MAGISTRATEOPD: VERIFY BEFORE DISTILLING FOR RELIABLE ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing On-Policy Distillation (OPD) methods combine policy learning from student-generated trajectories with dense token-level teacher guidance. Subsequent methods further strengthen this supervision by equipping a positive teacher with an evidence-centered zoom-in crop and constructing a negative teacher from a masked image. However, such dense supervision implicitly relies on the teacher responses being sufficiently reliable. For smaller models such as 4B-scale multimodal large language models(MLLMs), additional zoom-in evidence may improve the positive teacher's correct-response rate while its absolute accuracy remains relatively low, substantially weakening the effective supervision signal available to the student. In this paper, we propose MagistrateOPD, which applies pre-judge verification to both teacher responses before token-level on-policy distillation. The magistrate uses an MLLM to compare both teacher responses with the reference answer at the semantic level before admitting a sample into distillation. Pre-judge verification requires the positive-teacher response to match the reference answer and the negative-teacher response not to match it. In addition, we introduce a negative-teacher loss that pushes the student away from the masked-image negative-teacher distribution, thereby strengthening the negative-teacher supervision signal. We evaluate MagistrateOPD on V* Bench, ZoomBench, HR-Bench, and MME-RealWorld. MagistrateOPD achieves 92.15 on V* Bench, 84.12 on HR-Bench-4K, and 72.94 on MME-RealWorld-EN, as well as 61.89 on ZoomBench. Across all six evaluation splits, MagistrateOPD achieves an average score of 78.02, outperforming the base model, SFT, and OPSD by 13.72, 11.17, and 8.76 points, respectively.

1 INTRODUCTION

On-policy distillation (OPD) Agarwal et al. (2024) combines student-generated trajectories with dense token-level teacher guidance, aligning supervision with the student's current policy. On-policy self-distillation (OPSD) Zhao et al. (2026) instantiates teacher and student from the same model under different conditioning contexts, allowing privileged training-time information to guide the student. In multimodal OPD, this privileged information can take the form of additional visual evidence. Vision-OPD Yuan et al. (2026) conditions the teacher on an evidence-centered crop to transfer privileged regional perception to a full-image student, while OPD-V Aniri et al. (2026) further constructs a masked-image negative teacher alongside the zoom-in positive teacher.

However, these multimodal OPD methods implicitly assume that providing stronger or privileged visual information is sufficient to produce reliable teacher supervision. In practice, this assumption can fail, particularly for smaller models such as 4B-scale MLLMs. On the 6.2K synthetic training set Aniri et al. (2026), the zoom-in positive teacher can still produce incorrect responses despite receiving additional visual evidence. As shown in Figure 1 (upper right), only 45.83% of samples satisfy the positive-correct and negative-incorrect condition at the start of training. Using such responses as distillation targets can misdirect the student across many tokens of its on-policy trajectories. The validity of both teacher responses therefore determines whether their complementary supervision should be applied to a training sample.

A natural alternative is to filter the training set before optimization and retain only samples that satisfy both teacher-response conditions. However, such static filtering evaluates sample validity

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

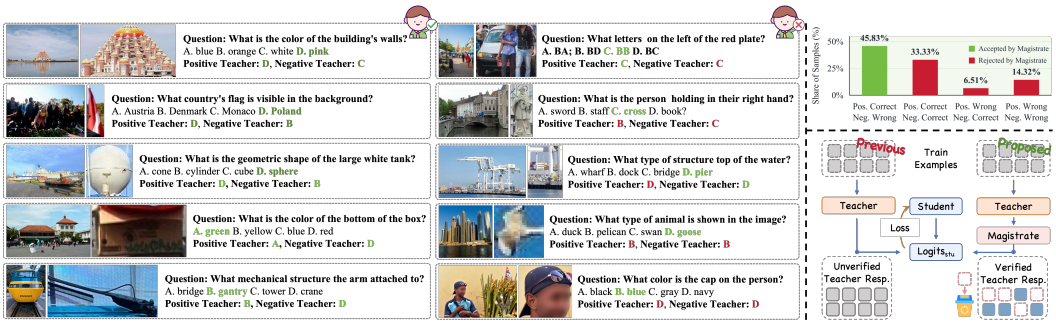


Figure 1: Motivation for MagistrateOPD's response-level pre-judge verification in on-policy distillation. The left panel shows selected valid and invalid Qwen3.5-4B teacher-response pairs on the 6.2K synthetic training set Aniri et al. (2026). A sample is valid only when the positive-teacher response is correct and the negative-teacher response is incorrect. The upper-right panel shows that only 45.83% of samples meet this validity criterion at training step 0. The lower-right panel contrasts unverified distillation, in which teacher responses can supervise the student regardless of semantic validity, with MagistrateOPD, which admits only samples verified as valid by the magistrate.

only under the initial teacher and preferentially retains samples that are already easy at the beginning of training. Because the teacher itself evolves during on-policy training, samples that are initially invalid can later become useful supervision as the teacher improves. Teacher-response reliability should therefore be evaluated dynamically rather than treated as a fixed property of the training set.

To this end, we propose MagistrateOPD, which applies pre-judge verification to both teacher responses during training. As illustrated in Figure 1 (lower right), this gate prevents unverified teacher responses from supervising the student. The magistrate uses an MLLM to compare both teacher responses with the reference answer at the semantic level. A sample is admitted into distillation only when the positive-teacher response is semantically correct and the negative-teacher response is semantically incorrect. Pre-judge verification therefore requires the magistrate to assess semantic equivalence rather than solve the original fine-grained visual problem. By repeating pre-judge verification throughout training, MagistrateOPD dynamically adapts the usable supervision set to the evolving teacher instead of relying on one-time offline data filtering.

Inspired by contrastive learning Tian et al. (2020), we further introduce a **negative-teacher loss** as a counterpart to the positive-teacher distillation loss. While the positive-teacher loss pulls the student toward the verified positive-teacher distribution, the negative-teacher loss pushes it away from the masked-image negative-teacher distribution. This bidirectional supervision enables the student to learn jointly from what should be approached and what should be avoided. Consequently, both positive and negative teachers provide explicit optimization signals rather than using the negative teacher only for token selection.

We evaluate MagistrateOPD on V* Bench Wu & Xie (2024), ZoomBench Wei et al. (2026), HR-Bench Wang et al. (2025), and MME-RealWorld Zhang et al. (2024). MagistrateOPD achieves 92.15 on V* Bench, 84.12 on HR-Bench-4K, and 72.94 on MME-RealWorld-EN, exceeding OPSD Zhao et al. (2026) by 7.86, 6.87, and 11.74 points, respectively. On ZoomBench, MagistrateOPD achieves 61.89, exceeding OPSD by 4.38 points. MagistrateOPD achieves an overall average score of 78.02 in our evaluation, representing an 8.76-point improvement over OPSD.

Our main contributions are summarized as follows:

- **Pre-Judge Verification:** We introduce pre-judge verification to check both teacher responses dynamically before admitting a sample into on-policy distillation.
- **Negative-Teacher Loss:** Inspired by contrastive learning, we introduce a negative-teacher loss that explicitly pushes the student away from the masked-image negative-teacher distribution.

2 Related Work

2.1 Multimodal Large Language Models and Fine-Grained Perception

Modern multi-modal large language models(MLLMs) build on advances in visual representation learning and vision–language alignment. CLIP (Radford et al., 2021) learns transferable visual representations from image–text contrastive supervision, while Flamingo (Alayrac et al., 2022) connects pretrained visual and language components to support few-shot learning from interleaved multimodal inputs. BLIP-2 (Li et al., 2023) bridges frozen image encoders and language models through a lightweight querying transformer. LLaVA (Liu et al., 2023b) and its improved baseline (Liu et al., 2023a) demonstrate the effectiveness of visual instruction tuning for general-purpose multimodal interaction. Subsequent models expand these capabilities through visual grounding and text reading in Qwen-VL (Bai et al., 2023), dynamic-resolution processing in Qwen2-VL (Wang et al., 2024) and Qwen2.5-VL (Bai et al., 2025), and scaled visual representations with progressive language alignment in InternVL (Chen et al., 2023).

Despite these advances, fine-grained perception remains challenging when the answer depends on small or visually crowded regions. V* (Wu & Xie, 2024) introduces guided visual search and a benchmark targeting such details; HR-Bench (Wang et al., 2025) and MME-RealWorld (Zhang et al., 2024) further expose difficulties with high-resolution and real-world visual evidence. Zooming without Zooming (Wei et al., 2026) transfers region-grounded supervision to full-image prediction through region-to-image distillation and introduces ZoomBench. These studies motivate our use of privileged visual evidence during training.

2.2 Reinforcement Learning and Preference Optimization

Reinforcement learning provides an alternative route to post-training language and multimodal models. Proximal Policy Optimization (PPO) (Schulman et al., 2017) optimizes a surrogate policy objective from sampled interactions, and InstructGPT (Ouyang et al., 2022) applies reinforcement learning from human feedback to improve instruction following. Direct Preference Optimization (DPO) (Rafailov et al., 2023) instead learns directly from preferred and rejected responses through a classification objective, without explicitly fitting a reward model. For reasoning, DeepSeek-Math (Shao et al., 2024) introduces Group Relative Policy Optimization (GRPO), while DeepSeek-R1 (DeepSeek-AI et al., 2025) demonstrates the potential of large-scale reinforcement learning with verifiable feedback.

In the multimodal setting, LLaVA-RLHF (Sun et al., 2023) augments reward modeling with factual information to improve visual alignment, and RLAIF-V (Yu et al., 2024) uses open-source AI feedback to improve response trustworthiness. DeepEyes (Zheng et al., 2025) trains active visual perception through reinforcement learning, allowing the model to request additional visual evidence during reasoning. These approaches connect learning quality to feedback quality. MagistrateOPD uses correctness feedback to determine whether a teacher pair is admissible; on admitted examples, optimization receives dense token-distribution targets along student-generated trajectories. Its verification gate therefore serves a different role from a scalar reward or a sequence-level preference label.

2.3 Knowledge Distillation and On-Policy Self-Distillation

Knowledge distillation (Hinton et al., 2015) transfers predictive information from a teacher to a student, with sequence-level distillation (Kim & Rush, 2016) extending this idea to teacher-generated outputs in sequence modeling. Training on model-induced states also connects to imitation learning: DAgger (Ross et al., 2011) addresses the distribution shift caused by sequential decisions by collecting supervision under the learner's induced state distribution. For language models, MiniLLM (Gu et al., 2023) uses reverse KL divergence with on-policy optimization, while generalized knowledge distillation (Agarwal et al., 2024) queries the teacher on student-generated prefixes to address training–inference mismatch. DistiLLM (Ko et al., 2024) studies skew KL objectives and adaptive off-policy reuse of student-generated outputs. Rethinking On-Policy Distillation (Li et al., 2026b) further identifies teacher–student compatibility and the teacher's additional capabilities as important factors in successful OPD.

On-policy self-distillation (OPSD) (Zhao et al., 2026) uses the same model under different conditioning contexts, allowing privileged information to provide dense supervision on the student's own trajectories without a separately trained external teacher. Multimodal variants construct this information from visual inputs. Vision-OPD (Yuan et al., 2026) conditions the teacher on an evidence-centered crop and the student on the full image. Visual-OPSD (Li et al., 2026a) transfers reasoning from a teacher conditioned on privileged visual thoughts, while RP-OPSD (Zhu et al., 2026) exploits the information gap between high- and low-resolution views.

OPD-V (Aniri et al., 2026) adds a masked-image negative teacher alongside the zoom-in positive teacher and uses their relative token probabilities to define a modality-balance trust region. This criterion selects positions favored by the privileged visual context, but does not establish the semantic correctness of either complete teacher response. MagistrateOPD complements this token-level selection with dynamic response-level verification and an explicit negative-teacher loss.

2.4 Verification and Learning from Negative Signals

Training verifiers (Cobbe et al., 2021) can select mathematical solutions, while process supervision (Lightman et al., 2023) evaluates intermediate reasoning steps. STaR (Zelikman et al., 2022) iteratively improves reasoning by training on generated rationales that lead to correct answers. LLM-as-a-judge studies (Zheng et al., 2023) establish the usefulness of model-based evaluation while documenting biases and reasoning limitations. Our magistrate performs a narrower, reference-conditioned semantic-equivalence check on both teacher responses. This check is repeated as the teacher evolves and controls admission to distributional distillation; it does not certify every reasoning step or every token-level teacher target.

Negative learning signals also have a substantial history. Unlikelihood training (Welleck et al., 2019) explicitly lowers the probability of undesirable generations, and contrastive representation distillation (Tian et al., 2020) transfers structure through a contrastive objective. In multimodal generation, visual contrastive decoding (Leng et al., 2023) contrasts distributions from original and distorted images to mitigate hallucinations at inference time. MagistrateOPD instead applies a distribution-level avoidance term during training: on verified examples and selected student-generated tokens, it increases Jensen–Shannon divergence from the masked-image negative teacher while matching the positive teacher. The response-level gate provides the correctness conditions under which these complementary teacher roles are used.

3 Method

MagistrateOPD applies response-level pre-judge verification and token-level selection before using teacher distributions to optimize the student. We introduce pre-judge verification to admit samples whose positive-teacher response matches the reference answer and whose negative-teacher response does not. For admitted samples, we retain the modality-balance trust region to select tokens where the positive teacher assigns higher probability than the negative teacher Aniri et al. (2026). We also introduce a negative-teacher loss that pushes the student away from the negative-teacher distribution while positive-teacher distillation draws it toward the positive teacher.

Given an original image I, a zoom-in image I^+, a masked image I^-, a textual query x, and a reference answer $a^\star$, the student observes I, while the positive and negative teachers observe I^+ and I^-, respectively. All three policies share the same model architecture, with the teacher parameters $\bar{\theta}$ updated from the student parameters θ following the underlying OPD training scheme.

As illustrated in Figure 2(a), the positive and negative teachers first generate textual responses under their respective visual conditions, and the magistrate uses an MLLM to verify their semantic relations to the reference answer. Only samples with a correct positive-teacher response and an incorrect negative-teacher response are admitted into subsequent distillation. For an admitted sample, the student generates an on-policy trajectory

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

and both teachers score the same student-generated prefixes under their respective visual inputs. Specifically, their token distributions are

$$q_t^+(\cdot) = p_{\bar{\theta}}(\cdot \mid I^+, x, y_{<t}), \qquad q_t^-(\cdot) = p_{\bar{\theta}}(\cdot \mid I^-, x, y_{<t}). \tag{2}$$

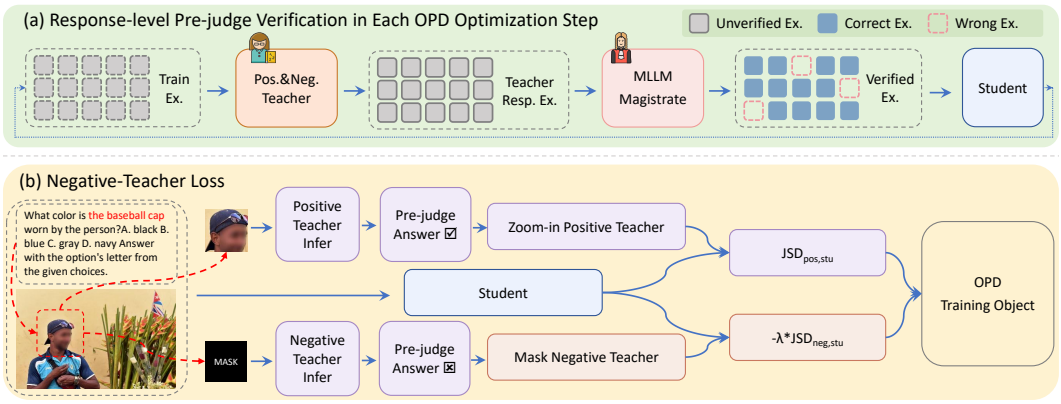


Figure 2: Pre-judge verification and the negative-teacher loss in MagistrateOPD. (a) At each OPD optimization step, the magistrate verifies both teacher responses before examples are used for distillation. (b) On verified examples, the training objective combines positive-teacher distillation and the negative-teacher loss.

At each token position, the two teachers assign log probabilities to the student-generated token y_t. We retain the modality-balance logit margin that compares these probabilities:

$$\delta_t^{\mathrm{MB}} = \log q_t^+(y_t) - \log q_t^-(y_t). \tag{3}$$

This margin defines the modality-balance trust region:

$$\mathcal{R}_{\mathrm{MB}}(y) = \left\{ t \mid \delta_t^{\mathrm{MB}} > 0 \right\}. \tag{4}$$

Only tokens with $\delta_t^{\mathrm{MB}} > 0$ enter token-level distillation. Pre-judge verification therefore selects samples by the complementary correctness conditions of the two teacher responses, while the modality-balance trust region selects eligible tokens within those samples. Figure 2(b) illustrates positive-teacher distillation and the negative-teacher loss on admitted samples. On selected tokens, the positive-teacher term draws the student toward the positive-teacher distribution, whereas the negative-teacher loss pushes it away from the negative-teacher distribution.

3.1 Dynamic Pre-Judge Verification

We apply pre-judge verification to both teacher responses before constructing token-level supervision so that each teacher has its intended semantic role. For each training sample, our positive and negative teachers first generate

$$\hat{a}^+ \sim p_{\bar{\theta}}(\cdot \mid I^+, x), \qquad \hat{a}^- \sim p_{\bar{\theta}}(\cdot \mid I^-, x). \tag{5}$$

We implement the magistrate with an MLLM and denote its semantic-equivalence judgment by $\mathcal{J}$:

$$s^+ = \mathcal{J}(\hat{a}^+, a^\star), \qquad s^- = \mathcal{J}(\hat{a}^-, a^\star), \tag{6}$$

where $\mathcal{J}(\hat{a}, a^\star) = 1$ indicates semantic equivalence and $\mathcal{J}(\hat{a}, a^\star) = 0$ otherwise. Our magistrate determines semantic equivalence between each teacher response and the reference answer without solving the original fine-grained visual task. We define the sample-level pre-judge gate as

$$g = s^+(1 - s^-), \qquad g \in \{0, 1\}. \tag{7}$$

Thus, $g = 1$ only when the positive-teacher response matches the reference and the negative-teacher response does not, and samples with $g = 0$ do not contribute to the distillation objective.

We repeat pre-judge verification at each training step using the current teacher parameters. As the teachers evolve with the student, a sample rejected earlier can later satisfy the response-level conditions. MagistrateOPD therefore updates the usable supervision set throughout training.

3.2 NEGATIVE-TEACHER LOSS

We use the negative teacher in two ways: its token probabilities help select positions through the modality-balance trust region, and its distribution provides an explicit avoidance signal at those positions. Pre-judge verification admits a sample only when the positive-teacher response matches the reference answer and the negative-teacher response does not. Inspired by contrastive learning Tian et al. (2020), we define a **negative-teacher loss** that pushes the student away from the negative-teacher distribution on admitted samples.

For each admitted sample, we apply the modality-balance trust region to select the tokens used for distillation. Let $p_{\theta,t}$ denote the student distribution at a selected token position t, and let $\mathrm{JSD}_{\mathrm{pos},t} = D_{\mathrm{JS}}(q_t^+, p_{\theta,t})$ denote its divergence from the positive teacher. We define the negative-teacher loss at the same position as

$$\mathcal{L}_{\mathrm{neg},t} = -\mathrm{JSD}_{\mathrm{neg},t} = -D_{\mathrm{JS}}(q_t^-, p_{\theta,t}). \tag{8}$$

Minimizing this term increases the divergence between the student and negative-teacher distributions on selected tokens.

We combine positive-teacher distillation, the negative-teacher loss, and reference regularization in the per-token training objective

$$\begin{aligned} \mathcal{L}_t &= \mathrm{JSD}_{\mathrm{pos},t} + \lambda\, \mathcal{L}_{\mathrm{neg},t} + \alpha\, \mathrm{KL}_{\mathrm{ref},t} \\ &= \mathrm{JSD}_{\mathrm{pos},t} - \lambda\, \mathrm{JSD}_{\mathrm{neg},t} + \alpha\, \mathrm{KL}_{\mathrm{ref},t}. \end{aligned} \tag{9}$$

Here $\mathrm{KL}_{\mathrm{ref},t}$ denotes the reference-policy anchoring term, and λ and α control the negative-teacher and reference terms, respectively.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation Details. We conduct all experiments on Qwen3.5-4B using the 6.2K fine-grained visual reasoning samples constructed by Yuan et al. (2026). The student observes the original image, while the positive teacher receives the zoom-in image and the negative teacher receives an 80% randomly masked version of the zoom-in image. We use Jensen–Shannon divergence for token-level distillation and adopt the established visual OPD training recipe for the remaining optimization settings. We use GPT-5.6-Luna to judge whether each teacher response is semantically equivalent to the reference answer during pre-judge verification. For MagistrateOPD, we set the negative-teacher loss coefficient to $\lambda = 1$ and the reference KL coefficient to $\alpha = 0.1$. Unless otherwise stated, MagistrateOPD uses dynamic pre-judge verification for both the positive and negative teachers.

Evaluation Benchmarks. We evaluate MagistrateOPD on V* Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench (Wang et al., 2025), and MME-RealWorld (Zhang et al., 2024). V* Bench evaluates fine-grained recognition of visually small targets, while ZoomBench focuses on visual questions requiring localized evidence from the full image. HR-Bench evaluates high-resolution visual understanding at 4K and 8K resolutions, and MME-RealWorld evaluates real-world multimodal perception in both English and Chinese. Following prior visual OPD studies, we report accuracy on each evaluation split. For compact presentation, HR-4K and HR-8K denote the 4K and 8K splits of HR-Bench, while MME-EN and MME-CN denote the English and Chinese splits of MME-RealWorld, respectively. The average score is computed over all six evaluation splits.

4.2 COMPARISON WITH MULTI-MODAL LARGE LANGUAGE MODEL AND OPD METHODS

We first assess MagistrateOPD's performance under the Qwen3.5-4B backbone and place its results in the context of multi-modal large language models(MLLMs). Table 1 reports scores on four fine-grained visual perception benchmarks for MagistrateOPD, representative closed- and open-source MLLMs, and OPD-based post-training methods. Within the Qwen3.5-4B group, we compare with SFT, OPSD (Zhao et al., 2026), Vision-OPD (Yuan et al., 2026), VA-OPD Liu et al. (2026), and OPD-V (Aniri et al., 2026) at the same model scale, and these results were reproduced locally by us. The MagistrateOPD result in Table 1 uses dynamic pre-judge verification for both teachers.

Table 1: Comparison on fine-grained visual perception benchmarks. Within the Qwen3.5-4B group, the best results are shown in bold and the second-best results are underlined.

Model	Param	V* Bench	ZoomBench	HR-4K	HR-8K	MME-EN	MME-CN	Average
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 (Single Forward Pass; Locally Reproduced Results)								
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	<u>59.76</u>	<u>82.25</u>	<u>80.50</u>	<u>69.54</u>	<u>69.95</u>	<u>75.17</u>
VA-OPD	4B	86.91	51.20	80.12	77.50	68.50	67.10	71.89
OPD-V	4B	<u>89.53</u>	58.93	82.12	77.88	<u>69.54</u>	<u>69.95</u>	74.66
MagistrateOPD	**4B**	**92.15**	**61.89**	**84.12**	**83.00**	**72.94**	**74.01**	**78.02**

MagistrateOPD achieves 92.15 on V* Bench, 61.89 on ZoomBench, 84.12 on HR-4K, 83.00 on HR-8K, 72.94 on MME-EN, and 74.01 on MME-CN, yielding an average score of 78.02. Relative to OPD-V, these results correspond to gains of 2.61, 2.96, 2.00, 5.12, 3.40, and 4.06 points on V* Bench, ZoomBench, HR-4K, HR-8K, MME-EN, and MME-CN, respectively. The average score improves by 3.36 points over OPD-V, and the method exceeds the 397B Qwen3.5 model by 0.58 points while using a 4B backbone.

4.3 Ablation Studies

4.3.1 Ablation of Pre-Judge Verification and Negative-Teacher Loss

We first evaluate the two key components of MagistrateOPD: pre-judge verification and the negative-teacher loss, including four variants with neither component, pre-judge verification only, negative-teacher loss only, and both components. The four variants use the same backbone, training data, and token-level modality-balance trust region; pre-judge verification controls sample admission based on both teacher responses, and the loss switch controls whether the negative-teacher term contributes to optimization.

Table 2: Ablation of the main components in MagistrateOPD.

Pre-Judge Verification	Neg. Loss	V* Bench	ZoomBench	HR-4K	HR-8K	MME-EN	MME-CN	Average
×	×	89.53	58.93	82.12	77.88	69.54	69.95	74.66
✓	×	89.01	60.47	83.00	78.00	69.71	70.19	75.06
×	✓	87.96	55.62	83.50	82.12	71.16	69.07	74.91
✓	✓	**92.15**	**61.89**	**84.12**	**83.00**	**72.94**	**74.01**	**78.02**

As shown in Table 2, the baseline, pre-judge verification alone, the negative-teacher loss alone, and the combined setting achieve average scores of 74.66, 75.06, 74.91, and 78.02, respectively.

The combined setting obtains the highest average score and reaches 92.15 on V* Bench, 84.12 on HR-Bench-4K, and 72.94 on MME-RealWorld-EN.

Adding the negative-teacher loss to pre-judge verification raises the average from 75.06 to 78.02, a 2.96-point gain. We therefore adopt both pre-judge verification and the negative-teacher loss in MagistrateOPD. The negative-teacher loss complements positive-teacher distillation by pushing the student away from the masked-image negative-teacher distribution. Without pre-judge verification, however, either teacher can provide a response with the wrong semantic role for its corresponding objective. Pre-judge verification removes these response-level failure modes, while the negative-teacher loss provides an additional token-level contrastive signal. The joint result supports their complementary effect.

4.3.2 Static Filtering vs. Dynamic Pre-Judge Verification

We evaluate whether dynamic pre-judge verification of both teacher responses is more useful than filtering the training set only once. All three variants retain the negative-teacher loss and share the same backbone, training data, optimization settings, and token-level modality-balance trust region. No verification admits every sample, static filtering checks both teacher responses before training, and dynamic pre-judge verification checks them at each training step.

Table 3: Final evaluation of sample-selection strategies.

Configuration	V* Bench	ZoomBench	HR-4K	HR-8K	MME-EN	MME-CN	Average
No Verification	87.96	55.62	83.50	82.12	71.16	69.07	74.91
Static Pre-Filtering	85.34	54.67	**86.00**	80.75	72.79	71.81	75.23
Dynamic Pre-Judge Verification	**92.15**	**61.89**	84.12	**83.00**	**72.94**	**74.01**	**78.02**

As shown in Table 3, the configurations with no verification, static filtering, and dynamic pre-judge verification achieve average scores of 74.91, 75.23, and 78.02, respectively. Dynamic pre-judge verification obtains 92.15 on V* Bench, 61.89 on ZoomBench, 84.12 and 83.00 on HR-Bench-4K and HR-Bench-8K, and 72.94 and 74.01 on MME-RealWorld-EN and MME-RealWorld-CN.

Dynamic pre-judge verification exceeds static filtering by 2.79 points on average under the matched negative-teacher loss setup. We therefore adopt dynamic pre-judge verification, which reevaluates both teacher responses as training progresses rather than fixing the admissible samples before training.

4.3.3 Pre-Judge Verification Criterion

We next examine whether the magistrate should check one or both teacher responses. The positive-only and negative-only settings keep the negative-teacher loss and the remaining training setup fixed while changing which teacher responses undergo pre-judge verification. Checking only the positive-teacher response admits a sample when it matches the reference answer, whereas checking only the negative-teacher response admits it when that response does not match the reference answer. The both-teacher setting requires both conditions. The row without pre-judge verification reproduces the component-ablation configuration with neither component and serves only as a common reference, rather than an isolated comparison of verification placement.

Table 4: Ablation of pre-judge verification across teachers.

Pre-Judge Verification	V* Bench	ZoomBench	HR-4K	HR-8K	MME-EN	MME-CN	Average
None	89.53	58.93	82.12	77.88	69.54	69.95	74.66
Neg. Teacher Only	87.96	57.51	80.00	76.12	70.19	69.78	73.59
Pos. Teacher Only	91.62	57.16	83.75	80.63	**73.67**	71.65	76.41
Both Teachers	**92.15**	**61.89**	**84.12**	**83.00**	72.94	**74.01**	**78.02**

As shown in Table 4, the four configurations achieve average scores of 74.66 without pre-judge verification, 73.59 with pre-judge verification of only the negative teacher, 76.41 with pre-judge

verification of only the positive teacher, and 78.02 with pre-judge verification of both teachers. Pre-judge verification of both teachers obtains the highest average among the four configurations.

The both-teacher setting exceeds the positive-teacher-only setting by 1.61 points on average under the matched loss setup. The main method applies pre-judge verification to both teachers, while the single-teacher settings provide criterion comparisons.

5 Conclusion

In this paper, we propose MagistrateOPD to improve the reliability of visual on-policy distillation through pre-judge verification of both teacher responses before applying dense token-level supervision. We introduce dynamic pre-judge verification to admit samples with a correct positive-teacher response and an incorrect negative-teacher response while retaining the modality-balance trust region for token-level selection. We also introduce a negative-teacher loss that explicitly pushes the student away from the masked-image negative-teacher distribution alongside positive-teacher distillation. MagistrateOPD achieves an average score of 78.02 on the evaluation benchmarks. The component ablation shows that combining pre-judge verification and the negative-teacher loss outperforms either component alone, while dynamic pre-judge verification exceeds one-time static filtering by 2.79 points on average. Overall, these results support pre-judge verification and complementary positive- and negative-teacher signals as a way to strengthen fine-grained visual on-policy distillation.

6 AI USE STATEMENT

Generative AI tools were used to assist with three aspects of this work. First, GPT-5.6 helped evaluate the correctness of model answers on our free-form question benchmark, including judging answer equivalence, flagging unsupported or hallucinated spans, and assisting in qualitative error analysis. Second, GPT-6 assisted with the implementation of part of the experimental code, including scaffolding evaluation scripts and data-processing utilities; all AI-generated code was read, run, debugged, and verified by the authors before use. Third, generative AI tools (notably GPT-6) assisted with language editing and polishing of the manuscript, rephrasing unclear sentences, toning down overstatements, and adapting the paper to the ICLR format. We have reviewed all AI-assisted content—including evaluation judgments, code, and text—and take full responsibility for the scientific claims, experimental results, and final manuscript.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*, 2022. URL https://arxiv.org/abs/2204.14198.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*, 2023. URL https://arxiv.org/abs/2312.14238.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-Policy Distillation of Large Language Models. *arXiv preprint arXiv:2306.08543*, 2023. URL https://arxiv.org/abs/2306.08543.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. *arXiv preprint arXiv:1606.07947*, 2016. URL https://arxiv.org/abs/1606.07947.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. *arXiv preprint arXiv:2402.03898*, 2024. URL https://arxiv.org/abs/2402.03898.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *arXiv preprint arXiv:2311.16922*, 2023. URL https://arxiv.org/abs/2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. URL https://arxiv.org/abs/2301.12597.

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

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, et al. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744*, 2023a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023b. URL https://arxiv.org/abs/2304.08485.

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. URL https://arxiv.org/abs/2605.21924.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

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

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, et al. Aligning Large Multimodal Models with Factually Augmented RLHF. *arXiv preprint arXiv:2309.14525*, 2023. URL https://arxiv.org/abs/2309.14525.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

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*, 2026.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural Text Generation with Unlikelihood Training. *arXiv preprint arXiv:1908.04319*, 2019. URL https://arxiv.org/abs/1908.04319.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. doi: 10.1109/CVPR52733.2024.01243.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, et al. RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. URL https://arxiv.org/abs/2405.17220.

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.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. *arXiv preprint arXiv:2203.14465*, 2022. URL https://arxiv.org/abs/2203.14465.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

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. URL https://arxiv.org/abs/2505.14362.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026.

A Limitation and Future Work

In the current implementation, pre-judge verification uses an online MLLM API to judge both teacher responses at each training step. Because distillation waits for this response-level decision, API request and inference latency can substantially increase iteration time and leave the training GPUs idle between updates. In future work, we plan to use JEV in place of the online MLLM API to make these judgments quickly during training. Such a replacement should preserve reliable sample admission while reducing verification latency and improving GPU utilization.