CV-OPD: CONTRASTIVE VISION ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) reveals versatility and convenience during VLM post-training, providing dense supervision for student from teacher. However, we observed an anomaly that the distilled VLM could still answer the questions accurately even when the related visual regions were masked or replaced with noise. Building on this observation, our quantitative and qualitative experiments demonstrate that OPD, which uses the answer as a hint, has failed to teach students to genuinely seek out and rely on visual evidence. Our theoretical analysis suggests that this failure arises from a *shortcut enabled by the lack of visual constraints* during distillation. To address the issue, we introduce **C**ontrastive **V**ision **O**n-**p**olicy **D**istillation (CV-OPD), which compares teacher distributions induced by positive and negative crops while conditioning on the same student prefix and augments optimization with a margin-based contrastive constraint, which is modulated with discrepancy gating across token and sample level. Specifically, CV-OPD first introduces a negative teacher unrelated to the answer and designs a Contrastive Divergence Loss (CDL) that constrains the feature distribution of student by contrasting the positive and negative teachers, correcting how the student exploits visual regions. The CV-OPD introduces a Discrepancy Aware Gate (DAG) that uses token-level and sample-level discrepancies to select visually relevant tokens for optimization, directing the student's attention toward visual content. Extensive experiments across a broad range of fine-grained visual benchmarks demonstrate that CV-OPD substantially alleviates shortcut and consistently improves the visual reasoning capabilities of existing VLMs. All the reproduction code will be released on `https://anonymous.4open.science/r/CV-OPD-9D04/`.

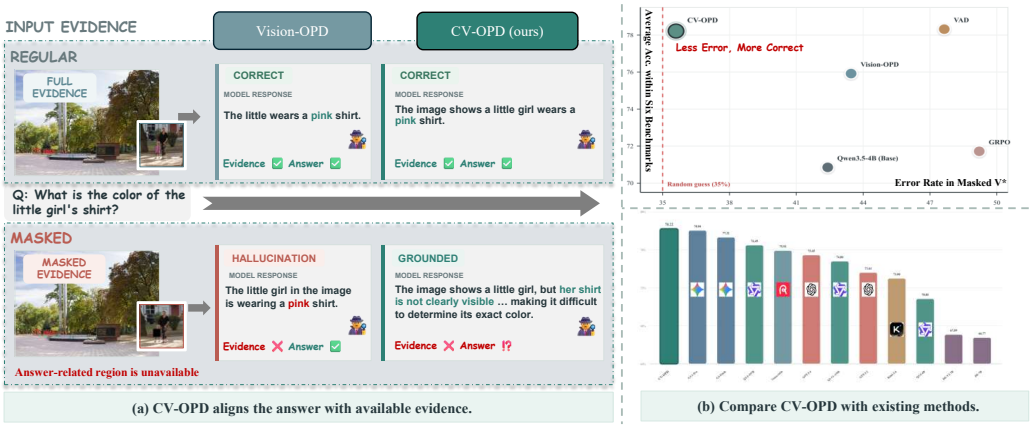


Figure 1: (a) The failure mode of Vision-OPD. The distilled VLM can generate correct answer even given wrong visual inputs. (b) Our CV-OPD shows less error with higher accuracy, which is even competitive while compared to closed-source VLMs.

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

1 INTRODUCTION

In recent LLM post-training, on-policy distillation (OPD) (Agarwal et al., 2024) has gained increasing attention as a means of addressing the lack of token-level supervision in reinforcement learning (Shao et al., 2024). By training the student to match the teacher's output distribution over its own on-policy generations, OPD effectively transfers the teacher's capabilities while alleviating the distribution mismatch between training and inference. A common practice is to equip the teacher with privileged information (Liu et al., 2026; Yuan et al., 2026), enabling it to identify and correct errors in the student's responses. These properties make OPD particularly well suited to handling the challenging samples encountered in Vision Language Model (VLM) post-training (Yuan et al., 2026; Sun et al., 2026).

Naturally, recent works, *e.g.*, Vision-OPD (Yuan et al., 2026), VA-OPD (Liu et al., 2026) and so on, have extended OPD to VLMs by providing the teacher with cropped or enlarged target regions to correct erroneous reasoning trajectories of the student, yielding substantial performance gains. However, as shown on Fig. 1, we uncover a significant issue of the distilled VLMs. In a simple diagnostic experiment, we mask the image regions relevant to the answer and evaluate the distilled model.

Surprisingly, Vision-OPD still generates correct answers without necessary visual evidence, achieving 43.46% and 48.69% accuracy on V^* (Wu & Xie, 2024) under black masking and Gaussian noise respectively, substantially exceeding the random rate for multiple-choice questions. This result suggests that directly exposing answer-relevant information to the teacher induces a **visual shortcut**, through which the VLM can bypass the underlying visual evidence to produce answers. In Sec. 2, we examine this phenomenon in detail and provide a formal derivation of its possible causes.

To address this limitation, we propose **C**ontrastive **V**ision **O**n-**p**olicy **D**istillation (CV-OPD), which constructs negative visual samples to explicitly discourage reliance on misleading regions. By contrasting positive and negative visual evidence, CV-OPD encourages the model to learn not only **what answer to produce**, but also **where to look and what to avoid**, thereby refining the solution space of response generation.

Specifically, CV-OPD introduces a **negative teacher** receives a selected image region that is unrelated with answer, and generates distractive content conditioned on the student's sampled trajectory. The difference between two teachers naturally captures model biases and style-related shortcuts that are irrelevant to the visual evidence. We then introduce a Contrastive Divergence Loss (CDL), which explicitly steers the student's token distribution away from the negative teacher with hinge margin. CDL encourages the model to not merely imitate the trajectory to answer the question, but to identify and reject visually irrelevant evidence. Finally, we propose a Discrepancy Aware Gate (DAG) to select token sequences that are highly relevant to the reasoning process. The constraint is applied only to these selected tokens, directing the optimization towards useful visual contents.

We evaluate CV-OPD on a broad range of fine-grained visual benchmarks following the previous work (Yuan et al., 2026), where it consistently delivers substantial improvements over vanilla OPD. More importantly, experiments in masked V^* show that CV-OPD is significantly more grounded in visual evidence.

In summary, our contributions can be concluded into three-fold:

- We uncover a previously overlooked shortcut phenomenon in OPD for VLMs and demonstrate that it primarily arises from directly exposing answer-relevant visual information to the teacher.
- We propose **C**ontrastive **V**ision **O**n-**p**olicy **D**istillation (CV-OPD), which leverages negative teachers constructed from randomly sampled visual regions. The proposed CDL and DAG encourage the model to learn not only the correct answer, but also an evidence-grounded reasoning trajectory for identifying its visual source.
- CV-OPD consistently improves performance across a broad range of visual benchmarks comapred with the base model. Furthermore, experiments in masked V^* demonstrate that CV-OPD substantially strengthens the reliance on visual evidence for VLMs.

2 OBSERVATION AND ANALYSIS

2.1 PRELIMINARY

On-policy distillation (Agarwal et al., 2024) provides dense supervision by leveraging a stronger teacher to reshape the token-level output distributions along trajectories sampled by the student. In VLM post-training, this paradigm typically equips the teacher with privileged visual hints to identify and correct errors in the student's generated responses. For each sample, giving the original visual content V and question Q, student first sample the t-th token $\pi_t(k)$ as:

$$\pi_t(k) = \pi_\theta(A_t = k \mid V, C_t), \tag{1}$$

where C_t is the context including V and sampled prefix $A_{<t}$ and A_t means the t-th token in output answer. k is the token index of the dictionary. Compared with student, the difference of privileged teacher is the extra visual hint of inputs. The distribution of teacher can be formulated as:

$$\tau_t^+(k) = \tau_\phi(A_t = k \mid C_t, H^+), \tag{2}$$

while the H^+ is the hint information, which is cropped or zoom-in region related with answers. Based on that, the standard optimization target of OPD in all data $\mathcal{D}$ can be formulated as:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{A_{1:T}\sim\pi_{\bar{\theta}}}\left[\sum_{t=1}^{T} w_t D_{\mathrm{JS}}(\pi_t, \tau_t^+)\right]. \tag{3}$$

The rollout distribution and teacher distribution are treated as stop-gradient during the optimization and only the parameters θ of student is optimized.

2.2 OBSERVATIONS

Although OPD appears promising, we identified the following observations in the context of VLMs.

Observation 1. *The distilled VLM can handle questions without visual evidence.*

We replace the answer-relevant regions in the V* (Wu & Xie, 2024) with black pixels or Gaussian noise and ensure no extra interference from ambiguity, then evaluate Vision-OPD (Yuan et al., 2026) on the processed images. As shown in Fig. 2 (a), Vision-OPD achieves 43.46% and 48.69% accuracy under black masking and Gaussian noise, respectively, exceeding the base model's 42.41% and 42.93%. Despite the removal of answer-relevant visual information, the distilled model becomes more accurate at predicting the answers. These results suggest that visual self-distillation may introduce shortcut learning (Geirhos et al., 2020), allowing the model to produce correct answers without properly grounding in visual evidence.

Observation 2. *The attention of distilled VLM is unrelated with vision content.*

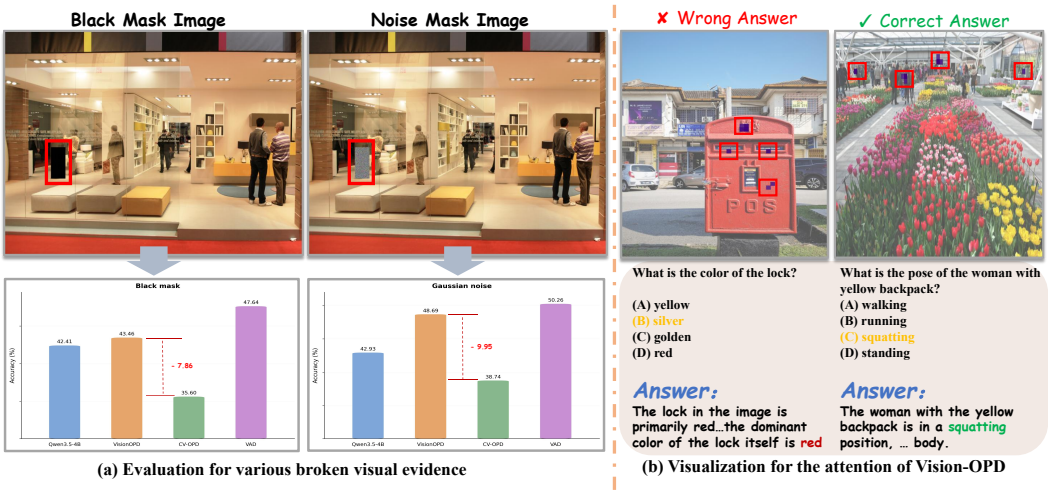
We visualize the attention of self-distilled Vision-OPD both on the correct and wrong situation. As shown in Fig. 2 (b), the model fails to consistently focus on the target regions in both correctly and incorrectly answered examples. Even a correct prediction does not necessarily mean that the model has identified the relevant visual evidence. This reveals a gap between answer accuracy and visual grounding in self-distilled VLMs. More visualizations and analysis can be found in A.4 and A.5.

2.3 ANALYSIS

These observations naturally lead us to a more fundamental question, **what drives the emergence of such shortcut during distillation?** We therefore formalize the distillation process and revisit its underlying optimization objective to identify the source of this behavior.

First, define H^+ as the hint for positive teacher. It's obviously the mutual information between H^+ and answer is higher than original image. According to the optimization target Eq. 3, due to the symmetry of JSD (Lin, 1991), what the student learns is the conditional Fréchet Mean of the privileged teacher. Therefore, student retains the same policy of privileged teacher token posterior:

$$\arg\max_k \pi^*_{t,k} \approx \arg\max_k \tau^+_{t,k}. \tag{4}$$



(a) Evaluation for various broken visual evidence

(b) Visualization for the attention of Vision-OPD

Figure 2: Diagnostic evaluation of visual shortcut. (a) We evaluate accuracy under both black masking and Gaussian noise, where answer-relevant visual regions are replaced. (b) We further visualize the attention (red rectangle) of Vision-OPD to find out whether the correct visual information is utilized.

Based on this, the gradient of JSD with respect to the student logits z_t can be expressed as:

$$\nabla_{z_t} D_{\mathrm{JS}}(\pi_t, \tau_t^+) = \frac{1}{2}\left(\mathrm{Diag}(\pi_t) - \pi_t \pi_t^\top\right) \log \frac{2\pi_t}{\pi_t + \tau_t^+}. \tag{5}$$

It's obviously that JSD will increase the logit of hint related target token. At the same time, the student distribution is decided with visual content V, question Q, and prefix $A_{<t}$ as shown on Eq. 1. Such gradients impose no constraint on what the student should optimize. Compared with the target from Eq. 2, because the mapping from V to H^+ is neither differentiable nor observable, optimization instead resorts to the full image and the fixed textual context. Consequently, the model learns to generate outputs by fitting the available conditions rather than actively identifying positive region R^+.

3 Contrastive Vision On-policy Distillation

To mitigate the shortcut during distillation, we propose Contrastive Vision On-policy Distillation (CV-OPD). As the crown jewels, CV-OPD introduces a **negative teacher** that receives randomly sampled visual regions irrelevant to the answer, and employs a Contrastive Divergence Loss (CDL) and Discrepancy Aware Gate (DAG) to encourage the model to distinguish visual evidence from irrelevant cues.

3.1 Overview

As shown in Fig. 3, given the vision inputs and text prompt, for each example, a frozen copy of the current student, π_θ, generates M on-policy trajectories from the same image:

$$A_{1:T_i}^{(i)} \sim \prod_{t=1}^{T_i} \pi_\theta\left(A_t^{(i)} \mid V, Q, A_{<t}^{(i)}\right), \qquad C_{i,t} = (Q, A_{<t}^{(i)}), \quad i = 1, \ldots, M. \tag{6}$$

Constructed on the student trajectory, the positive region R^+ is drawn from an answer related proposal. Besides, we construct an extra pipeline for the negative region R^-, which is sampled from the same image, with its height and width matched to the positive region and its overlap constrained by a threshold δ:

$$\begin{aligned} R^+ &\sim \rho^+(\cdot \mid V, Q), \\ R^- &\sim \rho^-(\cdot \mid V, Q, R^+), \quad \text{where } \mathrm{IoU}\left(R^+, R^-\right) \le \delta. \end{aligned} \tag{7}$$

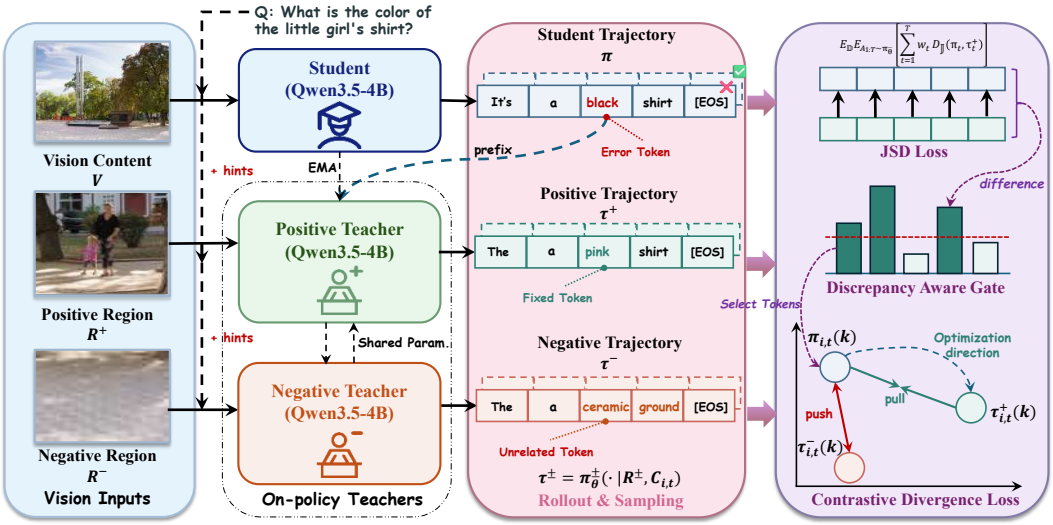


Figure 3: The pipeline of Contrastive Vision On-policy Distillation. CV-OPD introduces a negative teacher that provides unrelated visual information to fix the reliance of visual contents. Samples with at least one correct rollout are retained for training.

The ρ is the proposal mechanism to select the image region related to the conditions. The teacher branches share parameters and operate on the same prefix from student, differing only in the privileged crop. For the t-th token, the resulting teacher token distributions are:

$$\tau^+_{i,t} = \pi^+_\theta(\cdot|R^+, C_{i,t}), \tau^-_{i,t} = \pi^-_\theta(\cdot|R^-, C_{i,t}) \tag{8}$$

For the positive teacher distribution τ^+ and student distribution π, we rewrite the Jensen-Shannon divergence as the optimized target as shown on Eq. 3 with the average across all the rollouts from student. As for the negative teacher, we would take its advantages in the following sections.

3.2 Contrastive Divergence Loss

With the negative teacher in place, we introduce a contrastive divergence loss that explicitly encourages the model to capture the distributional differences induced by distinct visual prefixes, rather than indiscriminately aligning with the teacher's output.

Specifically, at each position (i, t), we compute the full-vocabulary Jensen–Shannon divergence between the student and each teacher distribution. The resulting positive and negative divergences are compared through a margin-based hinge loss:

$$\widetilde{d}^{\pm}_{i,t} = D_{\mathrm{JS}}\left(\pi_{i,t}, \tau^{\pm}_{i,t}\right), \qquad \mathcal{L}_{\mathrm{CDL}}(\theta) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{A_{1:T}\sim\pi_\theta}\left[\sum_{i=1}^{M} w_t(\mu + \widetilde{d}^{+}_{i,t} - \widetilde{d}^{-}_{i,t})\right]_+ . \tag{9}$$

Here, $[x]_+ = \max(x, 0)$ and $0 < \mu < \log 2$. Minimizing $\mathcal{L}_{\mathrm{CDL}}(\theta)$ draws the student toward the positive teacher while pushing it away from the negative teacher. Once $\widetilde{d}^{-}_{i,t} - \widetilde{d}^{+}_{i,t} \geq \mu$, the loss drops to zero, so the negative branch is not pushed away indefinitely.

In CDL, set $F(\cdot)$ is the Fisher Information Matrix in Eq. 5 and $u(\cdot, \cdot)$ is the log part. The gradient with respect to the student logits z_t can be expressed as:

$$\Delta z_t \propto F(\pi_t)\left[u(\pi_t, \tau^-_t) - u(\pi_t, \tau^+_t)\right]. \tag{10}$$

The exact form of the gradient is:

$$-\nabla_{z_t}\mathcal{L}_{CDL}(\theta) = \frac{1}{2}F(\pi_t)\log\frac{\pi_t + \tau^+_t}{\pi_t + \tau^-_t}, \tag{11}$$

which is proportional to $\tau_t^+ - \tau_t^-$. This finding indicates that CDL can itself amplify the sampling discrepancy between the positive and negative inputs, thereby helping the model learn to localize positive regions. The detailed proof of gradient can be viewed in A.1

3.3 Discrepancy Aware Gate

Go further, to alleviate gradient dilution arising from sampling noise and stylistic variations (Liang et al., 2026), we seek to concentrate the visual content related optimization signals of CV-OPD on tokens that are most relevant to the underlying visual evidence.

Therefore, we introduce Discrepancy Aware Gate (DAG), using the full-vocabulary divergence between student and positive teacher as the token-level evidence score to retain the useful tokens for the gradient backpropagation later within the batch:

$$\bar{d}_{\mathcal{B}} = \frac{1}{|\mathcal{T}_{\mathcal{B}}|} \sum_{(i,t)\in\mathcal{T}_{\mathcal{B}}} d_{i,t}, \qquad g_{i,t} = \mathbf{1}\left[d_{i,t} > \bar{d}_{\mathcal{B}}\right]. \tag{12}$$

Put simply, the contrastive term kicks in only at positions whose positive discrepancy exceeds the batch average. The gate $g_{i,t}$ selects sequence positions.

Beyond token-level filtering, we further consider the reliability of teacher supervision. Since the teacher samples conditioned on the student's prefix, the student's correctness directly influences the likelihood of obtaining a valid teacher correction, making low-quality student trajectories a potential source of noisy supervision. Let $r_i \in \{0, 1\}$ indicate whether the final answer from rollout i is correct. We define a sample-level reliability gate as

$$b = \mathbf{1}\left[\sum_{i=1}^{M} r_i > 0\right]. \tag{13}$$

An incorrect rollout can leave the teacher conditioning on an off-track prefix. If this happens for all M rollouts, we disable all the supervision term for that example.

3.4 Training Objective

Let $w_{i,t} \geq 0$ with $\sum_t w_{i,t} = 1$, and let $\lambda \geq 0$ weight the CDL term. Combining standard distillation and gated CDL gives

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\frac{b}{M}\sum_{i=1}^{M}\sum_{t=1}^{T_i}\left(w_{i,t}^{\mathrm{JSD}} d_{i,t} + \lambda g_{i,t} w_{i,t}^{\mathrm{CDL}}\left[\mu + \widetilde{d}_{i,t}^{+} - \widetilde{d}_{i,t}^{-}\right]_+\right)\right]. \tag{14}$$

The expectation covers training examples, positive and negative region sampling, and student rollouts. Since $b = \mathbf{1}[\sum_i r_i > 0]$, distillation is enabled only when at least one rollout is correct. In practice, we use $w_{i,t} = 1/T_i$.

4 Experiments

4.1 Setup

We train CV-OPD on Qwen3.5-4B (Qwen Team, 2026b) using 6.2K visual reasoning samples from Vision-OPD (Yuan et al., 2026). The primary distillation objective is Jensen–Shannon Divergence (JSD) with $\beta = 0.5$, computed over the Top-100 vocabulary entries and an aggregated tail bucket. An Exponential Moving Average (EMA) teacher tracks the student parameters at an update rate of 0.05. Each sample generates eight on-policy rollouts of up to 1024 tokens, sharing one negative crop for contrastive supervision. Samples with no correct rollouts are excluded from distillation.

Evaluation covers six fine-grained visual benchmarks: V* (Wu & Xie, 2024), ZoomBench (Wei et al., 2026a), HR-Bench (Wang et al., 2025) 4K and 8K, and MME-RealWorld (Zhang et al., 2024) EN and CN. All Qwen3.5-based models operate in non-thinking mode, and GPT-OSS-120B (OpenAI, 2025b) serves as the answer judge.

Table 1: Overall comparison on six fine-grained visual benchmarks. We report the models trained on the same data within the Qwen3.5-Based Models. Values are accuracies (%), and Avg_6 is the unweighted mean. Within the Qwen3.5-based block, best results are bold and second-best results are underlined; the cyan-green row denotes CV-OPD.

Model	Param.	$V^\star$	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg_6
Tool-Using and Region-Aware Models								
DeepEyes (Zheng et al., 2026)	7B	82.63	46.15	75.50	70.00	63.95	62.41	66.77
SenseNova-MARS (Chng et al., 2025)	8B	92.15	47.81	83.13	78.38	67.90	68.90	73.05
Vision-RL^2 (Shi et al., 2026)	7B	91.62	59.88	78.12	75.00	62.41	58.73	70.96
Closed-Source Models								
Gemini 3 Flash (Google DeepMind, 2025)	–	85.39	61.49	87.25	85.00	73.15	71.66	77.32
Gemini 3.1 Pro (Google DeepMind, 2026)	–	88.48	62.01	88.12	84.50	73.32	71.81	78.04
GPT-5.2 (OpenAI, 2025a)	–	84.59	54.71	84.64	78.48	71.82	67.43	73.61
GPT-5.4 (OpenAI, 2026)	–	85.83	57.94	86.15	78.10	74.16	70.50	75.45
Open-Source Models								
Qwen3.5-397B-A17B (Qwen Team, 2026a)	397B	84.23	58.30	91.25	84.37	74.17	66.64	76.49
Kimi-K2.5 (Kimi Team, 2026)	1T	85.86	56.33	81.87	75.38	71.51	68.40	73.23
Qwen3-VL-Instruct (Bai et al., 2025)	8B	84.82	42.96	79.63	75.25	63.19	64.61	68.41
Qwen3.5-Based Models								
Qwen3.5 (Qwen Team, 2026b)	4B	82.20	48.40	**86.38**	82.00	63.94	63.80	71.12
GRPO (Shao et al., 2024)	4B	81.68	56.57	76.75	75.62	71.10	68.58	71.72
VAD (Zhang et al., 2026)	4B	**91.62**	60.12	85.13	**83.38**	76.98	**71.44**	78.11
VA-OPD (Liu et al., 2026)	4B	86.91	60.36	81.62	78.12	73.29	67.42	74.62
V-Zero (Sun et al., 2026)	4B	88.48	56.45	84.88	78.75	73.49	70.45	75.42
Vision-OPD (Yuan et al., 2026)	4B	89.53	59.41	81.75	80.12	74.51	70.17	75.92
Decomposed OPD (Yoon et al., 2026)	4B	89.80	58.70	82.60	79.50	73.80	67.82	75.37
CV-OPD (Ours)	4B	**91.62**	**62.01**	84.38	**83.38**	**76.99**	70.93	**78.22**

4.2 Main Results

As shown on Tab. 1, most notably, CV-OPD consistently improves performance across domains when built upon the Qwen-3.5 backbone, demonstrating strong competitiveness against both open- and closed-source foundation models. Compared with Vision-OPD, CV-OPD improves the average performance across metrics by 2.3%. On the Zoom dataset, it consistently outperforms existing methods by approximately 2%. These results demonstrate that CV-OPD can attend to relevant visual regions more accurately and remains competitive with widely used proprietary models. Notably, on HR-Bench 4K, all distillation methods underperform the base model, which we attribute to advantages in the original pretraining for 4K inputs. **Regarding training cost**, the primary additional computational overhead comes from the negative teacher. However, enabled by KV caching and the design for sharing teacher, CV-OPD increases training time by only approximately 10% relative to the original Vision-OPD.

4.3 Ablation Study

We have conducted controlled ablations to understand the contribution of each component in CV-OPD. Specifically, we investigate three questions: (1) how to construct the pipeline, (2) how to construct the negative teacher, and (3) how to control the gate.

4.3.1 Q1: How to Construct Pipeline

We examine the contributions of the Contrastive Divergence Loss (CDL) and the Discrepancy Aware Gate (DAG). As shown in Table 2, removing CDL reduces Avg_6 from 78.22% to 76.81%. The contrast between positive and negative teachers helps the student distinguish answer-relevant visual evidence from irrelevant cues. Without DAG, the average drops to 76.88%, suggesting that infor-

Table 2: Ablation of CDL and DAG in the CV-OPD pipeline.

Method		Evaluation						
CDL	DAG	V*	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg_6
✗	✗	89.53	59.41	81.75	80.12	74.51	70.17	75.92
✗	✔	90.05	60.71	82.38	82.63	74.67	70.41	76.81
✔	✗	90.58	60.71	83.00	80.88	75.40	70.71	76.88
✔	✔	91.62	62.01	84.38	83.38	76.99	70.93	**78.22**

Table 3: Ablation of negative region selection strategies

Model	Param.	V*	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg_6
Base	4B	82.20	48.40	86.38	82.00	63.94	63.80	71.12
All Black	4B	91.62	59.76	81.88	81.62	74.94	70.73	76.76
Random Crop	4B	89.53	62.01	84.38	81.50	75.06	70.73	77.20
Neighbor Crop	4B	90.05	59.76	81.62	81.00	75.40	70.54	76.40
CV-OPD	4B	91.62	62.01	84.38	83.38	76.99	70.93	**78.22**

Table 4: Ablation of DAG threshold settings.

DAG	Threshold	V*	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg_6
✗	–	90.58	60.71	83.00	80.88	75.40	70.71	76.88
✔	rollout	90.05	61.54	83.25	82.13	75.05	70.47	77.08
✔	batch	91.62	62.01	84.38	83.38	76.99	70.93	**78.22**

mative token selection provides more effective supervision than treating all positions equally. The variant without either component achieves only 75.92%, while their combination reaches 78.22%. These results highlight the importance of precise visual evidence supervision and its alignment with informative tokens in the distillation pipeline.

4.3.2 Q2: How to Select the Negative Teacher

Besides CV-OPD, we explore three alternative strategies for constructing the negative teacher: an all-black image, a crop neighboring the target region, and a randomly sampled crop. In Random Crop, each rollout independently samples a negative region, while CV-OPD shares the same negative crop across all eight rollouts of each sample.

As shown in Table 3, Neighbor Crop achieves an Avg_6 of 76.40%, below the variant without CDL (76.81%). Regions near the target may still contain answer-relevant visual information, which can introduce conflicting supervision when the student is pushed away from the negative teacher. All Black achieves 76.76%, while Random Crop improves the average to 77.20%. CV-OPD further raises the performance to 78.22%. The performance gap between Random Crop and CV-OPD may stem from their different negative sampling strategies. By sharing the same negative crop across eight rollouts, CV-OPD reduces variation in contrastive gradients and provides more consistent supervision for optimization.

4.3.3 Q3: How to Control the Gate

We compare two thresholding strategies for DAG, using either the mean discrepancy of each rollout or the mean discrepancy across the batch. As shown in Table 4, the rollout-level threshold achieves an Avg_6 of 77.08%, only slightly above the variant without DAG (76.88%). For short responses, the average discrepancy can be sensitive to a few token positions, making it difficult to reliably identify tokens that require fine-grained contrastive supervision. Using the batch-level mean instead raises Avg_6 to 78.22%. The shared threshold provides a more consistent reference across rollouts,

allowing DAG to better capture informative tokens and direct contrastive supervision toward them. So a robust threshold is essential for controlling the gate, as it remains resilient to extreme cases and ensures stable, accurate token selection.

5 Related Work

5.1 On-Policy Distillation in Large Language Models

On-policy distillation (OPD) trains a student model to match the teacher's distribution on its own generated trajectories. GKD (Agarwal et al., 2024) demonstrates that this approach can reduce the distribution mismatch between training sequences and student outputs at inference time, thereby mitigating exposure bias. Prior work has explored student-generated training contexts, alternative divergence objectives, and efficient sample reuse to improve autoregressive distillation (Lin et al., 2020; Gu et al., 2024; Ko et al., 2024). On-policy self-distillation (OPSD) further removes the need for an external teacher by deriving supervision from the student itself. For example, SDFT (Shenfeld et al., 2026) conditions a self-teacher on expert demonstrations to provide stronger on-policy supervision. However, direct access to reference solutions can introduce answer leakage (Wei et al., 2026b), while the provided context can encourage shortcut-like rationalization and harm subsequent reasoning distillation (Lee et al., 2026). These findings highlight the importance of how privileged information is used in distillation.

5.2 On-Policy Distillation in Vision-Language Models

Recent studies have extended OPD to vision-language models by providing teachers with privileged visual information (Zhang et al., 2026; Liu et al., 2026; Yoon et al., 2026). VA-OPD (Liu et al., 2026) emphasizes visually informative tokens through visual advantage, enabling more targeted distillation of fine-grained visual knowledge. From a different perspective, Decomposed OPD (Yoon et al., 2026) disentangles language-prior and visual-grounding gradients to strengthen the model's reliance on visual evidence. Other approaches explore contrastive visual evidence to improve the quality of distillation supervision. V-Zero (Sun et al., 2026) uses paired positive and negative visual views to assess the evidence supporting student-generated trajectories and gate token-level distillation. VAD (Zhang et al., 2026) further investigates the visual attribution of teacher corrections by comparing evidence-present and evidence-removed views, reconstructing supervision targets from visually attributable distributional changes. These studies highlight the importance of distinguishing visual evidence from language priors during distillation. In particular, Vision-OPD (Yuan et al., 2026) uses an evidence-centered crop-conditioned teacher to supervise an original-image-conditioned student through token-level distillation. This regional-to-global formulation transfers the benefits of focused visual inputs to original-image inference. However,this approach also raises the question of whether students have learned to identify visual evidence independently. Building on this setting, we investigate how crop-guided supervision affects students' reliance on visual evidence and introduce contrastive teacher supervision to mitigate shortcut learning.

6 Conclusion

In this work, we introduced **C**ontrastive **V**ision **O**n-**p**olicy **D**istillation (CV-OPD) to mitigate optimization shortcuts induced by the direct exposure of answer-relevant information during distillation. CV-OPD incorporates a randomly sampled visual negative teacher and a margin-based **Contrastive Divergence Loss** (CDL) to encourage the model to distinguish relevant visual evidence from misleading cues and to learn evidence-grounded reasoning trajectories. In addition, CV-OPD further employs a Discrepancy Aware Gate (DAG) that combines token-level distributional discrepancies with sample-level reliability filtering to improve the precision of distillation supervision. These designs enable CV-OPD to achieve substantial improvements in both cross-domain generalization and hallucination mitigation, demonstrating the effectiveness of contrastive supervision for promoting visually grounded reasoning in VLMs.

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*, 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.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, Gao Huang, Dahua Lin, and Lewei Lu. SenseNova-MARS: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Google DeepMind. Gemini 3 Flash model card. Google DeepMind, December 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Google DeepMind. Gemini 3.1 Pro model card. Google DeepMind, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro.

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

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024.

Jungseob Lee, Seungyoon Lee, Suhyune Son, Dongyub Jude Lee, Sungbin Han, Sugyeong Eo, and Heuiseok Lim. Answer-conditioned chains of thought degrade verifiable-reasoning distillation in large language models. *arXiv preprint arXiv:2607.14552*, 2026.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133. Association for Computational Linguistics, 2020.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

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

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

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025b.

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

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

Qwen Team. Qwen3.5-4B. Hugging Face, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-4B.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. Region-level policy optimization for fine-grained mllm perception. *arXiv preprint arXiv:2609.19745*, 2026.

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.

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. doi: 10.1609/aaai.v39i8.32852.

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

Zhenlin Wei, Pu Jian, Yingzhuo Deng, Xiaohan Wang, Jiajun Chai, Zhexin Hu, Wei Lin, Shanbin Zhang, and Guojun Yin. AMR-SD: Asymmetric meta-reflective self-distillation for token-level credit assignment. *arXiv preprint arXiv:2605.18529*, 2026b.

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 (CVPR)*, pp. 13084–13094, 2024. doi: 10.1109/CVPR52733.2024.01243.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

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.

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

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.

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

A APPENDIX

AI USE STATEMENT

LLM was used solely for grammar correction, language polishing, and LaTeX formatting. The research ideas, methodology, experiments, analysis, and technical claims were developed and verified by the authors. All AI-assisted changes were reviewed and verified by the authors, who take full responsibility for the final content.

A.1 DETAILED DERIVATION OF CDL

Start from the Eq. 10, the gradient of k-th token is:

$$\begin{aligned}\frac{\partial \mathcal{L}_{CDL}}{\partial \pi_t(k)} &= \frac{1}{2}\log\frac{2\pi_t(k)}{\pi_t(k)+\tau_t^+(k)} - \frac{1}{2}\log\frac{2\pi_t(k)}{\pi_t(k)+\tau_t^-(k)} \\ &= \frac{1}{2}\log\frac{\pi_t(k)+\tau_t^-(k)}{\pi_t(k)+\tau_t^+(k)},\end{aligned} \tag{15}$$

which is the former format of the Eq. 11. It reveals that the optimization direction of the token t relies on the difference between positive distribution and negative distribution. In the local geometric case, the update direction degenerates to $\tau_t^+ - \tau_t^-$.

We define the π_t, τ_t^+, and τ_t^- belong to the neighbor domain of the latent distribution r_t like:

$$\pi_t = r_t + \delta_t^\pi, \qquad \tau_t^+ = r_t + \delta_t^+, \qquad \tau_t^- = r_t + \delta_t^-. \tag{16}$$

When the jitter norm of the perturbation is small, the local second-order expansion of JSD is:

$$D_{\mathrm{JS}}(\pi_t, \tau_t) = \frac{1}{8}(\pi_t - \tau_t)^\top \operatorname{Diag}(r_t)^{-1}(\pi_t - \tau_t) + \mathcal{O}(\|\pi_t - \tau_t\|_2^3). \tag{17}$$

Therefore, the gradient in Eq. 11 can be rewrite as:

$$\nabla_{z_t}\mathcal{L}_{CDL} = \frac{1}{4}(\tau_t^- - \tau_t^+) + \mathcal{O}(\|\delta_t\|_2^2), \tag{18}$$

which is proportional to $\tau_t^+ - \tau_t^-$.

Original JSD uses the teacher's absolute distribution as supervision. As for CDL, under a first-order approximation, CDL uses the difference in teacher distributions resulting from matched crop pairs as supervision.

A.2 TRAINING HYPERPARAMETERS

Our training settings are summarized in Table 5. CV-OPD is trained on eight A100 GPUs using Qwen3.5-4B, a batch size of 96, and eight on-policy rollouts per sample. The maximum prompt and response lengths are 8192 and 1024 tokens, respectively. Training runs for one epoch under AdamW optimization at a learning rate of 2×10^{-6}. The primary distillation objective is generalized JSD ($\beta = 0.5$), computed over Top-100+tail probability distributions. An EMA teacher tracks the student parameters at an update rate of 0.05. For contrastive distillation, both the margin and loss weight are set to 0.1. Negative crops are shared across rollouts of the same sample, subject to a maximum IoU of 0.1.

A.3 TRAINING ALGORITHM

The complete training procedure of CV-OPD is summarized in Algorithm 1. At each training step, the student generates multiple on-policy rollouts from the full image, while the EMA teacher provides positive and negative crop-conditioned distributions along the same prefix. A shared negative crop is used for all rollouts of each sample, and samples with no correct rollouts are excluded from distillation. The remaining trajectories are optimized using the main JSD distillation objective and the gated Contrastive Divergence Loss (CDL), followed by an EMA update of the teacher parameters.

A.4 ATTENTION VISUALIZATION

To examine whether CV-OPD can more accurately localize answer-relevant visual regions, we compare its attention distributions with those of Vision-OPD on fine-grained visual reasoning examples. As shown in Fig. 4, CV-OPD exhibits more concentrated attention on answer-relevant regions, while Vision-OPD shows relatively dispersed attention across the image. Based on the results, CV-OPD directs more attention to the visual details needed to answer the question. This tighter focus improves fine-grained visual understanding and reduces reliance on irrelevant cues and shortcut-based predictions.

Table 5: Training hyperparameters for CV-OPD.

Setting	**Value**	**Role**
GPU allocation	8 A100 GPUs	Training worker allocation
Backbone	Qwen3.5-4B	Student and EMA teacher model
Training samples	6,241	Number of visual reasoning examples
Training batch size	96	Number of samples per training batch
Actor update mini-batch size	96	Mini-batch size for actor updates
Rollouts per prompt	8	On-policy responses generated per sample
Maximum prompt length	8192 tokens	Prompt truncation limit
Maximum response length	1024 tokens	On-policy generation limit
Maximum model length	9216 tokens	Maximum sequence length
Training epochs	1	Number of passes over the training set
Learning rate	2×10^{-6}	Actor optimization learning rate
Optimizer	AdamW	Student parameter optimization
AdamW betas	$(0.9, 0.999)$	Optimizer momentum coefficients
Weight decay	0.01	Optimizer regularization
Gradient clipping	1.0	Maximum gradient norm
Learning rate scheduler	Constant	Learning rate schedule
Warmup steps	10	Learning rate warmup
Distillation objective	Generalized JSD	Main token-level distillation loss
JSD coefficient	0.5	Main distillation interpolation coefficient
Top-K size	100	Student-selected vocabulary support
Tail probability bucket	Enabled	Aggregation of remaining vocabulary mass
Distillation temperature	1.0	Temperature for the main distillation loss
Teacher parameter update	EMA of student weights	Teacher model update strategy
Teacher EMA update rate	0.05	Student-weight coefficient in EMA updates
Positive teacher image field	`bbox_images`	Source of answer-relevant crops
Negative crop sharing	Sample-wise	One negative crop shared across eight rollouts
Maximum negative crop IoU	0.1	Overlap constraint with the positive region
Contrastive loss	Full-vocabulary JSD	Positive-negative teacher comparison
Contrastive margin	0.1	Margin in the contrastive objective
Contrastive loss weight	0.1	Weight of the auxiliary contrastive loss
Token gate threshold	Batch-mean JSD	Selection of tokens for contrastive supervision
Sample-level filter	All-wrong filtering	Exclusion of samples with eight incorrect rollouts
Rollout correction mode	Token-level	Importance sampling correction
Rollout correction threshold	2.0	Importance ratio clipping threshold
GRPO loss weight	0.0	No additional GRPO policy loss
Checkpoint frequency	10 steps	Interval between saved actor checkpoints
Logging mode	Console, TensorBoard, W&B	Training metric recording

Algorithm 1 CV-OPD Training

Require: Training set $\mathcal{D}$, student policy π_θ, rollout count M, Top-K size K, JSD coefficient β, negative crop IoU threshold δ, contrastive margin μ, contrastive weight λ, EMA update rate τ
Ensure: Updated student parameters θ

Initialize teacher parameters $\phi \leftarrow \theta$
for each training step **do**
  Sample a mini-batch $\mathcal{B} \subset \mathcal{D}$
  Initialize retained trajectories $\mathcal{T} \leftarrow \emptyset$
  for each sample $x = (I, q, R^+, a) \in \mathcal{B}$ **do**
    Generate M on-policy rollouts $Y^{(1)}, \ldots, Y^{(M)} \sim \pi_\theta(\cdot \mid I, q)$
    Compute correctness rewards $r_i \leftarrow \mathrm{Reward}(Y^{(i)}, a)$
    if $\sum_{i=1}^{M} r_i = 0$ **then**
      continue
    end if
    Sample one negative crop R^- from I with $\mathrm{IoU}(R^+, R^-) \le \delta$
    for $i = 1, \ldots, M$ **do**
      Compute student distributions $\pi_{i,t}$ using the full image I
      Compute positive teacher distributions $\tau^+_{i,t}$ using R^+
      Compute negative teacher distributions $\tau^-_{i,t}$ using the shared R^-
      Compute full-vocabulary discrepancies $d_{i,t} \leftarrow D_{\mathrm{JS}}(\pi_{i,t}, \tau^+_{i,t})$
      Add valid response-token positions to $\mathcal{T}$
    end for
  end for
  if $\mathcal{T} = \emptyset$ **then**
    continue
  end if
  Compute batch-mean discrepancy $\bar{d} \leftarrow \frac{1}{|\mathcal{T}|} \sum_{(i,t)\in\mathcal{T}} d_{i,t}$
  Initialize losses $\mathcal{L}_{\mathrm{OPD}} \leftarrow 0$, $\mathcal{L}_{\mathrm{CDL}} \leftarrow 0$
  Initialize normalization factors $N \leftarrow 0$, $G \leftarrow 0$
  for each valid token position $(i, t) \in \mathcal{T}$ **do**
    Construct student-aligned Top-K+tail distributions $\tilde{\pi}_{i,t}$ and $\tilde{\tau}^+_{i,t}$
    $\mathcal{L}_{\mathrm{OPD}} \leftarrow \mathcal{L}_{\mathrm{OPD}} + w_{i,t} D_{\mathrm{JS},\beta}(\tilde{\pi}_{i,t}, \tilde{\tau}^+_{i,t})$
    $N \leftarrow N + w_{i,t}$
    if $d_{i,t} > \bar{d}$ **then**
      $d^+_{i,t} \leftarrow D_{\mathrm{JS}}(\pi_{i,t}, \tau^+_{i,t})$
      $d^-_{i,t} \leftarrow D_{\mathrm{JS}}(\pi_{i,t}, \tau^-_{i,t})$
      $\mathcal{L}_{\mathrm{CDL}} \leftarrow \mathcal{L}_{\mathrm{CDL}} + [\mu + d^+_{i,t} - d^-_{i,t}]_+$
      $G \leftarrow G + 1$
    end if
  end for
  $\mathcal{L} \leftarrow \dfrac{\mathcal{L}_{\mathrm{OPD}}}{N + \epsilon} + \lambda \dfrac{\mathcal{L}_{\mathrm{CDL}}}{G + \epsilon}$
  Update student parameters θ by minimizing $\mathcal{L}$
  Update teacher parameters $\phi \leftarrow (1 - \tau)\phi + \tau\theta$
end for

A.5 Attention Visualization in Black Mask

We analyzed the shortcut behavior of Vision-OPD under the black-mask setting, where the target region is replaced by black pixels. As shown in Fig. 5, Vision-OPD still gives the correct answer even though the relevant visual evidence is occluded. This suggests that the model can rely on hallucinated or shortcut-based reasoning rather than truly grounding its prediction in the target region. The attention visualization further supports this observation: Vision-OPD does not accurately localize the correct target position, yet it still predicts the expected answer. In contrast, CV-OPD gives a

more objective response by explicitly describing the visible scene and the ambiguity in the image, which indicates stronger reliance on observable visual evidence instead of unsupported inference.

A.6 Responses Under Insufficient Visual Evidence

We additionally present qualitative cases where answer-relevant visual evidence is occluded or corrupted. Despite the loss of visual information, CV-OPD explicitly acknowledges the missing evidence and describes the observable scene without inventing unsupported visual details.

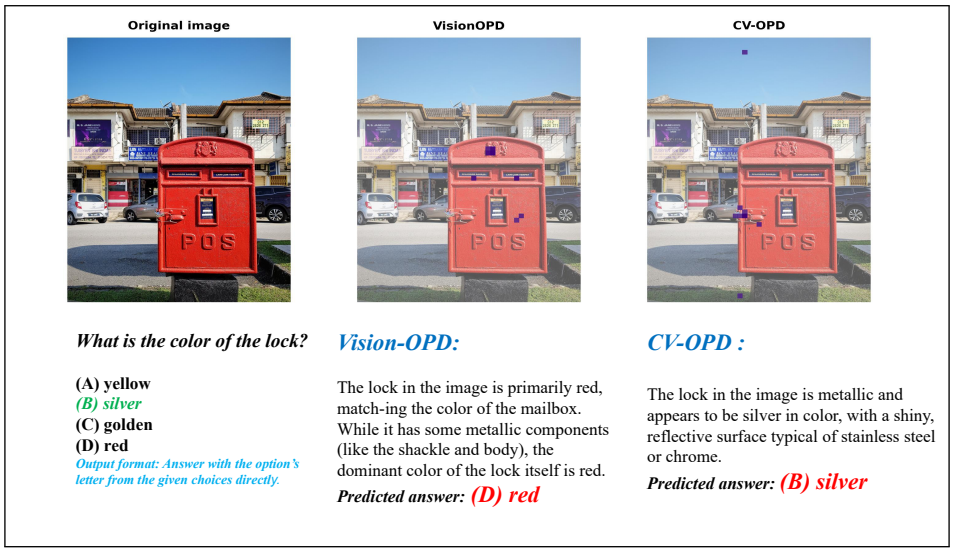


Figure 4: Qualitative comparison of visual attention and model responses between Vision-OPD and CV-OPD. Vision-OPD relies on the dominant color of the mailbox and predicts red, whereas CV-OPD focuses on the lock itself and identifies its metallic silver color.

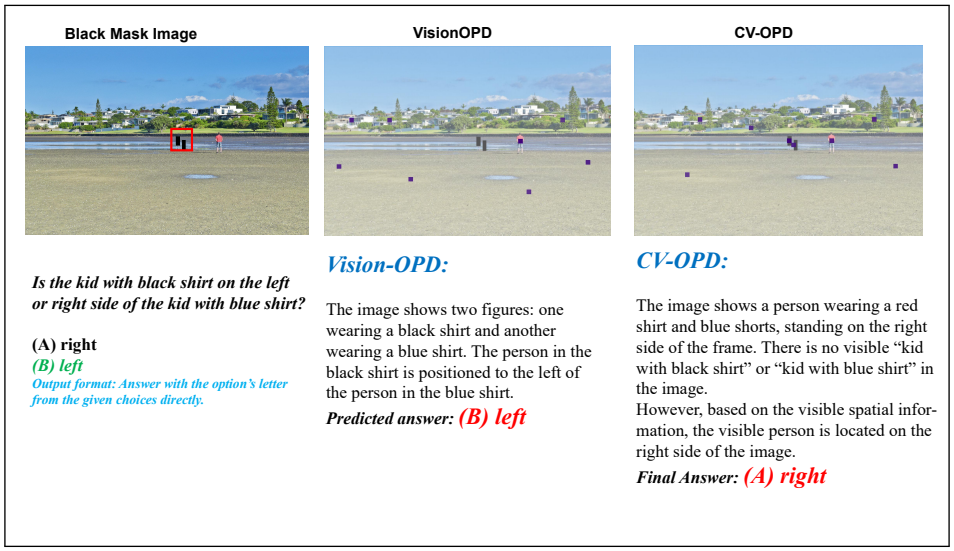


Figure 5: Attention visualization and model responses under the black-mask setting. After the answer-relevant region is replaced by black pixels, Vision-OPD still produces the expected answer despite failing to localize the relevant visual evidence, indicating shortcut-based prediction and hallucination. In contrast, CV-OPD explicitly describes the observable scene and its ambiguity, showing stronger reliance on available visual evidence.

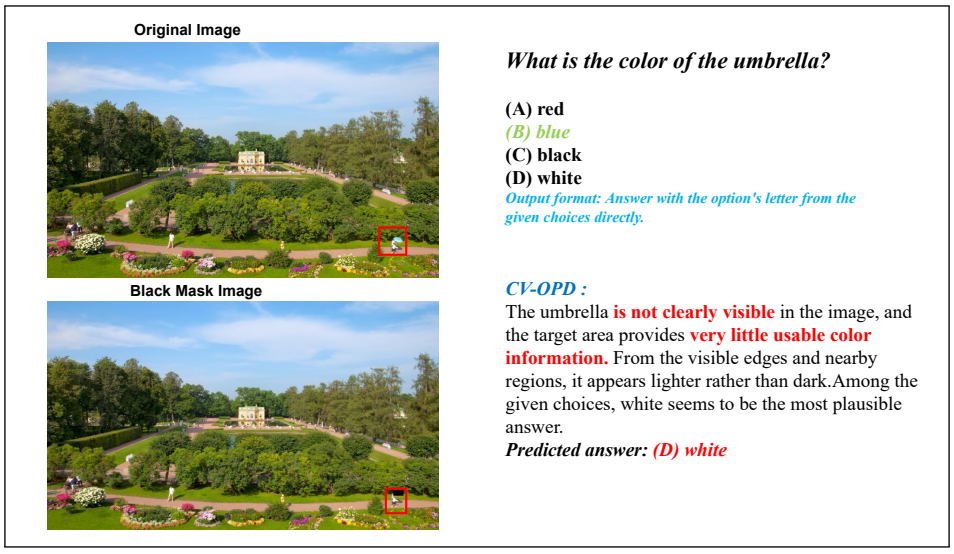


Figure 6: CV-OPD response under the black-mask setting. When the umbrella is not clearly visible, CV-OPD explicitly acknowledges the lack of reliable color information and limits its description to the remaining visual evidence, although its forced-choice prediction is incorrect.

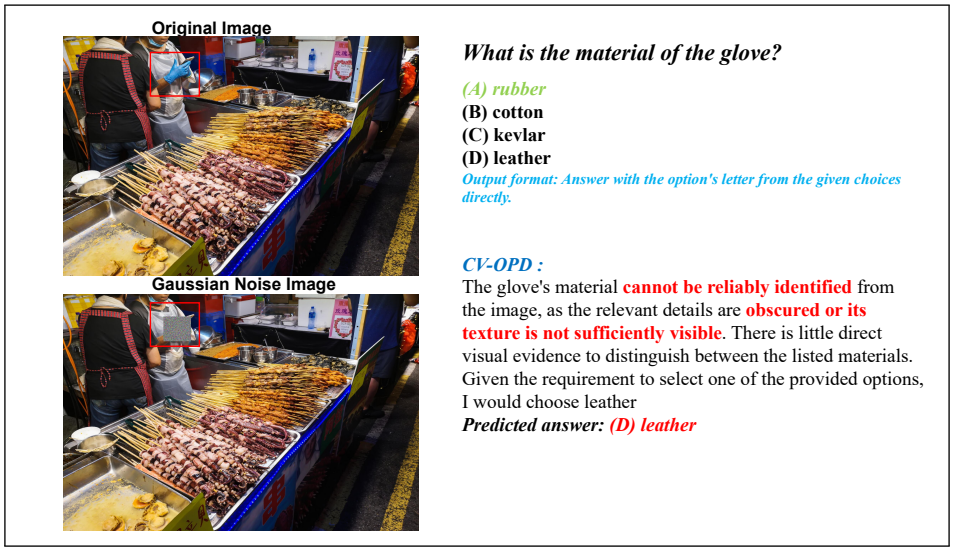


Figure 7: CV-OPD response under Gaussian noise. With the glove's material obscured by visual corruption, CV-OPD recognizes that the available evidence is insufficient for reliable identification, despite producing an incorrect final choice.