual evidence that supports the answer, including objects, attributes, colors, shapes, materials, counts,
readable text, or the key visual fact.

Do not label:
- Boilerplate or reasoning scaffolds such as "I see", "therefore", headings, or markdown numbering.
- Final answer choices, option letters, option text, or answer-selection statements.
- Distractors, search steps, articles, most prepositions, conjunctions, and punctuation.

Rules:
- Use the smallest meaningful spans.
- Prefer several short spans over a whole sentence.
- Character offsets must point into the exact response text.
- `text` must exactly equal `response[start:end]`.

Return JSON only:
{
  "spans": [
    {"start": 0, "end": 0, "label": "LOC", "text": ""}
  ]
}

Question: {question}
Answer: {answer}
```

Figure 6: Prompt for LOC/OBS span annotation with GPT-5.6 Luna; unlabeled text is treated as OTHER.

D.1 On-Policy Analysis

For the on-policy analysis, we randomly sample 200 question–answer pairs from the Vision-OPD-6K training set and generate reasoning trajectories with Qwen3.5-4B and Qwen3.5-9B using vLLM (Kwon et al., 2023) under the same prompting template as used during training. Generation uses greedy decoding with a maximum length of 2048 tokens. For each trajectory, we record the complete reasoning process together with the teacher and student token probabilities, and compute the statistics reported in Section 4.3 using the sampled-token log-ratio $u^t_{\mathrm{LV}}(Y_t)$.

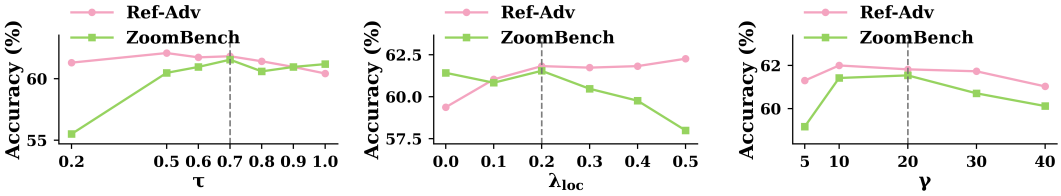
D.2 Reference-Trajectory Analysis

For the reference-trajectory analysis, we construct a separate set of 200 verified reasoning trajectories. Candidate trajectories are generated by GPT-5.6 Luna (OpenAI, 2026) using a prompting template that instructs the model to produce a concise reasoning process grounded in the visual evidence from the image before providing the final answer. The complete prompt is shown in Figure 5.

Each generated trajectory is manually verified, and only examples whose reasoning process and final answer are both correct are retained. We repeat this generation-and-verification procedure until 200 valid trajectories are collected. Teacher–student log-ratios are then computed under teacher forcing on the verified reference prefixes, yielding the statistics reported in Section 4.3.

Table 5: Human validation of the GPT-5.6 reasoning-behavior annotations on 120 sampled trajectories. Metrics are computed after mapping span-level annotations to individual tokens. "LOC–OBS Confusion" denotes the fraction of human-annotated LOC or OBS tokens that are assigned to the opposite category by GPT-5.6.

Metric	Score
Accuracy	0.890
Macro-F1	0.868
LOC–OBS Confusion	0.045



(a) Calibration threshold τ. (b) Localization weight λ_{loc}. (c) Calibration sharpness γ.

Figure 7: Hyperparameter sensitivity of PA-OPSD on Qwen3.5-4B, evaluated on ZoomBench and Ref-Adv. Each hyperparameter is varied independently while the others are fixed to their default values.

D.3 Reasoning Behavior Annotation

For both analyses, GPT-5.6 Luna identifies semantically coherent reasoning spans and labels them as LOC or OBS. Unlabeled text is treated as OTHER, yielding three functional categories:

- `LOC`: reasoning used to identify or spatially ground the question-relevant region;
- `OBS`: reasoning describing visual evidence within the localized region;
- `OTHER`: all remaining reasoning, including intermediate logical deductions and answer composition.

The span labels are subsequently mapped to individual tokens for all statistical analyses reported in Section 4.3. The annotation prompt specifies the semantic role of each category. The complete prompt is provided in Figure 6. The same annotation protocol and label definitions are used for both analyses.

To assess the reliability of the automatic reasoning-behavior annotations, we conduct a human audit on 120 randomly sampled reasoning trajectories. A human annotator independently labels each trajectory at the span level using the same definitions of LOC, OBS, and OTHER as those provided to GPT-5.6. The human and automatic span annotations are subsequently mapped to individual tokens using the same procedure as in our main analysis. We evaluate agreement using token-level accuracy and macro-F1, and the confusion rate between LOC and OBS.

As shown in Table 5, the automatic annotations achieve an overall token-level accuracy of 0.890 and a macro-F1 of 0.868, indicating substantial agreement with the human annotations. Moreover, confusion between LOC and OBS is only 4.5%, suggesting that errors in the automatic annotations rarely correspond to directly exchanging the two behavior categories whose optimization patterns are contrasted in Section 4.3. Overall, the human audit supports the use of the automatic annotations for our token-level behavioral analysis.

E Hyperparameter Sensitivity

Figure 7 examines the sensitivity of PA-OPSD to the calibration threshold τ, localization weight λ_{loc}, and calibration sharpness γ. We vary each hyperparameter independently while keeping the remaining settings fixed to their default values.

Input Image



Question

<image>
What is the primary color of the row of signs below the top green signs?
A. Yellow
B. Blue
C. White
D. Red

Answer of Vision-OPD

The primary color of the row of signs below the top green signs is **Red**.

Looking at the image, there's a prominent red sign with white text and graphics (possibly a restaurant or business sign) situated directly beneath the green signs. The red color is dominant in that row.

Correct Answer: **D. Red**

Answer of PA-OPSD

Looking at the image, we can see a row of signs below the top green signs. These signs are located on the right side of the street, just beneath the green signage that reads "2 98 0円" and similar text.

The signs directly below the green ones are **yellow** with black or dark lettering — they appear to be store or business signs (e.g., "占い館", "タマ", etc.).

Therefore, the correct answer is: **A. Yellow**

Figure 8: A qualitative comparison between Vision-OPD and PA-OPSD. Vision-OPD misidentifies the question-relevant evidence and produces an incorrect answer, whereas PA-OPSD identifies the relevant region and uses the corresponding visual evidence to answer correctly.

For the calibration threshold τ, performance remains relatively stable over a broad range, indicating that the method does not rely on a narrowly tuned transition point for attenuating negative local-view supervision. Similarly, varying γ produces only modest changes in performance, showing that PA-OPSD is not sensitive to the precise sharpness of the calibration function.

The localization weight λ_{loc} exhibits a clearer tradeoff. Increasing it initially improves localization performance on Ref-Adv, while an excessively large value reduces performance on ZoomBench. This behavior is consistent with the role of λ_{loc} in balancing complementary localization guidance against the observation supervision provided by the local-view teacher.

Overall, PA-OPSD remains stable across a broad range of hyperparameter settings. The default values $\tau = 0.7$, $\lambda_{\text{loc}} = 0.2$, and $\gamma = 20$ provide a favorable balance between fine-grained reasoning and localization performance.

F Qualitative Analysis

F.1 Localization Case Study

Figure 8 provides a concrete example of the localization difference between Vision-OPD and PA-OPSD. Vision-OPD refers to a red sign below the green signage and incorrectly predicts "D. Red". PA-OPSD instead identifies the relevant row of signs on the right side of the street, below the green signage, and determines their color from that evidence. This example illustrates how stronger local-

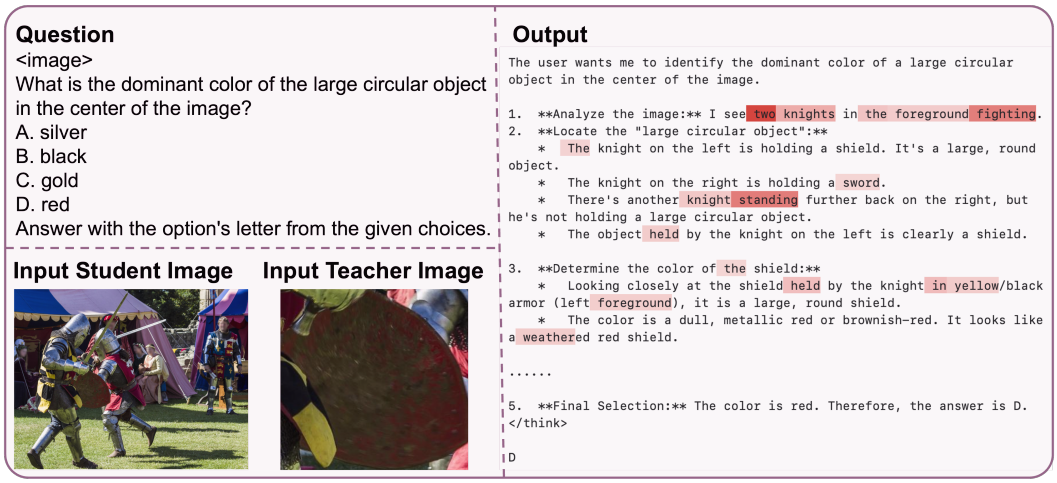


Figure 9: Token-level visualization of strongly negative optimization signals under local-view distillation. Tokens with local-view teacher–student log-ratio $u_{\mathrm{LV}} < -0.5$ are highlighted in red, with darker shading indicating more negative values.

ization behavior allows PA-OPSD to identify and use the question-relevant evidence in a visually cluttered scene.

F.2 Token-Level Diagnosis

Figure 9 illustrates the token-level pattern identified in Section 4.3 on Qwen3.5-4B. The question concerns the color of a large circular object, and the local crop provides a clear view of the shield itself while omitting much of the surrounding scene context. Among the tokens receiving strongly negative signals ($u_{\mathrm{LV}} < -0.5$), many occur when the model disambiguates the relevant object from other candidates in the full image, including phrases describing "two knights," the "foreground," another "knight standing" farther back, and the object "held" by the target knight. These tokens encode which object to attend to and how it is situated relative to other scene elements.

The visualization therefore provides a concrete example of why localization behavior is particularly vulnerable to partial-observation supervision: the local-view teacher observes the target shield clearly, but lacks the broader context needed to support the student's scene-level localization and spatial disambiguation. Although some strongly negative tokens are unrelated to localization, localization-related tokens are prominently represented among the highlighted positions, consistent with the distributional pattern in Figure 2b.