arXiv:2608.05131v2 [cs.CV] 6 Aug 2026

OPD-V: Visual On-Policy Self-Distillation with Modality Balance

Aniri[2,3], **Jinhe Bi**[1,2,3], **Peng Liao**[4], **Zengjie Jin**[2], **Volker Tresp**[2,3], **Fei Shen**[1], **Yunpu Ma**[2,3], **Tat-Seng Chua**[1]

[1]National University of Singapore, [2]Ludwig Maximilian University of Munich, [3]Munich Center for Machine Learning, [4]Sun Yat-sen University

GitHub Qwen3-VL-8B Qwen3-VL-4B Qwen3.5-9B Qwen3.5-4B

On-Policy Self-Distillation (OPSD) has become a standard post-training approach for improving visual reasoning in multimodal large language models (MLLMs). Existing methods draw privileged information from diverse input sources to guide self-distillation. Yet these designs overlook Modality Imbalance, a challenge inherent to MLLM reasoning. When textual information dominates generation, the model cannot fully integrate its multimodal input. Consequently, carefully designed privileged information remains underused, limiting the effectiveness of OPSD. To examine this limitation, we construct a Positive Teacher with the Zoom-In Image and a Negative Teacher with the Mask Image, which exhibit different degrees of Modality Imbalance. Changes in their reasoning correctness and token logits reveal that Modality Balance can itself serve as privileged information. Motivated by this finding, we introduce OPD-V, a visual OPSD paradigm that instantiates such information through the Positive Teacher and Negative Teacher. Positive Modality-Balance Logits Margins define a Modality-Balance Trust Region that selects the on-policy tokens used for self-distillation. Experiments across 6 benchmarks, 4 MLLM backbones, and 5 post-training methods show that OPD-V consistently improves reasoning performance while reducing training cost.

Date: August 7, 2026
Correspondence: bijinhe@outlook.com, cognitive.yunpu@gmail.com

1 Introduction

Multimodal large language models (MLLMs) support a broad range of visual understanding and reasoning tasks [1, 2, 3, 4]. On-Policy Self-Distillation (OPSD) has recently become a standard post-training approach for further improving their visual reasoning. OPSD samples trajectories from the current student and uses a copy of the same model to provide dense targets at student-visited prefixes [5, 6, 7]. This model copy receives privileged information available only during training, removing the need for a separately trained external teacher.

The quality of these targets depends on how the privileged information is constructed. A common strategy provides verified textual solutions or reference answers [7, 8, 9]. Recent visual OPSD methods instead derive privileged information from additional or transformed visual inputs. Vision-OPD conditions the teacher on an evidence-centered crop, whereas Visual-OPSD uses privileged visual thoughts [10, 11]. VA-OPD and VCSD create teacher contrasts by changing access to visual content [12, 13]. Although these methods construct the privileged signal in different ways, they do not account for Modality Imbalance, a challenge inherent to MLLM reasoning.

MLLMs generate each response from visual inputs, textual instructions, and autoregressive context. This joint conditioning creates a persistent asymmetry because strong textual priors can dominate generation even when the task depends on the image [14, 15, 16, 17]. We define this tendency as *Modality Imbalance*: the model relies disproportionately on textual context instead of integrating the multimodal input. Carefully designed privileged information can enrich the teacher input without revealing how the visual and textual modalities contribute to each prediction. Consequently, the model can underuse this information, limiting the effectiveness of OPSD. This limitation raises the central question of our work:

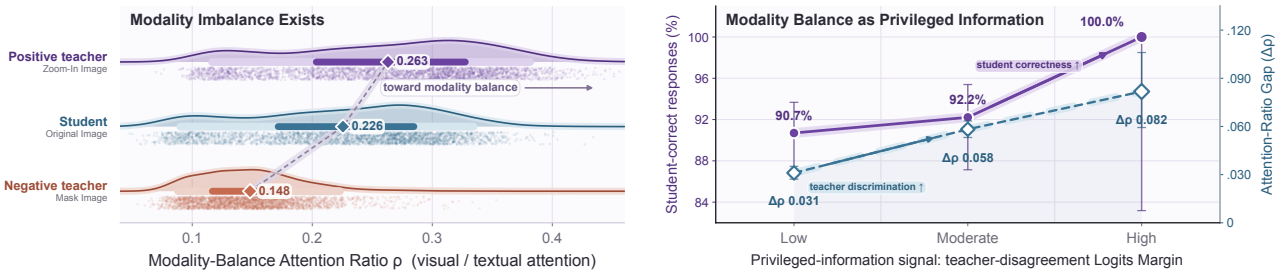


Figure 1: Modality Balance as privileged information, evaluated on 5K samples drawn from multiple domains. Left: the Modality-Balance Attention Ratio is lowest for the Negative Teacher with the Mask Image, intermediate for the student with the Original Image, and highest for the Positive Teacher with the Zoom-In Image. These matched conditions expose distinct degrees of Modality Imbalance. Right: as the Modality-Balance Logits Margin increases, the gap between the Positive and Negative Teacher attention ratios widens, accompanied by higher student correctness. For each margin interval b, the blue curve reports $\Delta\rho_b = |b|^{-1} \sum_{i \in b} [\rho(c^+; y_i) - \rho(c^-; y_i)]$, where y_i is the response trajectory for sample i, and c^+ and c^- denote the Zoom-In Image and Mask Image conditions, respectively.

How can privileged information be designed to remain effective for MLLM OPSD under Modality Imbalance?

The Present Work: We address this question by treating *Modality Balance itself* as privileged information for OPSD and introducing **OPD-V**. Unlike conventional privileged information like textual solutions or visual inputs, Modality Balance is not an explicit input that can be directly appended to the teacher context. It instead describes the model's internal allocation of visual and textual information during generation. The central challenge is therefore to express this internal condition through explicit teacher inputs and convert it into token-level supervision for OPSD.

Figure 1 demonstrates this conversion through matched student and teacher image conditions. We construct a Positive Teacher with the Zoom-In Image and a Negative Teacher with the Mask Image, while the student receives the Original Image. The Modality-Balance Attention Ratio measures visual attention relative to textual attention under each condition. As shown in the left panel, the ratio is lowest for the Negative Teacher, intermediate for the student, and highest for the Positive Teacher. This ordering shows that the matched image conditions expose distinct degrees of Modality Imbalance through explicit differences in the model input.

The right panel further examines how these differences affect teacher scoring. The Modality-Balance Logits Margin measures how differently the Positive Teacher and Negative Teacher score the student response. From low to high margin intervals, the gap between their Modality-Balance Attention Ratios widens, while student correctness increases. Teachers with different levels of Modality Balance thus provide a directional scoring distinction that can serve as privileged information for OPSD.

Building on this finding, we introduce **OPD-V**, a visual OPSD paradigm that uses Modality Balance as privileged information. Figure 2 summarizes how OPD-V implements this paradigm during training. The student receives the Original Image, the Positive Teacher receives the Zoom-In Image, and the Negative Teacher receives the Mask Image. All three distributions score the same student-generated tokens under matched textual context and on-policy prefixes. The Modality-Balance Logits Margin then compares the scores assigned by the Positive Teacher and Negative Teacher. Positive Modality-Balance Logits Margins define the Modality-Balance Trust Region used for self-distillation. Within this region, OPD-V applies Jensen–Shannon distillation from the Positive Teacher distribution to the student.

We evaluate OPD-V across 6 benchmarks, 4 MLLM backbones, and 5 post-training methods. These benchmarks span visually demanding settings and broader multimodal reasoning tasks across image scale, resolution, language, and real-world context. OPD-V consistently improves reasoning performance across the evaluated

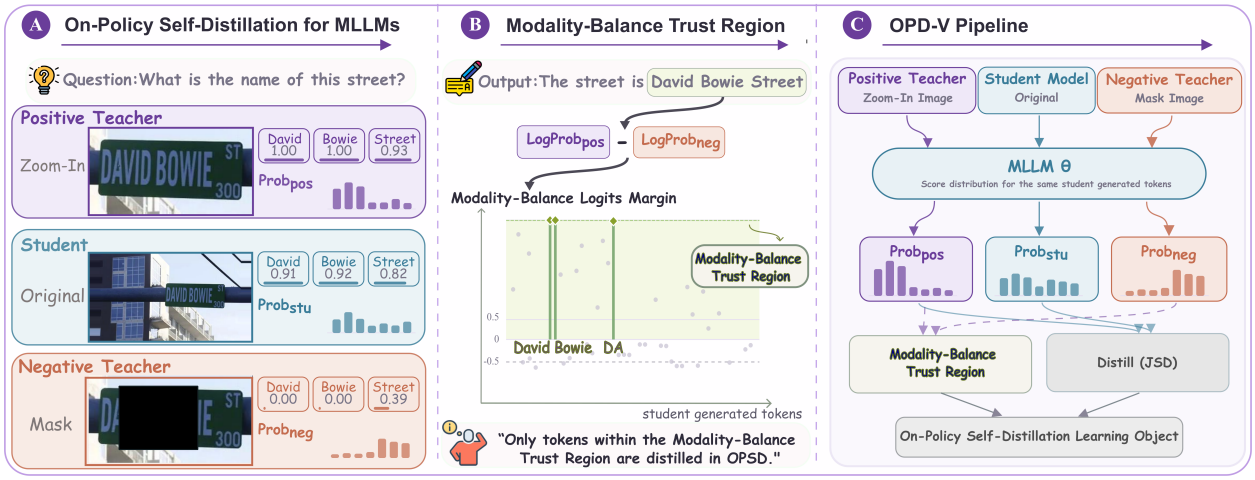


Figure 2: Overview of OPD-V. (A) For the same student-generated response, the Positive Teacher, student, and Negative Teacher score each on-policy token under the Zoom-In Image, Original Image, and Mask Image, respectively. (B) The tokenwise Modality-Balance Logits Margin compares the log probabilities assigned by the Positive Teacher and Negative Teacher; tokens with positive margins form the Modality-Balance Trust Region. (C) OPD-V applies Jensen–Shannon distillation from the Positive Teacher distribution to the student only within this region.

backbones and settings. Compared with standard OPSD, it also reduces step latency on both 4B and 9B backbones. These results show that OPD-V uses Modality Balance as privileged information to improve reasoning accuracy while reducing training cost.

Our contributions are summarized as follows:

- We identify Modality Imbalance as a limitation that restricts the effectiveness of privileged information in MLLM OPSD.
- We show that Modality Balance can serve as privileged information and propose OPD-V, which instantiates it through the Positive Teacher, Negative Teacher, and Modality-Balance Trust Region.
- We demonstrate consistent improvements across representative OPSD methods while reducing training cost.

2 Preliminaries

2.1 Task Formulation

We consider supervised post-training for multimodal reasoning. Each example contains an Original Image I, a textual query x, and a verified target $a^\star$ available only during training. The Original Image may include the target-region cue, and the corresponding visual focus instruction is included in the textual query. The trainable MLLM p_θ receives only the Original Image and textual query, and samples the on-policy trajectory

$$y = (y_1, \ldots, y_T) \sim p_\theta(\cdot \mid I, x). \tag{1}$$

Let $y_{<t} = (y_1, \ldots, y_{t-1})$ be the student-generated prefix before token y_t, and let $\mathcal{T}_y = \{1, \ldots, T\}$ be the corresponding token-index set. At position t, the student distribution is

$$p_{\theta,t}(\cdot) := p_\theta(\cdot \mid I, x, y_{<t}). \tag{2}$$

The verified target $a^\star$ is privileged information used only to construct training-time teacher supervision.

2.2 On-Policy Self-Distillation for MLLMs

On-policy self-distillation (OPSD) samples y from the current student and evaluates each student-generated prefix with a detached copy of the same MLLM [5, 6, 7]. Let $\bar{\theta}$ denote the detached teacher parameters. At token position t, the teacher receives the verified target $a^\star$ and scores the same student prefix:

$$q_t^{\text{priv}}(\cdot) := p_{\bar{\theta}}(\cdot \mid I, x, a^\star, y_{<t}). \tag{3}$$

Vanilla OPSD distills this privileged teacher distribution into the student along its on-policy trajectory:

$$\begin{aligned}\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{y\sim p_\theta(\cdot|I,x)}\Bigg[\frac{1}{|\mathcal{T}_y|} \\ \cdot \sum_{t\in\mathcal{T}_y} D_{\text{dist}}\Big(q_t^{\text{priv}}, p_{\theta,t}\Big)\Bigg].\end{aligned} \tag{4}$$

Here D_{dist} denotes the token-level distillation divergence, instantiated as D_{JS} in our implementation. Since each $y_{<t}$ is generated by the current student, supervision remains on-policy and gradients propagate only through $p_{\theta,t}$.

2.3 Modality Imbalance in MLLMs

During OPSD, each token prediction is conditioned on the visual input, textual query, and autoregressive prefix. Modality Imbalance denotes the tendency of an MLLM to depend disproportionately on textual context relative to visual input. Consequently, a prediction can be mainly supported by language priors even when the query requires visual information [15, 16, 17]. To measure this tendency over a complete response, let $\alpha_t^V(c)$ and $\alpha_t^T(c)$ denote the visual and textual attention masses at position t under model condition c. We define the Modality-Balance Attention Ratio as

$$\rho(c; y) = \frac{\sum_{t\in\mathcal{T}_y} \alpha_t^V(c)}{\sum_{t\in\mathcal{T}_y} \alpha_t^T(c) + \epsilon}. \tag{5}$$

where ϵ is a small constant for numerical stability. A larger $\rho(c; y)$ indicates stronger visual attention relative to textual attention over the on-policy trajectory. This ratio measures Modality Imbalance without requiring equal numerical contributions from the two modalities.

3 OPD-V

3.1 Overview

After measuring Modality Imbalance with the Modality-Balance Attention Ratio, OPD-V turns Modality Balance into OPSD supervision. For an on-policy trajectory $y \sim p_\theta(\cdot \mid I, x)$, the student first generates tokens from the Original Image. The Positive Teacher with the Zoom-In Image and the Negative Teacher with the Mask Image then score the same tokens under the matched textual query and prefix. Comparing their normalized logits gives the Modality-Balance Logits Margin. When this margin is positive, the token receives stronger support after the task-relevant region is magnified than after selected visual regions are masked, and is therefore placed in the Modality-Balance Trust Region. The objective distills the Positive Teacher distribution only within this selected region.

3.2 Positive Teacher with the Zoom-In Image and Negative Teacher with the Mask Image

The two teacher views are constructed sequentially. I^{zoom} is obtained by cropping and magnifying the task-relevant region. The Mask Image is then generated as $I^{\text{mask}} = \text{Mask}(I^{\text{zoom}})$ by replacing a random

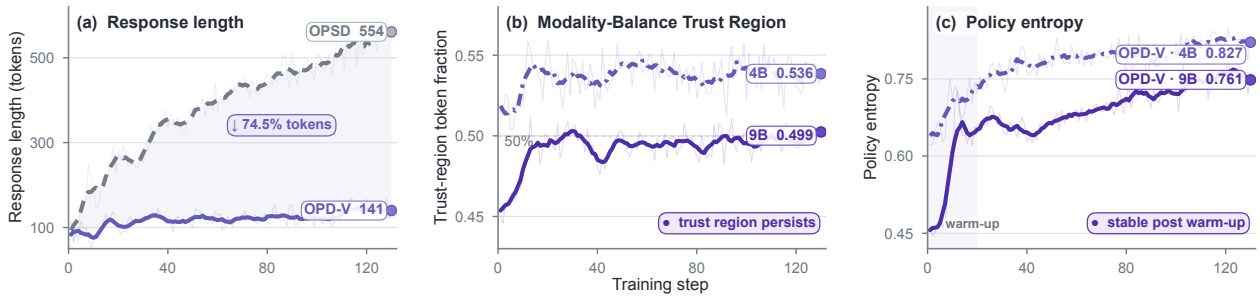


Figure 3: Training dynamics on Qwen3.5 backbones. Thin traces show logged values, and bold curves show seven-step rolling means. (a) Response length on the 4B backbone, where OPD-V maintains concise responses while OPSD exhibits late-stage length growth. (b) Fraction of on-policy tokens within the Modality-Balance Trust Region for OPD-V at 4B and 9B scales. (c) Policy entropy for OPD-V at 4B and 9B scales, both of which remain stable after warm-up.

rectangular region in the Zoom-In Image with black pixels. Both inputs are passed to the detached teacher with the same textual query and student-generated prefix:

$$\begin{aligned} q_t^+(\cdot) &= p_{\bar{\theta}}(\cdot \mid I^{\text{zoom}}, x, y_{<t}), \\ q_t^-(\cdot) &= p_{\bar{\theta}}(\cdot \mid I^{\text{mask}}, x, y_{<t}). \end{aligned} \tag{6}$$

The Positive Teacher supplies the distillation target, while the Negative Teacher supplies the paired comparison for the same student-generated token y_t.

3.3 Modality-Balance Trust Region

Since the two teacher distributions are evaluated at the same prefix, token selection can be based on their normalized logits. Let $\ell_t^+(v) = \log q_t^+(v)$ and $\ell_t^-(v) = \log q_t^-(v)$. For each on-policy token y_t, the Modality-Balance Logits Margin is

$$\delta_t^{\text{MB}} = \ell_t^+(y_t) - \ell_t^-(y_t). \tag{7}$$

A positive Logits Margin indicates that the same token receives more support from the Positive Teacher than from the Negative Teacher. These positions form the Modality-Balance Trust Region:

$$\mathcal{R}_{\text{MB}}(y) = \left\{ t \in \mathcal{T}_y \;\middle|\; \delta_t^{\text{MB}} > 0 \right\}. \tag{8}$$

Thus, $\mathcal{R}_{\text{MB}}(y)$ keeps the on-policy tokens whose support increases under the Zoom-In Image relative to the Mask Image.

3.4 OPD-V Objective

Within the Modality-Balance Trust Region, the Positive Teacher distribution serves as the Jensen–Shannon distillation target for the student distribution in Eq. (2). Let $r_t \in \{0, 1\}$ be the valid-token indicator, where $r_t = 1$ means that position t is a generated response token included in distillation. This gives the OPD-V objective:

$$\begin{aligned} \mathcal{L}_{\text{OPD-V}}(\theta) = \mathbb{E}\Bigg[& \frac{1}{\sum_{t \in \mathcal{T}_y} r_t} \\ & \cdot \sum_{t \in \mathcal{R}_{\text{MB}}(y)} r_t \delta_t^{\text{MB}} D_{\text{JS}}\big(q_t^+, p_{\theta,t}\big) \Bigg]. \end{aligned} \tag{9}$$

where D_{JS} denotes Jensen–Shannon divergence. Because every $t \in \mathcal{R}_{\text{MB}}(y)$ has $\delta_t^{\text{MB}} > 0$, the Modality-Balance Logits Margin scales the selected on-policy tokens during self-distillation.

Model	Params	V* Bench	ZoomBench	HR-Bench		MME-RealWorld		Average
				4K	8K	EN	CN	
Closed-Source Models (Single Forward Pass)								
GPT-5.2	-	79.06	50.89	81.12	78.38	72.60	68.80	71.81
GPT-5.4	-	76.96	52.66	84.00	77.88	74.20	70.93	72.77
Gemini-3.1-Pro	-	87.96	61.18	89.63	86.88	76.53	73.31	79.25
Gemini-3.5-Flash	-	89.01	61.42	89.12	86.62	75.31	73.97	79.24
Open-Source Baselines (Single Forward Pass)								
DeepEyes	7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
Thyme	7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyesV2	7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
SenseNova-MARS	8B	92.15	47.81	83.13	78.38	67.90	68.90	73.05
MiMo-VL-RL	7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
ZwZ	8B	87.96	56.69	83.63	81.75	66.57	68.09	74.12
MiniCPM-V-4.5	9B	70.68	42.60	69.63	61.50	62.65	61.64	61.45
GLM-4.6V	106B	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Qwen3-VL-Instruct	235B	91.10	56.09	86.13	80.38	71.74	69.04	75.75
Qwen3.5	397B	87.96	57.16	89.38	85.50	74.82	69.82	77.44
Kimi-K2.6	1T	88.48	53.14	81.88	78.00	69.22	66.13	72.81
Qwen3.5 (4B) Distillation Series & Ours								
Qwen3.5	4B	80.63	49.82	73.50	67.50	55.78	58.56	64.30
+ *SFT*	4B	83.25	42.60	76.50	72.12	64.10	62.50	66.85
+ *GRPO*	4B	85.34	48.20	78.25	74.50	66.50	65.10	69.65
+ *OPSD*	4B	84.29	57.51	77.25	72.88	61.20	62.40	69.26
+ *Vision-OPD*	4B	89.01	65.33	82.25	81.25	74.50	70.26	77.10
+ *VA-OPD*	4B	86.91	51.20	80.12	77.50	68.50	67.10	71.88
+ ***OPD-V (Ours)***	4B	**94.76** (+14.1%)	**65.92** (+16.1%)	**85.75** (+12.3%)	**83.25** (+15.8%)	**76.31** (+20.5%)	**74.06** (+15.5%)	**80.01** (+15.7%)

Table 1: Comparison with state-of-the-art MLLMs and matched Qwen3.5-4B post-training baselines across six multimodal benchmarks. All entries report accuracy (%), and Average is computed over V* Bench, ZoomBench, HR-Bench at 4K and 8K resolutions, and the English and Chinese subsets of MME-RealWorld. Results for a subset of the baseline models are taken from Vision-OPD [10]. The final group compares SFT, GRPO, OPSD, Vision-OPD, VA-OPD, and OPD-V using the same backbone and training data. Parenthetical values in the OPD-V row report absolute percentage-point gains over the Qwen3.5-4B base model; bold marks the best result within this matched post-training group.

3.5 Computational Efficiency

Compared with standard OPSD, OPD-V adds one Negative Teacher forward pass to score the same on-policy tokens under the Mask Image. The Modality-Balance Logits Margin and Modality-Balance Trust Region are computed tokenwise, while both teachers share the same EMA parameters; OPD-V therefore introduces no additional model parameters or student updates. In practice, shorter responses and more efficient teacher inputs offset this additional pass, as measured in Section 4.3.

4 Experiments

In this section, we conduct extensive experiments to address the following research questions: (**RQ1: Overall Effectiveness**) How does OPD-V perform when compared to existing OPSD methods? (**RQ2: Computational Efficiency**) Does OPD-V introduce significant computational overhead? (**RQ3: Mechanistic Validity**) How sensitive is OPD-V to the proposed Dual Teacher Guidance Dynamics and the choice of Visual Transformation Strategy Variants?

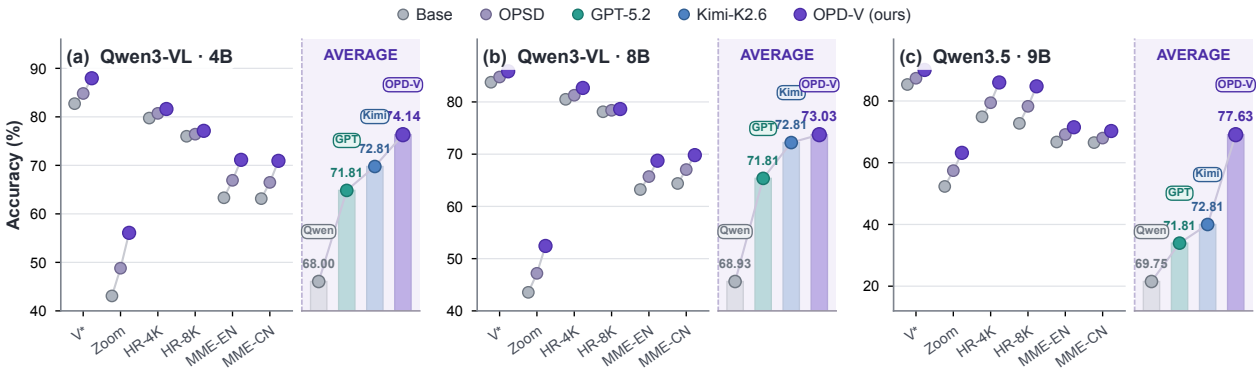


Figure 4: Absolute accuracy of OPD-V across diverse model architectures (Qwen3-VL-Instruct 4B/8B and Qwen3.5 9B). Connected points compare the Base model, OPSD, and OPD-V on each benchmark. Model panels use independently zoomed y-axes to make within-backbone differences visible without altering the reported scores. The Average cards are independently scaled and annotate exact values for the Base model, GPT-5.2, Kimi-K2.6, and OPD-V.

4.1 Experimental Settings

4.1.1 Model Training and Implementation Details

We instantiate our proposed **OPD-V** framework across multiple vision-language backbones, including Qwen3.5-4B, Qwen3.5-9B, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct [18, 19], utilizing a curated synthetic dataset of 6.2K visual reasoning samples from Vision-OPD [10]. For policy alignment, we employ Jensen-Shannon Divergence (JSD) with $\beta = 0.5$ as the divergence objective. To overcome the high memory overhead inherent to full-vocabulary logit distillation [7], we adopt a top-K distillation strategy that evaluates only the top $K = 100$ logits from the student model alongside their corresponding teacher probability distributions, supplemented by a tail-probability adjustment. In our experimental setup, tokens beyond the top 100 account for less than 1×10^{-13} of the cumulative probability mass, making this top-K approximation virtually lossless while significantly reducing memory requirements. The teacher parameters are maintained using an Exponential Moving Average (EMA) of the student weights. Throughout training, the maximum on-policy generation length is set to 1024 tokens, and models are trained for 1 epoch.

4.1.2 Evaluation Benchmarks

We evaluate OPD-V on six established benchmarks that connect visually demanding inputs to broader multimodal problem solving. Each requires models to extract task-relevant visual evidence and use it for question answering, contextual understanding, or reasoning. V* Bench [20] tests the localization and recognition of small targets in high-resolution, visually crowded scenes, whereas ZoomBench [21] examines whether models can recover subtle regional evidence from a full image across multiple fine-grained VQA dimensions. HR-Bench [22] extends image understanding to native 4K and 8K resolutions, where relevant information can be lost through conventional downsampling. MME-RealWorld [23] provides the broadest setting, spanning diverse high-resolution tasks and real-world scenarios that require both detailed observation and contextual reasoning. Together, the suite provides a general multimodal evaluation across complementary challenges in evidence localization, scale, resolution, language, and real-world context.

4.1.3 Baselines

Following Vision-OPD [10], we compare **OPD-V** with two groups of baselines:

1. **State-of-the-Art MLLMs:** In accordance with Table 1, we organize these baselines by model accessibility. (a) The closed-source group comprises the Google Gemini family, represented by Gemini-3.1-Pro [24] and Gemini-3.5-Flash [25], and the OpenAI GPT family, represented by GPT-5.4 [26] and GPT-5.2 [27]. (b) The open-source group covers a broad range of model scales. Its compact

Figure 5: Wall-clock step time for Qwen3.5 4B and 9B backbones. Pale traces show raw measurements and haloed curves show seven-step rolling means; dashed lines denote OPSD and solid lines denote OPD-V. Paired checkpoint connectors emphasize the time gap. Mean step latency drops from 352s to 240s (31.8%) for 4B and from 451s to 340s (24.7%) for 9B.

7B–9B systems include Thyme [28], DeepEyes [29], DeepEyesV2 [30], MiMo-VL-RL [31], ZwZ [21], SenseNova-MARS [32], and MiniCPM-V-4.5 [33]; the larger-scale systems comprise GLM-4.6V [34], Kimi-K2.6 [35], Qwen3-VL-Instruct [19], and Qwen3.5 [18].

2. **Alternative Optimization Strategies:** Using the same Qwen3.5-4B backbone and training data, we evaluate: (a) *SFT*, supervised fine-tuning on the reference targets; (b) *GRPO* [36], reinforcement learning with binary ground-truth outcome verification; (c) *OPSD* [7], on-policy self-distillation conditioned on the reference targets; (d) *Vision-OPD* [10], which constructs privileged information using evidence-centered crops; and (e) *VA-OPD* [12], which contrasts teacher predictions under different access to visual content.

For this matched comparison, all post-training methods use the same 6.2K training set and the ground-truth target protocol from Vision-OPD [10]. All models are trained and evaluated in non-thinking mode.

4.2 RQ1: Overall Effectiveness

This analysis compares predictive performance and training dynamics across post-training strategies, model families, and parameter scales.

Takeaway 1: OPD-V improves performance across matched baselines, architectures, and scales. Table 1 shows that OPD-V raises the Qwen3.5-4B average from 64.30% to 80.01%, an absolute gain of 15.71 percentage points. It also exceeds every matched post-training baseline: the strongest alternative, Vision-OPD, reaches 77.10%, leaving a 2.91-point gap. OPD-V improves upon Vision-OPD on all six reported evaluations, so the aggregate gain spans the benchmark suite. The same pattern extends beyond Qwen3.5-4B. As shown in Figure 4, OPD-V improves Qwen3-VL-4B from 68.00% to 74.14%, Qwen3-VL-8B from 68.93% to 73.03%, and Qwen3.5-9B from 69.75% to 77.63%. These two model families differ in how multimodal information is represented and propagated. Qwen3-VL uses Interleaved-MRoPE and DeepStack to inject multi-level ViT features into the language backbone, whereas Qwen3.5 is pretrained with early vision–text fusion and adopts a 3:1 hybrid language backbone that interleaves Gated DeltaNet linear-attention layers with full-attention layers [19, 18]. The consistent improvements therefore span different mechanisms for visual feature integration, positional modeling, and token interaction, rather than only parameter scale.

Despite having only 4B parameters, the resulting Qwen3.5-4B model reaches 80.01% average accuracy, outperforming every closed-source model listed in Table 1. It also surpasses substantially larger open-source systems, including the 397B Qwen3.5 model at 77.44% and the 1T-parameter Kimi-K2.6 model at 72.81%. These comparisons show that using Modality Balance as privileged information enables a compact model to achieve the strongest average performance among the reported single-forward-pass models across a general suite of multimodal tasks.

Takeaway 2: OPD-V maintains concise responses throughout training. Response length provides a behavioral view of reasoning efficiency, because shorter successful trajectories can reflect more direct problem

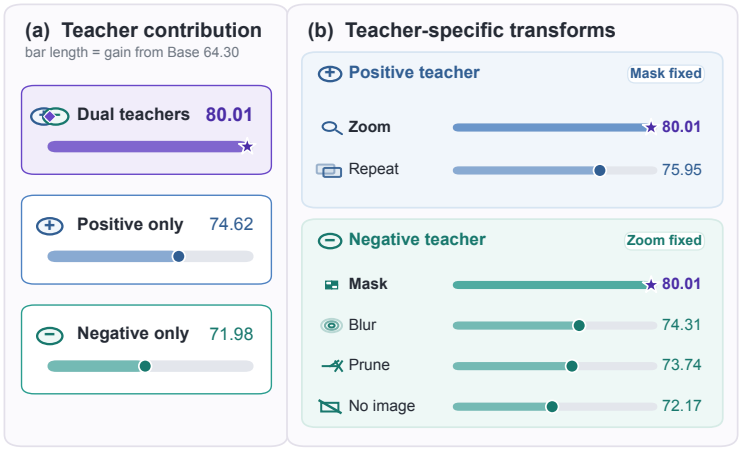


Figure 6: Ablation studies on OPD-V components. Bar lengths encode gains from the 64.30% base model, and labels report absolute accuracy. (a) Combining the Positive Teacher and Negative Teacher yields the highest accuracy (80.01%). (b) Teacher-specific image-operation ablations: the Positive Teacher comparison holds Mask Image fixed and evaluates Repeat Image against Zoom-In Image, whereas the Negative Teacher comparison holds Zoom-In Image fixed and evaluates No Image, Prune, Blur, and Mask Image. Zoom-In Image and Mask Image form the strongest tested pair.

solving [37, 38]. Figure 3(a) separates OPD-V from standard OPSD after the early training stage. The OPSD trajectory grows progressively longer, whereas OPD-V remains comparatively short and stable. Over the final ten recorded steps, the mean response length is 140.9 tokens for OPD-V and 553.8 tokens for OPSD, a 74.5% reduction. OPD-V therefore pairs the performance improvement in Table 1 with a compact response regime throughout training. The shorter responses also reduce token-level computation, which contributes directly to the efficiency result in RQ2.

Takeaway 3: OPD-V preserves stable policy entropy after warm-up. Figure 3(c) shows closely bounded entropy trajectories for both Qwen3.5 scales after warm-up. Their final rolling means remain at 0.827 for 4B and 0.761 for 9B. The shared pattern indicates that OPD-V preserves a non-degenerate policy distribution across model sizes. Together with the sustained accuracy gains, this stability shows that OPD-V maintains an active learning signal throughout training.

Takeaway 4: The Modality-Balance Trust Region persists across training and model scales. Figure 3(b) tracks the fraction of on-policy tokens whose Modality-Balance Logits Margin is positive. After warm-up, the trust-region fraction averages 53.8% for Qwen3.5-4B and 49.6% for Qwen3.5-9B, remaining near one half of each rollout. The region thus stays selective throughout optimization: the Positive Teacher and Negative Teacher consistently distinguish a subset of tokens for self-distillation. Similar fractions at 4B and 9B further show that this selection behavior is retained across scales.

4.3 RQ2: Computational Efficiency

Takeaway: OPD-V reduces step time through shorter rollouts and more efficient privileged-input construction. Figure 5 shows that mean step latency decreases from 352 s to 240 s on Qwen3.5-4B and from 451 s to 340 s on Qwen3.5-9B, corresponding to reductions of 31.8% and 24.7%, respectively. Two stages account for this difference. First, the shorter OPD-V responses in Figure 3(a) reduce the tokens processed during generation, student and teacher forward passes, and backpropagation. This effect is most visible on the 4B backbone, where the final-ten-step response length is 74.5% lower than that of OPSD. Second, the form of privileged information changes the cost of constructing and evaluating teacher inputs. Standard OPSD inserts the reference target into an answer-hint prompt and reconstructs the multimodal teacher sequence. OPD-V instead builds the Positive Teacher and Negative Teacher inputs by applying the Zoom-In Image and Mask Image operations while retaining the matched textual context. In the 9B timing logs,

teacher-batch construction and preprocessing decrease from 79.2 s for OPSD to 15.3 s for OPD-V. Although OPD-V evaluates two teachers, their combined forward time is 27.1 s, below the 40.4 s required by the single OPSD teacher. Thus, the number of teacher passes alone does not determine training cost; response length and the construction of privileged inputs jointly explain why OPD-V trains faster.

4.4 RQ3: Mechanistic Validity

Takeaway 1: The Positive Teacher and Negative Teacher provide complementary supervision. Figure 6(a) shows that each teacher is effective in isolation. The Negative Teacher alone improves the base model from 64.30% to 71.98%, while the Positive Teacher alone reaches 74.62%. Combining them yields 80.01%, exceeding the stronger single-teacher variant by 5.39 percentage points and the Negative-Teacher-only variant by 8.03 points. This additional gain establishes the complementarity of the two teachers. The Positive Teacher supplies the target distribution, whereas the Negative Teacher determines which token positions enter the Modality-Balance Trust Region; the combined result aligns with these distinct roles in the OPD-V objective.

Takeaway 2: Zoom-In Image and Mask Image form the strongest tested operation pair. Figure 6(b) varies one teacher operation at a time. With Mask Image fixed for the Negative Teacher, replacing Zoom-In Image with Repeat Image lowers accuracy from 80.01% to 75.95%. With Zoom-In Image fixed for the Positive Teacher, Mask Image reaches 80.01%, compared with 74.31% for Blur, 73.74% for Prune, and 72.17% for No Image. These comparisons reveal two properties of the selected pair. Zoom-In Image changes the prominence of the task-relevant region, whereas repeating the Original Image leaves its visual composition unchanged. Mask Image modifies selected regions while preserving the remaining image context, whereas No Image removes that context entirely. Among the tested operations, the resulting Zoom-In–Mask produces the most effective Modality-Balance Trust Region and the highest final accuracy.

5 Conclusion

This work identifies Modality Balance as privileged information for OPSD and introduces OPD-V, which uses the Positive Teacher and Negative Teacher to define a Modality-Balance Trust Region. Across six benchmarks and four MLLM backbones, OPD-V consistently improves reasoning performance while reducing training cost.

References

[1] Xuanle Zhao, Qiushi Sun, Jingyu Xiao, Xuexin Liu, Haoyue Yang, Qiaosheng Chen, Xianzhen Luo, Jing Huang, Yufeng Zhong, Lei Chen, Shuai Fu, Zhenlin Wei, Jinhe Bi, Lei Jiang, Haibo Qiu, Siqi Yang, Peng Shi, Jian Hu, and Zhixiong Zeng. Beyond nl2code: A structured survey of multimodal code intelligence, 2026. URL https://arxiv.org/abs/2606.15932.

[2] Xiaowen Ma, Yunpu Ma, Chenyang Lin, Sikuan Yan, Jinhe Bi, Zixuan Cao, Yijun Tian, Volker Tresp, and Hinrich Schuetze. Self-evolving multi-agent systems via textual backpropagation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9918–9951, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.483. URL https://aclanthology.org/2026.findings-acl.483/.

[3] Gengyuan Zhang, Jinhe Bi, Jindong Gu, Yanyu Chen, and Volker Tresp. Spot! revisiting video-language models for event understanding. *arXiv preprint arXiv:2311.12919*, 2023.

[4] Xingyue Huang, Rishabh, Gregor Franke, Ziyi Yang, Jiamu Bai, Weijie Bai, Jinhe Bi, Zifeng Ding, Yiqun Duan, Chengyu Fan, Wendong Fan, Xin Gao, Ruohao Guo, Yuan He, Zhuangzhuang He, Xianglong Hu, Neil Johnson, Bowen Li, Fangru Lin, Siyu Lin, Tong Liu, Yunpu Ma, Hao Shen, Hao Sun, Beibei Wang, Fangyijie Wang, Hao Wang, Haoran Wang, Yang Wang, Yifeng Wang, Zhaowei Wang, Ziyang Wang, Yifan Wu, Zikai Xiao, Chengxing Xie, Fan Yang, Junxiao Yang, Qianshuo Ye, Ziyu Ye, Guangtao Zeng, Yuwen Ebony Zhang, Zeyu Zhang, Zihao Zhu, Bernard Ghanem, Philip Torr, and Guohao Li. Loong:

Synthesize long chain-of-thoughts at scale through verifiers, 2025. URL https://arxiv.org/abs/2509.03059.

[5] 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, pages 21246–21263, 2024.

[6] 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.

[7] 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.

[8] Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. Dopd: Dual on-policy distillation, 2026. URL https://arxiv.org/abs/2606.30626.

[9] Jinhe Bi, Chennan Zhou, Zengjie Jin, Aniri, Shuo Lu, Wenke Huang, Hu Cao, Xun Xiao, Zhihong Zhu, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. Reflectrl: Learning from golden negative trajectories via reflective-to-direct reasoning, 2026. URL https://arxiv.org/abs/2608.03972.

[10] 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.

[11] 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*, 2026.

[12] 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.

[13] 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.

[14] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15230–15250, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.739/.

[15] Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Generalizing from SIMPLE to HARD visual reasoning: Can we mitigate modality imbalance in VLMs? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48192–48244. PMLR, 2025. URL https://proceedings.mlr.press/v267/park25j.html.

[16] Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*, 2025.

[17] Yu Zhang, Mufan Xu, Xuefeng Bai, Kehai Chen, Pengfei Zhang, Yang Xiang, and Min Zhang. Instruction anchor: Dissecting the mechanistic dynamics of modality arbitration, 2026. URL https://arxiv.org/abs/2602.03677.

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

[19] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[20] 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*, pages 13084–13094, 2024.

[21] 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.

[22] 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, pages 7907–7915, 2025.

[23] 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.

[24] Google. Gemini 3.1 pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026.

[25] Google. Gemini 3.5 flash. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026.

[26] OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

[27] OpenAI. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025.

[28] 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*, 2025.

[29] 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.

[30] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

[31] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

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

[33] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

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

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

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

[37] Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. *ArXiv*, abs/2502.07266, 2025. URL https://api.semanticscholar.org/CorpusID:276259519.

[38] Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don't overthink it. preferring shorter thinking chains for improved LLM reasoning, 2026. URL https://openreview.net/forum?id=nhUlA8iMkD.

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

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

[41] Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Cot-kinetics: A theoretical modeling assessing lrm reasoning process. *ArXiv*, abs/2505.13408, 2025. URL https://api.semanticscholar.org/CorpusID:278769227.

[42] Jinhe Bi, Aniri, Minglai Yang, Xingcheng Zhou, Wenke Huang, Sikuan Yan, Yujun Wang, Zixuan Cao, Michael Färber, Xun Xiao, Volker Tresp, and Yunpu Ma. EchoRL: Reinforcement learning via rollout echoing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=A6az59SGtF.

[43] Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *arXiv preprint arXiv:2606.19120*, 2026.

[44] Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *ArXiv*, abs/2502.12119, 2025. URL https://api.semanticscholar.org/CorpusID:276421326.

[45] Tianfan Peng, Yuntao Du, Pengzhou Ji, Shijie Dong, Kailin Jiang, Mingchuan Ma, Yijun Tian, Jinhe Bi, Qian Li, Wei Du, Feng Xiao, and Lizhen Cui. Can visual input be compressed? a visual token compression benchmark for large multimodal models, 2025. URL https://arxiv.org/abs/2511.02650.

[46] Yujun Wang, Aniri , Jinhe Bi, Soren Pirk, and Yunpu Ma. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12): 10306–10314, Mar. 2026. doi: 10.1609/aaai.v40i12.38000. URL https://ojs.aaai.org/index.php/AAAI/article/view/38000.

[47] Kailin Jiang, Ning Jiang, Yuntao Du, Yuchen Ren, Yuchen Li, Yifan Gao, Jinhe Bi, Yunpu Ma, Qingqing Liu, Xianhao Wang, Yifan Jia, Hongbo Jiang, Yaocong Hu, Bin Li, and Lei Liu. Mined: Probing and updating with multimodal time-sensitive knowledge for large multimodal models, 2025. URL https://arxiv.org/abs/2510.19457.

[48] Kailin Jiang, Hongbo Jiang, Ning Jiang, Zhi Gao, Jinhe Bi, Yuchen Ren, Bin Li, Yuntao Du, Lei Liu, and Qing Li. Kore: Enhancing knowledge injection for large multimodal models via knowledge-oriented augmentations and constraints, 2025. URL https://arxiv.org/abs/2510.19316.

Appendix

A Related Work

A.1 From SFT to On-Policy Self-Distillation

Supervised fine-tuning (SFT) trains a model on reference responses with teacher-forced prefixes. The objective is simple and stable, yet the supervision remains tied to fixed reference trajectories rather than the states produced by the current student. Knowledge distillation changes the target from a single reference token to a teacher distribution [39, 40, 41, 42]. On-Policy Distillation (OPD) brings this distributional supervision to student-generated trajectories: the student first samples its own response, and the teacher evaluates the prefixes actually visited by the student [5, 6]. Compared with SFT, OPD exposes model errors on the model's own rollouts and provides token-level feedback at those states.

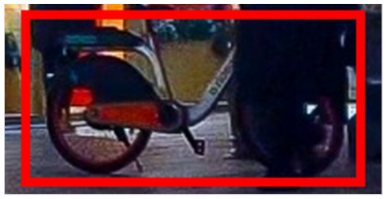
On-Policy Self-Distillation (OPSD) keeps the on-policy structure while replacing the external teacher with a detached copy of the same model [7]. At each token position, the teacher scores the same student-generated prefix, and the distillation loss updates the student distribution at that visited state. This makes the teacher condition central: the trajectory determines where supervision is applied, while the condition determines what signal is distilled. In MLLMs, the condition includes the image input because each prediction depends on the Original Image, textual query, and autoregressive prefix. Recent visual OPD and OPSD methods use transformed visual inputs or visual reasoning signals to make teacher supervision more informative [10, 11, 12, 13, 43]. OPD-V adds a Modality Balance criterion before applying self-distillation, favoring teacher support that reflects stronger use of visual evidence rather than confidence driven mainly by textual context.

Algorithm 1 OPD-V Training

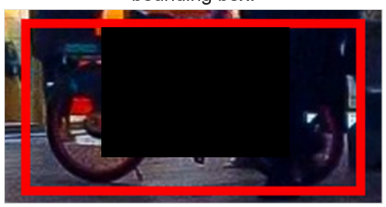
Require: Training set $\mathcal{D}$, student model p_θ, rollout count n, EMA update rate τ
Ensure: Updated student parameters θ

Initialize teacher parameters $\bar{\theta} \leftarrow \theta$
for each training step **do**
  Sample a mini-batch $\mathcal{B} \subset \mathcal{D}$
  Initialize mini-batch loss $\mathcal{L}_{\mathcal{B}} \leftarrow 0$
  for each $(I, x) \in \mathcal{B}$ **do**
    Obtain I^{zoom} from the training data
    $I^{\text{mask}} \leftarrow \text{Mask}(I^{\text{zoom}})$
    for $j = 1, \ldots, n$ **do**
      Sample rollout $y = (y_1, \ldots, y_T) \sim p_\theta(\cdot \mid I, x)$
      Initialize rollout loss $\mathcal{L}_y \leftarrow 0$
      for each $t \in \mathcal{T}_y$ **do**
        Set $r_t \leftarrow 1$ for generated response tokens
        Compute $p_{\theta,t}$, q_t^+, and q_t^-
        $\delta_t^{\text{MB}} \leftarrow \log q_t^+(y_t) - \log q_t^-(y_t)$
        if $\delta_t^{\text{MB}} > 0$ and $r_t = 1$ **then**
          $D_t \leftarrow D_{\text{JS}}(q_t^+, p_{\theta,t})$
          $\mathcal{L}_y \leftarrow \mathcal{L}_y + \delta_t^{\text{MB}} D_t$
        end if
      end for
      $\mathcal{L}_{\mathcal{B}} \leftarrow \mathcal{L}_{\mathcal{B}} + \mathcal{L}_y / \sum_{t \in \mathcal{T}_y} r_t$
    end for
  end for
  Update θ by minimizing $\mathcal{L}_{\mathcal{B}} / (n|\mathcal{B}|)$
  Update teacher parameters $\bar{\theta} \leftarrow (1 - \tau)\bar{\theta} + \tau\theta$
end for

Student input image
Original Image



Positive teacher input image
Zoom-In Image cropped from the red bounding box.



Negative teacher input image
Mask Image constructed by randomly masking the Zoom-In Image.

Prompt image: Student input image.

Image placeholder: `<image>`

Question: What is the color of the reflector attached to the rear fender?

Visual focus instruction: Only focus on the objects inside the red bounding box in the image to answer this question.

Answer choices:
A. white
B. black
C. red
D. yellow

Output format: Answer with the option's letter from the given choices.

Ground-truth answer: C

Figure 7: Training example for a rear-fender reflector question.

A.2 Modality Imbalance in MLLMs

Modality Imbalance denotes the tendency of an MLLM to depend disproportionately on textual context relative to visual input during response generation. Prior studies show that textual information can dominate multimodal prediction even when the query requires visual evidence [14, 15, 16, 17, 44, 45, 46, 47, 48]. This creates a training-time ambiguity for OPSD. A confident teacher distribution may reflect visual evidence, language priors, or the textual query, so the relevant visual content can remain underused even when token-level supervision appears strong.

Once OPSD uses teacher distributions as token-level supervision, this ambiguity affects token selection. Selecting tokens by teacher confidence alone can preserve supervision driven mainly by textual context, consistent with the shortcut supervision observed in multimodal OPSD [43]. OPD-V makes Modality Balance part of the self-distillation criterion, so token selection depends on whether teacher support reflects stronger use of visual evidence rather than confidence alone.

B OPD-V Algorithm

In algorithm 1 for OPD-V: each prompt first yields n on-policy rollouts, and teacher scoring is then aligned to the same visited prefixes. At token position t, $p_{\theta,t}$, q_t^+, and q_t^- are calculated under the shared prefix. The Positive Teacher with the Zoom-In Image provides the Jensen–Shannon distillation target, while the Negative Teacher with the Mask Image contributes the paired comparison for the Modality-Balance Trust Region. Only generated response tokens with $r_t = 1$ enter the loss, and the detached teacher is updated by EMA after the student update.

Benchmark	Samples	Scoring Rule
V* Bench	191	Rule-based answer matching is applied first; unresolved responses are compared with the verified answer by a Yes/No judge.
ZoomBench	845	Generic answer matching handles normalized responses first; unresolved cases use the verified-answer Yes/No judge.
HR-Bench 4K	800	Multiple-choice matching checks direct answer matches and option-letter matches before the Yes/No judge handles unresolved responses.
HR-Bench 8K	800	Multiple-choice matching follows the HR-Bench 4K rule.
MME-RealWorld EN	23609	Multiple-choice matching and the Yes/No judge are applied to the English subset.
MME-RealWorld CN	5917	The Chinese subset uses the MME-RealWorld multiple-choice scoring rule.

Table 2: Evaluation protocol for benchmarks.

B.1 Loss and Top-K Distillation Details

Token selection uses the Modality-Balance Logits Margin δ_t^{MB} on the student-generated token y_t. A positive Logits Margin indicates that the same token receives more support from the Positive Teacher than from the Negative Teacher, and these positions form the Modality-Balance Trust Region. Within this region, δ_t^{MB} scales the selected on-policy tokens during self-distillation. The distillation target remains distributional: D_t compares the Positive Teacher with the Zoom-In Image and student distributions, rather than only the probabilities assigned to y_t.

Full-vocabulary logit distillation is memory intensive for MLLMs, so the implementation first takes the student top-K token set at each response position:

$$\mathcal{K}_t = \mathrm{TopK}(p_{\theta,t}, K)\,, \qquad K = 100. \tag{10}$$

The Positive Teacher with the Zoom-In Image is evaluated on the same set $\mathcal{K}_t$. For each distribution $s \in \{p_{\theta,t}, q_t^+\}$, the top-$K$ probabilities are retained and one tail bucket stores the remaining vocabulary mass:

$$s_t^{\mathrm{tail}} = 1 - \sum_{v \in \mathcal{K}_t} s(v). \tag{11}$$

The $(K+1)$-dimensional form preserves the probability assigned to tokens outside $\mathcal{K}_t$ without materializing the full vocabulary. With q_t^+ and $p_{\theta,t}$ represented by their tail-adjusted top-K distributions, per-token distillation term expands as

$$\begin{aligned} D_t &= D_{\mathrm{JS}}\big(q_t^+, p_{\theta,t}\big) \\ &= \frac{1}{2} D_{\mathrm{KL}}\big(q_t^+ \,\|\, m_t\big) + \frac{1}{2} D_{\mathrm{KL}}(p_{\theta,t} \,\|\, m_t)\,, \\ m_t &= \frac{1}{2}\left(q_t^+ + p_{\theta,t}\right). \end{aligned} \tag{12}$$

The Negative Teacher with the Mask Image enters only through δ_t^{MB}, while the Positive Teacher with the Zoom-In Image remains the distillation target.

Loss reduction follows the valid-token count rather than the sum of margin weights. The normalization stays tied to generated response tokens, while δ_t^{MB} controls the strength of the selected distillation signal.

B.2 Image Operations

OPD-V changes the teacher image input while keeping the textual query, on-policy prefix, and teacher parameters fixed. The Zoom-In Image and Mask Image are derived from the same Original Image, which includes the target-region cue. The Positive Teacher with the Zoom-In Image receives I^{zoom}. In the experiments, I^{zoom} is taken from the evidence-centered crop supplied with the Vision-OPD training data [10]. The crop is defined by the bounding box around the task-relevant region and is magnified before image processing, increasing the prominence of the visual evidence required by the query.

Backbone	Params	Model Identifier	VLM and Perception Stack
Qwen3.5-4B	4B	Qwen/Qwen3.5-4B	Qwen3.5 multimodal backbone with early vision–text fusion and a 3:1 hybrid language backbone that interleaves Gated DeltaNet linear-attention layers with full-attention layers.
Qwen3.5-9B	9B	Qwen/Qwen3.5-9B	Qwen3.5 multimodal backbone with the same vision–text fusion design as Qwen3.5-4B, scaled to the 9B parameter setting.
Qwen3-VL-4B-Instruct	4B	Qwen/Qwen3-VL-4B-Instruct	Qwen3-VL vision-language stack with Interleaved-MRoPE and DeepStack, which injects multi-level ViT features into the language backbone.
Qwen3-VL-8B-Instruct	8B	Qwen/Qwen3-VL-8B-Instruct	Qwen3-VL vision-language stack with the same Interleaved-MRoPE and DeepStack design as Qwen3-VL-4B-Instruct, scaled to the 8B parameter setting.

Table 3: Backbone configurations used in the OPD-V experiments.

The Negative Teacher with the Mask Image receives I^{mask}, computed as $I^{\text{mask}} = \text{Mask}(I^{\text{zoom}})$. The Mask Image is constructed from the evidence-centered cropped Original Image I^{zoom} by replacing a randomly selected rectangular region with black pixels before image processing, forming a visually weakened condition while preserving most image context. The Positive Teacher with the Zoom-In Image and the Negative Teacher with the Mask Image share the same textual query, on-policy prefix, and EMA parameters, so their distributional difference reflects the image operation applied to the teacher input.

C Training Prompt Example

Figure 7 presents one training sample in the same field structure used by the prompt: prompt image, image placeholder, question, visual focus instruction, answer choices, output format, and ground-truth answer. The left column marks which image condition is paired with the shared textual prompt in the student and teacher passes.

D Model Configuration Details

Table 3 separates the four MLLM backbones by model identifier, parameter scale, and VLM/perception stack. This makes the comparison across Qwen3.5 and Qwen3-VL explicit, including the early vision–text fusion design in Qwen3.5 and the Interleaved-MRoPE with DeepStack design in Qwen3-VL.

The training split contains 6241 visual reasoning samples from the dataset [10]. Each sample provides the Original Image, textual query, verified target, and the dataset field used as I^{zoom}. The Mask Image is generated from the evidence-centered cropped Original Image I^{zoom} during training, and evaluation uses only the Original Image and textual query.

E Evaluation Protocol Details

Evaluation is run in non-thinking mode. Each response is normalized by deterministic matching before unresolved cases are passed to a Yes/No judge. For multiple-choice tasks, direct option matching and first-letter option matching are checked first. Table 2 pairs each benchmark with its sample count and scoring rule.

The reported Average is the unweighted mean over the six benchmark accuracies:

$$\text{Average} = \frac{1}{6} \sum_{b \in \mathcal{B}_{\text{eval}}} \text{Acc}_b. \tag{13}$$

Setting	Value	Role
GPU allocation	1 node, 4 H200 GPUs	Training worker allocation
Training batch size	48	Mini-batch size per training step
Actor update mini-batch size	48	Mini-batch size for actor update
Rollouts per prompt	8	Number of on-policy responses sampled per prompt
Maximum prompt length	8192 tokens	Prompt truncation limit
Maximum response length	1024 tokens	On-policy generation limit
Training epochs	1	Number of passes over the training set
Policy loss mode	OPD-V	Self-distillation objective used for actor update
Distillation objective	Jensen–Shannon divergence	Token-level distribution matching
Top-K size	100	Candidate set for distillation distributions
Teacher parameter update	EMA of student weights	Detached teacher update rule
Teacher EMA update rate	0.05	Student-weight coefficient in the EMA update
Teacher construction	Always on	Teacher inputs are built at every training step
Positive teacher image field	bbox_images	Source of the Zoom-In Image
Negative teacher image operation	Mask Image	Random rectangular masking of the Zoom-In Image
Rollout correction mode	Token	Token-level rollout correction
Rollout correction threshold	2.0	Clipping threshold for rollout correction
Checkpoint frequency	-1	No default checkpoint saving during training
Validation frequency	-1	No default validation during training
Logging mode	Offline W&B	Logs are synced after training

Table 4: Training hyperparameters for OPD-V.

This benchmark-level mean differs from sample-weighted overall accuracy, which pools correct predictions over all 32162 examples before computing a single percentage.

F Hyperparameters and Compute

Table 4 records the parameters that control training scale and the teacher–student update schedule. The training batch size and actor update mini-batch size are both 48, so each step uses the sampled mini-batch directly for the actor update. With $n = 8$, every prompt contributes eight on-policy responses before the mini-batch loss is reduced by $n|\mathcal{B}|$ as in Algorithm 1.

Sequence limits separate input context from generated trajectories. The 8192-token prompt limit bounds the Original Image, textual query, and prompt formatting, while the 1024-token response limit caps rollout length and the number of token positions scored by the student and teachers. The top-K parameter is set to $K = 100$, which fixes the candidate set size used by the tail-adjusted distillation distributions described in Section B.

The EMA update rate $\tau = 0.05$ is the coefficient on the current student weights in the detached teacher update; equivalently, each update retains 0.95 of the previous teacher parameters. Token-level rollout correction uses threshold 2.0 before the actor update. Periodic checkpointing and validation are disabled by the -1 settings, and offline W&B logging stores run statistics without adding online synchronization during training.