TTRSD: TEST-TIME REINFORCEMENT LEARNING WITH SELF-DISTILLATION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time reinforcement learning enables vision-language models (VLMs) to adapt using unlabeled inputs. However, repeated sampling under fixed visual conditions can reinforce shared perceptual errors, while sequence-level rewards fail to isolate visual perception—the foundational bottleneck that anchors multimodal reasoning—risking the degradation of pre-trained reasoning capabilities. We propose TTRSD, a test-time reinforcement learning framework combining multi-view answer-level self-distillation with visual contrastive token selection. A shared policy aggregates teacher predictions across original, cropped, and downsampled views into an answer distribution. Student trajectories generated from the original image receive rewards based on the support for their final answers in this distribution. To allocate this feedback precisely toward perceptual bottlenecks, we compare the log-probabilities of the same sampled tokens under original and visually ablated inputs while holding their textual prefixes fixed, selecting visually sensitive positions for policy-gradient updates. TTRSD separates update direction, determined by group-relative advantages, from update position, determined by visual sensitivity, without requiring ground-truth labels, external verifiers, or a separate teacher. With only 20 unlabeled adaptation samples, TTRSD improves performance across seven benchmarks and three VLMs, raising InternVL3-2B's MMMU accuracy from 35.79% to 49.32% (+13.53%), demonstrating cross-dataset generalization while preserving inherent reasoning integrity.

1 INTRODUCTION

Vision-language models (VLMs) have made substantial progress in visual recognition, question answering, and multi-step reasoning (Zhu et al., 2025; Bai et al., 2025; Li et al., 2024). Adapting these models to target tasks in deployment settings without labeled data, however, remains challenging. Test-time reinforcement learning (TTRL) offers a promising approach: it repeatedly samples responses to unlabeled inputs, converts agreement among the model's own answers into rewards, and updates the policy accordingly. TTRL explores this paradigm for language reasoning (Zuo et al., 2026), TTRV extends it to VLMs (Singh et al., 2026), and TTPO further investigates test-time policy optimization with pseudo-label supervision (Wang et al., 2026a). These studies demonstrate the potential of learning from a model's own predictions, while highlighting the construction and use of supervision as central challenges in unlabeled adaptation.

Representative approaches construct pseudo-supervision through majority voting or answer frequencies, as illustrated in Figure 8(a). Their effectiveness depends on both the quality of this supervision and how it guides policy updates. Yet repeated sampling under the same visual condition may reproduce shared perceptual errors, allowing incorrect answers to receive strong support and become further reinforced. A central challenge is that answer agreement does not necessarily indicate correct visual understanding, a problem compounded by the fact that additional sampling provides limited diversity in the underlying visual evidence. On-policy self-distillation (OPSD), which derives supervision from a shared model operating under different information conditions, inspires us to broaden supervision sources (Zhao et al., 2026; Kaur et al., 2026; Yu et al., 2026). However, privileged teacher information, such as ground-truth answers, is generally unavailable during unlabeled

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.

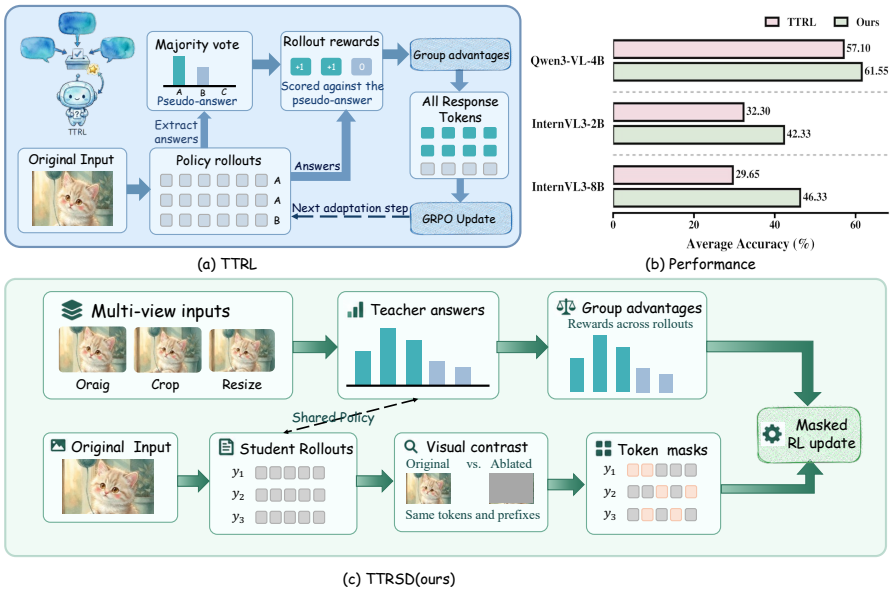


Figure 1: Overview and performance of TTRSD. (a) TTRL derives rewards from majority-vote pseudo-labels and applies sequence-level advantages to all response tokens. (b) TTRSD improves average accuracy over TTRL across seven benchmarks on three VLM backbones. (c) TTRSD uses a shared policy to construct rewards from multi-view teacher answers and selects visually sensitive tokens through original–ablated prediction comparisons, enabling targeted policy-gradient updates.

test-time adaptation. Motivated by this limitation, we construct an online teacher using different visual presentations of the same input. By aggregating predictions across views with different spatial coverage and resolutions, we obtain an answer-level self-distillation signal for reasoning over the original image.

Beyond the source of supervision, allocating answer-level feedback across reasoning tokens presents a second challenge. Multimodal reasoning interleaves visual evidence extraction and textual deductions, whose tokens differ in their dependence on visual input. Crucially, errors in visual perception can lead to reasoning errors, while accurate visual grounding supports subsequent reasoning and correct answers. Yet, GRPO broadcasts the same trajectory-level advantage to all response positions without accounting for these differences. In test-time adaptation, this feedback is estimated from the model's own predictions on limited unlabeled data. Since final-answer correctness does not necessarily reflect the validity of intermediate visual evidence, applying such feedback throughout the response may penalize valid grounding or reinforce hallucinated evidence. Furthermore, indiscriminately updating all tokens with sparse and noisy signals exacerbates the stability-plasticity dilemma, risking catastrophic forgetting of the model's robust, pre-trained textual reasoning capabilities. Prior work highlights the complementary roles of perception and reasoning (Li et al., 2026) and explore fine-grained optimization using visual dependency (Huang et al., 2026) or token-level teacher predictions (Agarwal et al., 2024; Zhao et al., 2026), but these methods typically rely on verifiable rewards or available on-policy teachers. These insights motivate a selective allocation of self-generated feedback during unlabeled adaptation. Inspired by RLSD's separation of update direction and token-level magnitude (Yang et al., 2026), we use group-relative advantages to determine update directions and visual sensitivity to select update positions. We identify these positions by comparing token log-probabilities under original and visually ablated inputs with identical textual prefixes, concentrating policy-gradient updates on predictions more strongly influenced by the image.

As illustrated in Figure 8(c), we propose TTRSD, a test-time reinforcement learning framework combining multi-view answer-level self-distillation with visual contrastive token selection. First, a shared policy serves as both teacher and student. We aggregate teacher predictions from the original, cropped, and downsampled images into an empirical answer distribution. Student trajectories generated from the original image then receive rewards based on voting support for their final

answers. Second, inspired by the conditional comparison in visual contrastive decoding (Leng et al., 2024), we compare the log-probabilities of the same sampled student tokens under original and visually ablated inputs while holding textual prefixes fixed. The absolute differences serve as a visual sensitivity score to select positions for policy-gradient updates. Ultimately, group-relative advantages determine the update direction, while visual sensitivity determines the targeted positions. Reference-policy regularization over all response tokens further constrains overall distribution drift. The entire procedure requires no ground-truth answers, external verifiers, or separate teacher models.

With only 20 unlabeled adaptation samples, TTRSD improves performance across seven benchmarks and three VLMs, consistently outperforming the TTRL baseline (8(b)). Maximum gains over zero-shot baselines reach 7.13% on MMMU for Qwen3-VL-4B, 14.14% on WeMath for InternVL3-2B, and 15.44% on MathVista for InternVL3-8B. Notably, the adapted 8B model reaches 71.84% on MathVista, exceeding the reported results of GPT-4o (71.60%) and Gemini-2.0-Flash (70.46%) (Achiam et al., 2023; Team et al., 2023). Furthermore, positive transfer across evaluated source–target pairs demonstrates strong cross-dataset generalization, while comprehensive ablations confirm the independent effectiveness of both multi-view supervision and visual token selection.

Our contributions are:

- We propose TTRSD, a test-time reinforcement learning framework that uses multi-view answer-level self-distillation to construct rewards without labels or a separate teacher.
- We introduce visual contrastive token selection to decouple update direction from update position: teacher answer support determines group-relative advantages, while visual sensitivity selects tokens for policy-gradient updates.
- TTRSD improves performance across diverse VLMs and benchmarks with limited unlabeled data and demonstrates cross-dataset generalization.

2 Related Work

Vision–language models (VLMs). Vision–language models broadly include dual-encoder models trained with contrastive learning for cross-modal retrieval and recognition (Radford et al., 2021; Jia et al., 2021), and large multimodal models (LMMs) that couple vision encoders with language models for open-ended perception and reasoning (Alayrac et al., 2022; Wang et al., 2026b; Wu et al., 2026). Recent open LMMs, such as Qwen3-VL (Bai et al., 2025) and InternVL3 (Zhu et al., 2025), have substantially improved OCR, spatial perception, long-context understanding, and multi-step visual reasoning. Nevertheless, their performance remains sensitive to visual distribution shifts, image quality, and task formats at deployment, while frozen models cannot directly exploit unlabeled test inputs to correct such errors. This motivates test-time training (TTT) (Sun et al., 2020), particularly test-time reinforcement learning (TTRL) (Zuo et al., 2026), which adapts pretrained models using self-generated supervision from test samples. Following this direction, our method constructs supervision from the same model's predictions across multiple visual views, without requiring test-set labels, external verifiers, or an independently trained teacher model.

Post-training for VLMs. Post-training improves language and multimodal models through instruction tuning and reward- or preference-based optimization, including RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), and GRPO (Shao et al., 2024). Multimodal RL methods refine credit assignment at the token level. VPPO (Huang et al., 2026) reshapes trajectory advantages and filters gradients according to visual dependency, while PGPO (Ye et al., 2026) and GCPO (Zhang et al., 2026) calibrate token updates using visual grounding, evidence relevance, or token criticality. A complementary direction studies on-policy distillation: OPD matches teacher predictions on student-generated prefixes (Agarwal et al., 2024), OPSD uses a shared model under asymmetric contexts (Zhao et al., 2026), and RLSD combines reward directions with token-level discrepancies between teacher and student predictions (Yang et al., 2026). In contrast, TTRSD uses unlabeled test inputs to construct multi-view answer supervision and direct updates toward visually sensitive tokens.

Test-time training. Recently, test-time training (TTT) has emerged as a paradigm that directly adapts model parameters using unlabeled deployment inputs (Sun et al., 2020). Early studies mainly focused on unimodal classifiers and addressed distribution shifts through self-supervised surrogate objectives such as entropy minimization (Wang et al., 2020). Subsequent work extended test-time adaptation to vision-language models, but primarily considered dual-encoder architectures through prompt

tuning, image-encoder adaptation, or embedding-space updates (Shu et al., 2022). More recently, TTRL, TTRV, and TTPO have used answer consensus across sampled responses to construct online rewards, extending test-time policy optimization to generative models (Zuo et al., 2026; Singh et al., 2026; Wang et al., 2026a). Building on this progression, we propose a view-asymmetric on-policy distillation method that constructs answer-level pseudo-supervision from transformed teacher views and optimizes original-view student trajectories through group-relative reinforcement learning.

3 METHOD

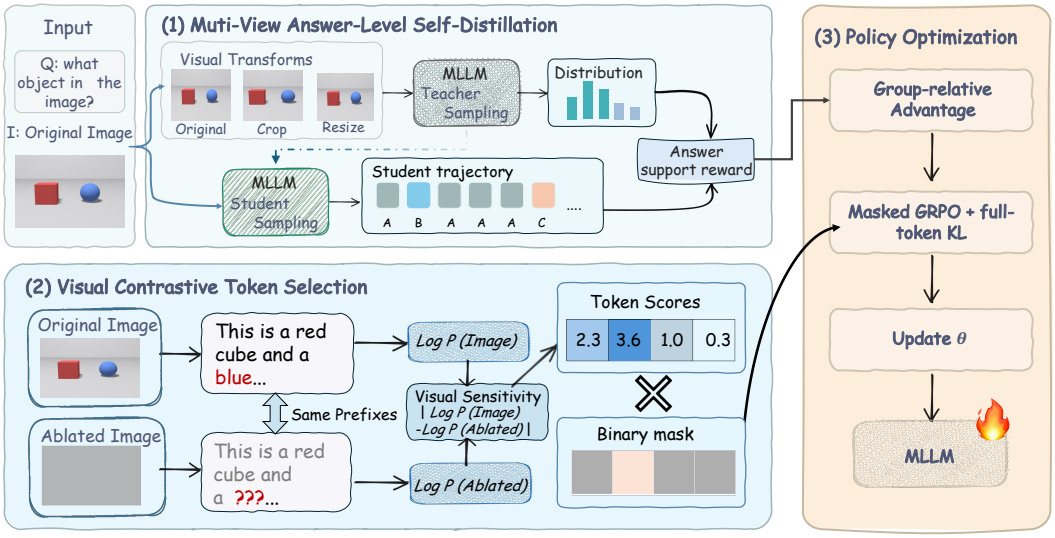


Figure 2: Overview of TTRSD. (1) Multi-view answer-level self-distillation aggregates teacher predictions into an answer distribution to reward original-view student trajectories. (2) Visual contrastive token selection compares the log-probabilities of the same student tokens under original and ablated images, selecting visually sensitive positions. (3) Policy optimization combines group-relative advantages with token masks for GRPO updates, while applying KL regularization over all valid response tokens. Teacher and student share the policy parameters throughout adaptation.

3.1 PROBLEM FORMULATION AND OVERVIEW

Let π_θ denote a vision-language model and let $\mathcal{D}_{\text{adapt}} = \{x_n\}_{n=1}^{N}$ be an unlabeled test-time adaptation set, where each input $x = (q, I)$ consists of a question q and an image I. We adapt π_θ using model-generated supervision, without ground-truth answers or a separately trained teacher for optimization. The shared policy serves as student and teacher under the following visual contexts:

$$x^S = [q; I], \qquad x^{T,v} = [q; g_v(I)], \quad v \in \mathcal{V}, \tag{1}$$

where g_v is a visual transformation, including the identity. The student generates reasoning trajectories from the original image, while the teacher generates responses across views. As shown in Fig. 2, TTRSD combines two mechanisms. ***Multi-view answer-level self-distillation*** aggregates teacher predictions into an empirical answer distribution and rewards student trajectories according to the support for their final answers. ***Visual contrastive token selection*** compares the likelihoods of the same student tokens under original and image-ablated inputs to identify visually sensitive positions. Together, answer-level supervision determines the group-relative update direction, while visual contrast determines the positions receiving the policy-gradient signal.

3.2 MULTI-VIEW ANSWER-LEVEL SELF-DISTILLATION

We construct answer-level supervision from the shared policy's predictions across visual views and use it to guide student trajectories generated from the original image.

Teacher-view construction. We consider three visual views, $\mathcal{V} = \{\text{orig}, \text{crop}, \text{down}\}$:

$$I^{(\text{orig})} = I, \qquad I^{(\text{crop})} = \text{Crop}_{0.85}(I), \qquad I^{(\text{down})} = \text{Resize}_{0.70}(I). \tag{2}$$

The center crop retains approximately 85% of each spatial dimension. Resizing scales both dimensions by 0.70, with each resulting dimension lower-bounded by 64 pixels. These transformations vary spatial coverage and resolution to provide alternative visual conditions for estimating answer support.

View-asymmetric sampling. At adaptation iteration k, we sample N_S student trajectories from the original-view context and K_T teacher trajectories from each teacher-view context:

$$\begin{aligned} y_i^S &\sim \pi_{\theta_k}(\cdot \mid x^S), \quad i = 1, \dots, N_S, \\ y_j^{T,v} &\sim \pi_{\theta_k}(\cdot \mid x^{T,v}), \quad j = 1, \dots, K_T, \quad v \in \mathcal{V}. \end{aligned} \tag{3}$$

Let $e(\cdot)$ extract the final answer from a completion, with $a_i^S = e(y_i^S)$ and $a_j^{T,v} = e(y_j^{T,v})$.

Answer-level supervision. The $N_T = |\mathcal{V}|K_T$ teacher answers define an empirical distribution:

$$\widehat{q}_{T,k}(a \mid x) = \frac{1}{|\mathcal{V}|K_T} \sum_{v \in \mathcal{V}} \sum_{j=1}^{K_T} \mathbb{I}[a_j^{T,v} = a]. \tag{4}$$

We reward each student trajectory according to the support for its final answer:

$$r_i^{\text{teach}} = \widehat{q}_{T,k}(a_i^S \mid x). \tag{5}$$

This reward preserves support for alternative answers rather than assigning binary rewards based on agreement with the modal teacher answer. It transfers multi-view teacher supervision to the original-view student through policy optimization, without direct token-distribution matching.

3.3 Visual Contrastive Token Selection

Answer-level rewards do not distinguish response tokens by their dependence on visual input. Inspired by visual contrastive decoding (Leng et al., 2024), we use prediction changes under visual ablation to select visually sensitive positions for policy-gradient updates.

Visual contrastive scoring. For a student trajectory $y = (y_1, \dots, y_L)$ sampled from x^S, we construct an ablated context $x_\varnothing^S = [q; I_\varnothing]$ by zeroing the processed pixel values while preserving the image-tile structure and image-token layout. Both forward passes score the same sampled tokens with identical textual prefixes; no additional trajectory is generated. We define token-level visual sensitivity as

$$\Delta_t = \left| \log \pi_\theta(y_t \mid x^S, y_{<t}) - \log \pi_\theta(y_t \mid x_\varnothing^S, y_{<t}) \right|. \tag{6}$$

Δ_t measures sensitivity to visual ablation in either direction, rather than token correctness. The ablated forward pass uses no gradients, and both log-probabilities are detached for mask construction. The original-image log-probabilities retain gradients for policy optimization.

Token-selective updates. Let L denote the number of valid response tokens and τ_ρ the $\lceil \rho L \rceil$-th largest score. The selection mask is

$$M_{\rho,t} = \mathbb{I}[\Delta_t \geq \tau_\rho], \qquad t = 1, \dots, L. \tag{7}$$

We apply this rule independently to each student trajectory with $\rho = 0.2$, retaining ties at the threshold. The mask selects positions for the policy-gradient objective, while the group-relative advantage determines the update direction. Reference-policy regularization remains active over all valid response tokens.

3.4 Policy Optimization

Following TTRV (Singh et al., 2026), we combine teacher support with a diversity-control reward:

$$r_i = r_i^{\text{teach}} - \lambda_H \bar{H}_{T,k}(x), \tag{8}$$

where $\bar{H}_{T,k}(x)$ is the normalized entropy of the empirical teacher answer distribution. And We set $\lambda_H = 0.75$ in all experiments. We first compute group-relative advantages across the N_S student trajectories for each input, then apply visual token selection to obtain token-level advantages:

$$A_i = \frac{r_i - \mathrm{mean}_j(r_j)}{\mathrm{std}_j(r_j) + \epsilon}, \qquad j = 1, \ldots, N_S. \tag{9}$$

We then apply the visual mask:

$$\widetilde{A}_{i,t} = M_{\rho,i,t} A_i. \tag{10}$$

Selected tokens retain the sequence-level advantage, while unselected tokens receive zero policy-gradient signal. Thus, answer support determines the relative update direction, and visual sensitivity determines the update positions. We incorporate $\widetilde{A}_{i,t}$ into the GRPO surrogate (Shao et al., 2024):

$$\mathcal{L}(\theta) = -\frac{\sum_{i,t} \ell_{\mathrm{clip}}\big(u_{i,t}(\theta), \widetilde{A}_{i,t}\big)}{\sum_{i,t} M_{\rho,i,t}} + \beta \mathcal{L}_{\mathrm{KL}}(\theta), \tag{11}$$

where $u_{i,t}$ is the current-to-old policy probability ratio and ℓ_{clip} is the clipped surrogate, implemented with dual clipping. Sums are taken over valid response positions in each optimization microbatch. $\mathcal{L}_{\mathrm{KL}}$ regularizes all valid response tokens to constrain overall distribution drift, independently of the visual mask, and we set $\beta = 0.001$. No auxiliary distillation loss is introduced.

4 Experiments

4.1 Setup

Implementation. We adapt InternVL3-2B, InternVL3-8B (Zhu et al., 2025), and Qwen3-VL-4B (Bai et al., 2025) on 20 unlabeled examples for eight epochs using GRPO, with a learning rate of 5×10^{-7}, a clipping parameter of 0.2, and a KL regularization coefficient of 10^{-3}. All methods, including TTRL and TTRV, use identical rollout and training configurations to ensure a fair comparison. Each rollout batch undergoes one epoch of policy optimization. Ground-truth answers are used only for evaluation and analysis.

Baselines. We consider three categories of baselines. For closed-source models, we include GPT-4o (Achiam et al., 2023)and Gemini-2.0-Flash (Team et al., 2023). For open-source models, we report MiniCPM-V 2.6 (Yao et al., 2024), InternLM-XComposer2-VL-7B (Dong et al., 2024), Phi-3.5-Vision (Haider et al., 2024), and LLaVA-1.5-7B (An et al., 2025). For test-time training, we compare against TTRL (Zuo et al., 2026), which constructs pseudo-rewards through majority voting, and TTRV (Singh et al., 2026), which employs answer-frequency rewards with diversity regularization. The Zero-shot baseline, TTRL, and our method all use chain-of-thought prompting, while TTRV is evaluated both with and without it. All adaptation methods use the same adaptation samples, training epochs, and total rollout budget.

Evaluation. We evaluate on held-out sets from seven benchmarks: WeMath (Qiao et al., 2025), LogicVista (Xiao et al., 2024), MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), MMMU (Yue et al., 2024), and MME-R (Yuan et al., 2025). Zero-shot and post-adaptation evaluations use identical prompts and greedy decoding for each model. We report accuracy (%), its unweighted mean across benchmarks, and gains over zero-shot in percentage points.

4.2 Main results

Table 1 presents the results of TTRSD on seven vision–language benchmarks covering visual mathematics, logical reasoning, and general multimodal understanding. TTRSD consistently improves over zero-shot performance across all benchmarks and backbones, increasing the average accuracy of Qwen3-VL-4B, InternVL3-2B, and InternVL3-8B by 4.14%, 9.96%, and 7.32%, respectively. For example, InternVL3-2B improves by 14.14% on WeMath and 13.95% on LogicVista, while InternVL3-8B improves from 56.40% to 71.84% on MathVista. Compared with TTRL and TTRV, TTRSD achieves the highest average accuracy at all three model scales, demonstrating consistent gains across tasks and model sizes.

Table 1: Main results on seven vision–language benchmarks. We report accuracy (%) and the seven-benchmark average (Avg.). The best result within each adaptation backbone is shown in **bold**. Missing published results are denoted by "–".

Backbone	Method	WeMath	LogicVista	MathVista	MathVerse	MathVision	MMMU	MME-R	Avg.
Closed-source Model									
GPT-4o	–	50.60	64.40	71.60	49.90	43.80	70.70	30.20	54.46
Gemini-2.0-Flash	–	47.42	–	70.46	43.65	47.82	69.30	–	–
Open-source Model									
MiniCPM-V 2.6	–	38.62	27.54	60.60	38.30	23.40	49.80	19.34	–
InternLM-XComposer2-VL-7B	–	12.67	–	57.60	25.90	14.54	43.00	–	–
Phi-3.5-Vision	–	11.20	25.10	43.90	36.02	14.90	43.09	–	–
LLaVA-1.5-7B	–	–	29.24	25.40	26.02	8.52	31.30	12.10	
Test-time Training Method									
Qwen3-VL-4B	TTRL (w/ CoT)	58.28	51.30	81.06	66.51	**43.39**	58.63	40.51	57.10
	TTRV (w/o CoT)	56.32	41.64	71.69	48.92	27.91	52.57	37.42	48.07
	TTRV (w/ CoT)	60.06	51.06	80.00	63.25	42.86	59.25	40.82	56.76
	Zero-shot (w/ CoT)	71.30	56.43	79.96	60.74	35.95	53.32	44.15	57.41
	TTRSD(w/ CoT)	**73.58**	**57.75**	**81.68**	**67.17**	40.01	**60.45**	**50.22**	**61.55**
	Δ	+2.28	+1.32	+1.72	+6.43	+4.06	+7.13	+6.07	+4.14
InternVL3-2B	TTRL (w/ CoT)	12.76	24.11	57.45	49.10	19.58	47.45	15.66	32.30
	TTRV (w/o CoT)	33.13	31.74	62.19	48.51	12.42	42.98	19.87	35.83
	TTRV (w/ CoT)	27.64	23.48	61.43	47.44	23.76	48.44	25.31	36.79
	Zero-shot (w/ CoT)	23.07	25.30	58.47	43.34	19.22	35.79	21.41	32.37
	TTRSD(w/ CoT)	**37.21**	**39.25**	**66.94**	**49.63**	**27.87**	**49.32**	**26.07**	**42.33**
	Δ	+14.14	+13.95	+8.47	+6.29	+8.65	+13.53	+4.66	+9.96
InternVL3-8B	TTRL (w/ CoT)	35.69	8.51	59.57	35.78	16.60	36.02	15.35	29.65
	TTRV (w/o CoT)	43.39	20.54	56.73	21.31	26.11	46.31	32.26	35.24
	TTRV (w/ CoT)	49.19	27.09	64.29	47.09	27.35	39.50	30.06	40.65
	Zero-shot (w/ CoT)	35.83	32.42	56.40	50.05	25.42	44.05	28.92	39.01
	TTRSD(w/ CoT)	**49.54**	**37.20**	**71.84**	**52.82**	**31.90**	**47.02**	**34.00**	**46.33**
	Δ	+13.71	+4.78	+15.44	+2.77	+6.48	+2.97	+5.08	+7.32

TTRSD also enables smaller models to compete with larger counterparts. With only 2B parameters, adapted InternVL3-2B achieves an average accuracy of 42.33%, surpassing both zero-shot InternVL3-8B (39.01%) and its best TTRV result (40.65%). Meanwhile, the 4B Qwen3-VL-4B outperforms the listed open-source comparison models wherever results are available across the seven benchmarks, and exceeds the reported closed-source results on WeMath, MathVista, MathVerse, and MME-R. Its average accuracy reaches 61.55%, compared with GPT-4o's 54.46%. These results highlight the potential of test-time adaptation to improve visual reasoning and make smaller models competitive across diverse tasks.

4.3 Ablation Study

Cross-Dataset Generalization. We further evaluate whether the benefits of TTRSD transfer across datasets. For each source-target pair, we adapt InternVL3-8B exclusively on source data, without using target inputs or labels during adaptation. We compare the adapted and zero-shot models on the same target test set using identical prompts and greedy decoding. As shown in Fig. 3, TTRSD improves accuracy across all three transfer settings, with gains of 3.83 percentage points on MathVerse→WeMath and 2.78 points on WeMath→MathVista. The largest gain occurs on WeMath→MMMU, where accuracy increases from 44.05% to 53.29%, a gain of 9.24 percentage points. These results show that the adaptation benefits extend beyond the source dataset, supporting the cross-dataset generalization of TTRSD.

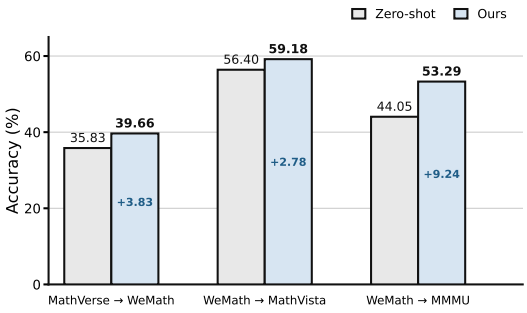


Figure 3: Cross-dataset generalization of TTRSD with InternVL3-8B. Bars compare zero-shot and source-adapted accuracy without target adaptation data. Arrows indicate transfer direction; gains are in percentage points.

Table 2: Ablation studies on InternVL3-2B. (a) Component ablation: MVSD denotes multi-view answer-level self-distillation, and VCTS denotes visual contrastive token selection. (b) Effect of the token selection ratio ρ on MathVista. All results are accuracy (%).

(a) Component ablation

MVSD	VCTS	MathVista	MathVision	LogicVista
✓	✓	**66.94**	**27.87**	**39.25**
–	✓	62.44	25.19	37.92
✓	–	61.78	27.02	31.15
–	–	61.43	23.76	23.48

(b) Token selection ratio

ρ	0.0	0.1	0.2	0.3	0.5
Acc.	58.47	63.02	**66.94**	63.43	58.68

Ablation Study on Components. We evaluate the contributions of multi-view answer-level self-distillation (MVSD) and visual contrastive token selection (VCTS) on MathVista, MathVision, and LogicVista using InternVL3-2B. Removing MVSD replaces multi-view teacher sampling with original-view sampling; removing VCTS applies the policy-gradient objective to all valid response tokens. We also evaluate the variant without either component, keeping the remaining training and evaluation settings fixed. As shown in Table 2(a), the full method achieves the highest accuracy on all three benchmarks. Removing MVSD reduces accuracy by 4.50, 2.68, and 1.33 percentage points, respectively, while removing VCTS leads to drops of 5.16, 0.85, and 8.10 points. Removing both components further degrades performance, with the largest drop of 15.77 points on LogicVista. These results support the effectiveness of both components and the benefit of combining them.

Effect of the token selection ratio. We vary the token selection ratio ρ on MathVista using InternVL3-2B. As shown in Table 2(b), accuracy increases from 63.02% at $\rho = 0.1$ to 66.94% at $\rho = 0.2$, then decreases to 63.43% and 58.68% at $\rho = 0.3$ and 0.5, respectively. Thus, $\rho = 0.2$ achieves the best performance among the evaluated ratios and is used as our default setting.

5 Analysis

Why transformed teacher views help. Figure 4 compares pseudo-label accuracy from individual teacher views and multi-view voting. On MathVision, multi-view voting achieves 22.2%, outperforming the original (12.4%), cropped (15.3%), and resized (12.1%) views. On LogicVista, it reaches 36.2%, compared with 33.9%, 32.1%, and 30.9%, respectively. Individual transformations do not consistently improve accuracy: cropping helps on MathVision, while both transformations underperform the original view on LogicVista. Nevertheless, aggregating their predictions outperforms every individual view on both benchmarks. This suggests that transformed views can provide complementary predictions even when they are less accurate in isolation, supporting multi-view aggregation for more accurate answer-level supervision.

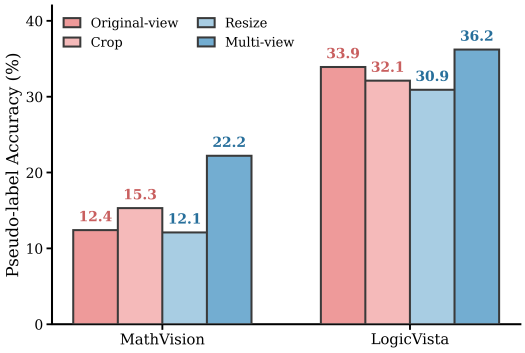


Figure 4: Multi-view teacher voting improves pseudo-label accuracy over individual views on both benchmarks.

Why visual credit assignment helps. Answer-level rewards reflect final prediction correctness but fail to differentiate tokens by their reliance on visual evidence. We evaluate whether visual sensitivity offers a principled criterion for targeted credit assignment. First, we verify the discriminative power of our selection rule. As shown in Table 3, at $\rho = 0.2$, the selected tokens exhibit mean absolute log-probability differences $9.23\times$–$19.11\times$ higher than unselected tokens across six benchmarks. This substantial separation confirms that our contrastive scoring effectively isolates steps directly grounded in visual context from general textual reasoning.

Table 3: Token-level visual sensitivity at $\rho = 0.2$. Values are mean absolute log-probability differences between original and visually ablated inputs.

Benchmark	All	Selected	Unselected
WeMath	0.1709	0.7012	0.0367
LogicVista	0.4283	1.5862	0.1262
MathVista	0.3825	1.4089	0.0974
MathVerse	0.3538	1.1875	0.1286
MathVision	0.2036	0.7295	0.0677
MMMU	0.2839	1.0563	0.0837

Table 4: Token-selection ablation on MathVista. Selection methods retain 20% of valid response tokens; other training settings are fixed.

Selection rule	Acc. (%)	Δ (pp)
Visual (ours)	**66.94**	**+5.16**
Random	62.45	+0.67
High entropy	61.43	-0.35
All tokens	61.78	0.00

Second, we examine whether this criterion improves adaptation beyond merely reducing the update budget. Table 4 compares visual selection against random, high-entropy, and all-token baselines on MathVista under an identical 20% token budget. Visual selection achieves 66.94% accuracy, outperforming random and high-entropy selection by 4.49 pp and 5.51 pp, respectively, and surpassing all-token optimization by 5.16 pp.

These comparisons highlight the importance of the token-selection criterion. **High-entropy selection** underperforms all-token optimization by 0.35 pp, indicating that prioritizing uncertain tokens does not improve adaptation in this setting. **Random selection** yields a modest gain of 0.67 pp, showing that reducing the number of updated positions can provide some benefit even without visual guidance. Visual selection achieves a larger gain of 5.16 pp and outperforms both alternatives under the same 20% token budget. These results support visual dependence as an effective criterion for selecting which response positions receive answer-level policy-gradient feedback.

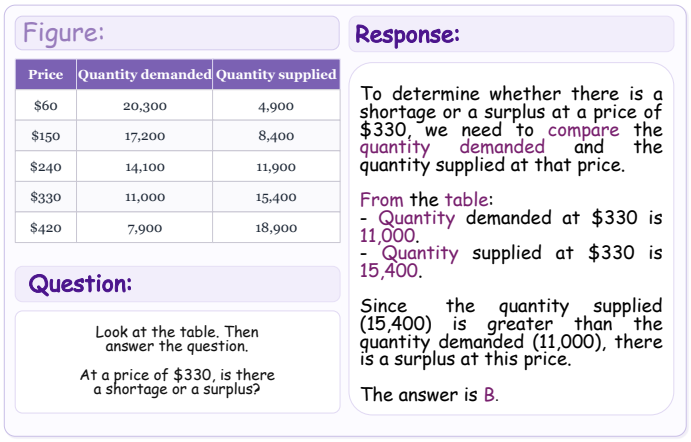


Figure 5: Visual token selection in a table-reading example. Purple tokens are selected for policy-gradient updates based on their sensitivity to visual ablation. More visualizations in Appendix.

Case study. Figure 5 visualizes token selection in a table-reading example, with selected tokens shown in purple. The selection captures tokens in the initial extraction of "11,000" and "15,400", together with descriptive tokens such as "quantity" and "table". Later repetitions of these values remain unselected, consistent with their availability in the textual prefix. This example illustrates how visual sensitivity distinguishes the same content at different positions in a reasoning trajectory.

6 Conclusion

We presented TTRSD, a test-time reinforcement learning framework that enables VLMs to improve from their own predictions without ground-truth labels, external verifiers, or a separate teacher. By combining multi-view answer-level self-distillation with visual contrastive token selection, TTRSD uses teacher answer support to determine update direction and visual sensitivity to select update positions. Experiments across seven benchmarks and three VLMs demonstrate consistent performance gains and effective cross-dataset generalization. This work suggests a path toward VLM self-evolution through multi-view self-supervision and selective policy updates.

Ethics Statement

Our research is conducted entirely within the domain of multimodal reasoning. We exclusively use publicly available academic benchmarks comprising mathematical, logical, and general visual reasoning problems. No new data were collected for this study, and no human subjects, personally identifiable information, or private user data were involved. The objective of this work is to improve the reasoning capabilities of vision–language models through adaptation on unlabeled data. Given its focus on academic benchmarks and abstract problem solving, we are not aware of any direct ethical concerns arising from the proposed methodology or experimental findings.

Reproducibility Statement

To facilitate reproducibility, we provide a detailed description of all experimental components in the paper and appendices. Section 3 presents the complete formulation of multi-view pseudo-label generation, trajectory rewarding, and visual token selection, including the image transformations and optimization objective. Section 4.1 describes the evaluated models, benchmarks, adaptation protocol, baselines, and evaluation settings. Appendix A and Table 6 provide the complete training configuration and key hyperparameters. Appendix B reports the multi-seed experiments, while Appendix C examines scalability with larger adaptation sets. The exact student, teacher, and evaluation prompt templates are provided in Appendix D. All datasets used in our experiments are publicly available.

AI Use Statement

Generative AI tools were used solely to assist with translating portions of the manuscript into English and improving linguistic clarity. All AI-assisted text was carefully reviewed and revised by the authors. These tools were not used to formulate the research questions, develop the methodology, design or conduct experiments, analyze or interpret the results, or draw scientific conclusions. All references were identified and verified by the authors. The authors take full responsibility for the content of this submission.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, Atabak Ashfaq, David Majercak, Wen Wen, Dongwoo Kim, Ziyi Yang, et al. Phi-3 safety post-training: Aligning language models with a" break-fix" cycle. *arXiv preprint arXiv:2407.13833*, 2024.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 103291–103325, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*, 2026.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogério Feris, Sivan Doveh, James Glass, and M Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33153–33163, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Aozhe Wang, Zhengxi Lu, Jianze Wang, Shangke Lv, Ying Liu, Weiming Lu, Jun Xiao, Yueting Zhuang, Hua Yang, Qianglong Chen, et al. Ttpo: Test-time policy optimization. *arXiv preprint arXiv:2608.27448*, 2026a.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

Shuning Wang, Zhiheng Wu, YiNuo Lu, Naiming Liu, Chen Jia, Bowen Liu, Shuo Nie, Weijie Zhu, and Yumeng Zhang. See more, think deeper: Query-expanded visual evidence and answer-clue guided reflection for long video understanding. *arXiv preprint arXiv:2606.09064*, 2026b.

Zhiheng Wu, Tong Wang, Shuning Wang, Naiming Liu, and Yumeng Zhang. See further, think deeper: Advancing vlm's reasoning ability with low-level visual cues and reflection. *arXiv preprint arXiv:2604.24339*, 2026.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *arXiv preprint arXiv:2604.01840*, 2026.

Xin Yu, Liuchen Liao, Yiwen Zhang, Yingchen Yu, Lingzhou Xue, and Qinzhen Guo. Preference-based self-distillation: Beyond kl matching via reward regularization. *arXiv preprint arXiv:2605.05040*, 2026.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, et al. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Guohui Zhang, Hu Yu, Xiaoxiao Ma, JingHao Zhang, Yaning Pan, Mingde Yao, Jie Xiao, Linjiang Huang, Jie Huang, and Feng Zhao. Group critical-token policy optimization for autoregressive image generation. In *International Conference on Learning Representations*, volume 2026, pp. 51697–51719, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 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.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

A Implementation Details

We provide the complete training configuration in Table 5. All experiments employ on-policy training with bfloat16 precision, FSDP, gradient checkpointing, and padding removal. We train for eight epochs using an actor learning rate of 5×10^{-7}, a PPO clipping range of 0.2, one PPO epoch per batch, and a KL-loss coefficient of 10^{-3}. During training, responses are sampled with a temperature of 1.0 and top-p of 1.0, whereas evaluation uses greedy decoding. We set the maximum prompt and completion lengths to 7,524 and 3,072 tokens, respectively. The hard visual mask retains the top 20% most visually sensitive completion tokens for policy-gradient updates, while KL regularization remains active over all valid completion tokens. No ground-truth reward, confidence filtering, or self-paced sample selection is used in the principal experiments.

Table 5: Training hyperparameters used in all experiments.

Hyperparameter	**Value**
Actor learning rate	5×10^{-7}
Training epochs	8
PPO epochs per batch	1
PPO clipping range	0.2
KL-loss coefficient	1×10^{-3}
Training batch size	32
Rollouts per input	16
Sampling temperature	1.0
Top-p	1.0
Maximum prompt length	7,524
Maximum response length	3,072
Numerical precision	bfloat16
Visual-token ratio ρ	0.2

B Multi-Seed Experiments

To evaluate the robustness of our method to stochastic variation, we conduct multi-seed experiments on five benchmarks using Qwen3-VL-4B. For each benchmark, we repeat the complete adaptation procedure with five independent random seeds. The adaptation examples, evaluation splits, prompts, and training hyperparameters are fixed across runs, while the random seed controls stochastic rollout generation, data ordering, and policy optimization. Table 6 reports the mean accuracy and standard deviation over the five runs.

The results demonstrate that TTRSD produces consistent performance across different random seeds. The standard deviation does not exceed 1.26 percentage points on any benchmark, and remains below one percentage point on four of the five benchmarks. Overall, the limited variation across independent runs indicates that the reported performance is robust to training stochasticity and is not driven by a particular favorable random seed.

Table 6: Multi-seed results of TTRSD with Qwen3-VL-4B. Accuracy (%) is reported as the mean and standard deviation over five independent random seeds.

Benchmark	**Accuracy (%)**
LogicVista	57.58 ± 0.17
MathVerse	67.01 ± 0.16
MathVision	39.64 ± 0.37
MathVista	80.93 ± 0.75
WeMath	72.96 ± 1.26

C Scaling to Larger Adaptation Sets

To further examine the scalability of our method, we investigate how its performance changes as the amount of unlabeled adaptation data increases. Specifically, we compare adaptation using 20 and 100 questions on WeMath, MMMU, and LogicVista with InternVL3-2B. All experiments follow the same multi-view construction, chain-of-thought prompting, pseudo-label generation, and visual token optimization procedures.

As shown in Table 7, increasing the adaptation set consistently improves performance across all three benchmarks. Accuracy increases from 37.21% to 38.14% on WeMath, from 49.32% to 50.21% on MMMU, and from 39.25% to 41.78% on LogicVista, corresponding to gains of 0.93, 0.89, and 2.53 percentage points, respectively. These results indicate that our method can effectively exploit additional unlabeled examples rather than relying on a small collection of favorable adaptation instances. A larger adaptation set exposes the model to a broader range of visual content and reasoning patterns, thereby providing more diverse self-supervision for policy optimization.

The performance improvement is accompanied by additional computational cost, since a larger adaptation set requires more multi-view rollout generation, pseudo-label estimation, and policy updates. Therefore, the number of adaptation examples introduces a practical trade-off between adaptation performance and computational efficiency. Overall, the consistent improvements observed across all three benchmarks demonstrate that multi-view pseudo-labeling and visual token selection remain effective as the amount of unlabeled adaptation data increases.

Table 7: Effect of scaling the unlabeled adaptation set from 20 to 100 questions with InternVL3-2B. Accuracy is reported in percent.

Benchmark	**20 Questions**	**100 Questions**	**Δ**
WeMath	37.21	38.14	0.93
MMMU	49.32	50.21	0.89
LogicVista	39.25	41.78	2.53

D Results on Perception-Oriented and Multi-Disciplinary Benchmarks

Beyond mathematical reasoning, we further evaluate our method on perception-oriented and multi-disciplinary benchmarks, including texture classification (DTD), multi-disciplinary perception (MMStar and SEED-Bench), and real-world scene understanding (RealWorldQA). As reported in Table 8, our method consistently improves over the zero-shot baseline on all of these benchmarks, with a particularly large gain of $+50.8$ points on DTD. These results indicate that the benefit of on-policy adaptation with our method is not confined to mathematical reasoning, but generalizes across diverse multimodal task categories.

Table 8: Results on perception-oriented and multi-disciplinary benchmarks with InternVL3-2B. Accuracy (%) is reported before and after adaptation.

Benchmark	**Zero-shot**	**Ours**	Δ
DTD	37.12	**87.94**	+50.8
MMStar	47.97	**51.11**	+3.1
SEED-Bench	69.85	**70.82**	+1.0
RealWorldQA	63.75	**64.57**	+0.8

E Computational Overhead Analysis

The only component our method adds over the TTRV baseline is one additional teacher-forced forward pass under the blank-image condition; no extra autoregressive rollout is required. We profile the wall-clock cost of each stage of a training step on a single NVIDIA A100 GPU with InternVL3-2B

Table 9: Wall-clock breakdown of one training step for a single sequence (InternVL3-2B, bfloat16, one NVIDIA A100 GPU; averaged over ten runs). The blank-image scoring pass is the only component added by our method.

Stage	Time	Share
Rollout generation (111 tokens, autoregressive)	13,170 ms	96.5%
Real-image scoring forward (no grad)	94.1 ms	0.7%
Blank-image scoring forward (no grad, ours)	94.2 ms	0.7%
Policy forward + backward	224.4 ms	1.6%
Optimizer update	63.4 ms	0.5%
Total	13,646 ms	100%

in bfloat16, using a representative sequence (a 919-token prompt covering three image tiles and a 111-token response), and report the average over ten runs in Table 9. The blank-image forward pass costs 94.2 ms, essentially identical to its real-image counterpart (94.1 ms), since the two differ only in the visual input. In practice, autoregressive rollout generation dominates the training step, accounting for over 96% of the end-to-end time, whereas the blank-image pass contributes less than 1%. The overhead introduced by visual sensitivity estimation is therefore negligible.

F Limitations and Future Work

Dependence on multi-view voting. Our method relies on multi-view responses to construct pseudo-labels. When all views produce correlated errors or no rollout yields the correct answer, the resulting supervision may be unreliable. Dynamically weighting views according to their agreement could mitigate this limitation.

Fixed visual token selection. We retain a fixed proportion of visually sensitive tokens based on the difference between real-image and blank-image likelihoods. However, the optimal proportion may vary across questions. Future work could adapt the selection ratio according to visual dependence and question difficulty.

Domain scope and efficiency. Our experiments focus on vision–language reasoning tasks with extractable answers. Extending the method to open-ended tasks may require semantic answer clustering or learned verification. Moreover, multi-view sampling and counterfactual forward passes introduce additional computation. Dynamically allocating views and rollouts according to pseudo-label confidence could improve adaptation efficiency.

G Prompt Template

We first present the general prompt template used for rollout generation and evaluation. The image, question, and answer choices provided by each benchmark are inserted into the corresponding placeholders. We then provide one representative prompt from each benchmark to illustrate the specific question and answer-choice formats.

G.1 General Prompt Template

Chain-of-Thought Prompt

```
<Image>
<Question>
<Answer Choices>
Reason concisely using the visual evidence and the choices.
Use enough steps to avoid guessing, but keep the reasoning
focused.  End with a separate final line exactly in the form:
The answer is X.
```

G.2 BENCHMARK-SPECIFIC EXAMPLES

WeMath Prompt

```
<Image>

<Question>
As shown in the diagram, there is a circular dining table
with an approximate diameter of 8 dm.  To cover this dining
table with a square tablecloth, the store offers three sizes
of tablecloths.  The side length of the tablecloth should be
at least (    ).

<Answer Choices>
A. 15 decimeters
B. 7 decimeters
C. 9 decimeters
D. No correct answer

Reason concisely using the visual evidence and the choices.
Use enough steps to avoid guessing, but keep the reasoning
focused.  End with a separate final line exactly in the form:
The answer is X.
```

LogicVista Prompt

```
<Image>

<Question>
The scale is balanced.  What weight is the weight with the
question mark?  Select from A, B, C, D, and E.

<Answer Choices>
A. 1 lb
B. 2 lb
C. 8 lb
D. 10 lb
E. 25 lb

Reason concisely using the visual evidence and the choices.
Use enough steps to avoid guessing, but keep the reasoning
focused.  End with a separate final line exactly in the form:
The answer is X.
```

MathVista Prompt

```
<Image>

<Question>
In the figure above, triangle ABC is inscribed in the circle
with center O and diameter AC. If AB = AO, what is the degree
measure of angle ABO?

<Answer Choices>
A. 15 degrees
B. 30 degrees
C. 45 degrees
D. 60 degrees
```

E. 90 degrees

Reason concisely using the visual evidence and the choices. Use enough steps to avoid guessing, but keep the reasoning focused. End with a separate final line exactly in the form: The answer is X.

MathVerse Prompt

<Image>

<Question>
Emile is observing a wind turbine. The vertical distance between the ground and the tip of one of the turbine's blades, in meters, is modeled by $H(t)$, where t is the time in seconds. What is the meaning of the highlighted segment?

<Answer Choices>
A. The turbine's center is 35 meters above the ground.
B. The turbine completes a single cycle in 35 seconds.
C. The length of the blade is 35 meters.
D. The turbine has 35 blades.

Reason concisely using the visual evidence and the choices. Use enough steps to avoid guessing, but keep the reasoning focused. End with a separate final line exactly in the form: The answer is X.

MathVision Prompt

<Image>

<Question>
$ABCD$ is a square, and M and N are the midpoints of BC and CD, respectively. Then $\sin\theta =$

<Answer Choices>
A. $\frac{\sqrt{5}}{5}$
B. $\frac{3}{5}$
C. $\frac{\sqrt{10}}{5}$
D. $\frac{4}{5}$
E. None of these

Reason concisely using the visual evidence and the choices. Use enough steps to avoid guessing, but keep the reasoning focused. End with a separate final line exactly in the form: The answer is X.

MMMU Prompt

<Image>

<Question>
A hash table of length 10 uses open addressing with the hash function $h(k) = k \bmod 10$ and linear probing. After inserting

six values into an empty hash table, the table is as shown in the image. Which option gives a possible order in which the key values could have been inserted?

<Answer Choices>
A. 46, 42, 34, 52, 23, 33
B. 34, 42, 23, 52, 33, 46
C. 46, 34, 42, 23, 52, 33
D. 42, 46, 33, 23, 34, 52

Reason concisely using the visual evidence and the choices. Use enough steps to avoid guessing, but keep the reasoning focused. End with a separate final line exactly in the form: The answer is X.

MME-Reasoning Prompt

<Image>

<Question>
As shown in the diagram, the flower color of a certain plant is controlled by two pairs of independently inherited alleles on autosomes: D, d and R, r. Which of the following statements is incorrect?

<Answer Choices>
A. There are four genotypes that can produce purple flowers.
B. When plants with genotype $DdRr$ are self-crossed, the proportion of purple-flowered offspring that can stably inherit the trait is $\frac{1}{6}$.
C. When a $Ddrr$ plant is crossed with a $ddRR$ plant, half of the offspring have blue flowers and half have purple flowers.
D. When a $DDrr$ plant is crossed with a $ddRr$ plant, half of the offspring have blue flowers and half have purple flowers.

Reason concisely using the visual evidence and the choices. Use enough steps to avoid guessing, but keep the reasoning focused. End with a separate final line exactly in the form: The answer is X.

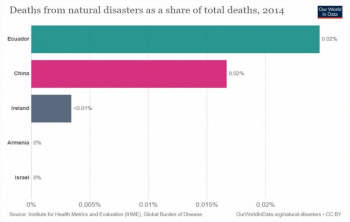
H Additional Case Studies

As a complement, we provide additional qualitative case studies, including positive cases and a failure case, for a more comprehensive analysis of our method. For each example, the same greedy response is re-scored with the real image and with a blank image, and the tokens in the top-20% of Δ are highlighted, visualizing how the selected tokens correspond to the visual evidence in the image. These examples also present the behavior of our method in both successful and failed cases, providing complementary insight for a comprehensive understanding of our method.

I Overall Adaptation Algorithm

We summarize the complete adaptation procedure of TTRSD through three components. Algorithm 1 constructs answer-level pseudo-supervision from multiple teacher views. Algorithm 2 identifies the response tokens whose probabilities are most sensitive to the visual input. Algorithm 3 combines these components to optimize original-view student trajectories using GRPO.

Figure:



Question:

Is the sum of two lowest bar is greater then the largest bar?

(A) Yes (B) No

Response:

To determine if the sum of the two lowest bars is greater than the largest bar, we need to add the values of the two lowest bars and compare it to the value of the largest bar. The two lowest bars are from Ireland and Armenia, both with a value of 0%.The largest bar is from Ecuador, with a value of 0.02%.

Adding the values of the two lowest bars: 0% (Ireland) + 0% (Armenia) = 0%

Comparing this sum to the value of the largest bar: 0% < 0.02% (Ecuador). Therefore, the sum of the two lowest bars is not greater than the largest bar.

The answer is B. ✅

Figure 6: **Case Study 1:** numeric extraction from a chart.

Figure:



Question:

How many seconds left to cross the street for a pedestrian?

A. 13 B. 11 C. 12 D. 14

Response:

The pedestrian signal shows a countdown timer indicating "11", which suggests there are left to cross the street. The other options are incorrect. because:- A. 13 is not visible on the signal. C. 12 is not visible on the signal. D. 14 is not visible on the signal.

The answer is B. ✅

Figure 7: **Case Study 2:** object counting.

Figure:



Question:

Is the orange larger than the car?

(A) Yes (B) No

Response:

To determine whether the orange is larger than the car, we need to compare their sizes.

The orange in the image appears to be quite large, almost as big as the cars in the background. The cars are smaller in comparison to the orange. Therefore, the answer is A. ❌

Figure 8: **Case Study 3 (failure):** a perceptual illusion.

Algorithm 1 Multi-View Teacher Distribution

Require: Example $x = (q, I)$, policy π_θ, teacher prompt p_T, samples per view K_T
Ensure: Teacher answer distribution $q_T(\cdot \mid x)$

1: Construct the teacher-view set

$$\mathcal{V}(I) \leftarrow \{I,\ \mathrm{Crop}_{0.85}(I),\ \mathrm{Resize}_{0.70}(I)\}$$

2: Initialize $\mathcal{Y}_T \leftarrow \varnothing$
3: **for all** $I^{(v)} \in \mathcal{V}(I)$ **do**
4: Sample teacher rollouts

$$\mathcal{Y}_T^{(v)} \leftarrow \left\{ y_{v,j}^T \sim \pi_\theta(\cdot \mid p_T, q, I^{(v)}) \right\}_{j=1}^{K_T}$$

5: $\mathcal{Y}_T \leftarrow \mathcal{Y}_T \cup \mathcal{Y}_T^{(v)}$
6: **end for**
7: Construct the teacher answer distribution

$$q_T(a \mid x) \leftarrow \frac{1}{|\mathcal{Y}_T|} \sum_{y \in \mathcal{Y}_T} \mathbb{I}[e(y) = a]$$

8: **return** $q_T(\cdot \mid x)$

Algorithm 2 Counterfactual Visual Token Selection

Require: Question q, image I, response y, policy π_θ, selection ratio ρ
Ensure: Visual-token mask M^ρ

1: Construct the blank-image counterfactual $I_\varnothing$ by zeroing the image pixels while preserving the visual-token layout
2: **for** $t = 1, \ldots, |y|$ **do**
3: Compute visual sensitivity

$$\Delta_t \leftarrow |\log \pi_\theta(y_t \mid q, I, y_{<t}) - \log \pi_\theta(y_t \mid q, I_\varnothing, y_{<t})|$$

4: **end for**
5: Select the largest $\lceil \rho|y| \rceil$ scores

$$M^\rho \leftarrow \mathrm{TopMask}\left(\{\Delta_t\}_{t=1}^{|y|}, \lceil \rho|y| \rceil \right)$$

6: **return** M^ρ

The teacher distribution is estimated exclusively from transformed teacher-view responses. Student responses therefore do not participate in constructing their own training targets, reducing the direct self-confirmation that arises from single-view majority voting.

The resulting mask measures visual dependence rather than token uncertainty. It is applied only to the policy-gradient component, whereas KL regularization remains active over the complete valid response sequence.

In the reported configuration, we use $N_S = 16$ student rollouts and $K_T = 16$ samples for each of the three teacher views, yielding 48 teacher votes and 64 total responses per question. We set $\rho = 0.2$, $\lambda_H = 0.75$, $\beta = 10^{-3}$, and use an actor learning rate of 5×10^{-7}. The model is optimized for eight epochs with a PPO clipping range of 0.2 and one PPO epoch per batch. Only original-view student trajectories receive policy-gradient updates, while teacher-view trajectories are used exclusively to construct pseudo-supervision.

Algorithm 3 TTRSD: Test-Time On-Policy Self-Distillation

Require: Unlabeled adaptation set $\mathcal{D}_{\text{adapt}}$; initial policy π_{θ_0}; student and teacher prompts p_S, p_T; rollout numbers N_S, K_T; coefficients ρ, λ_H, β

Ensure: Adapted policy π_θ

Initialize $\theta \leftarrow \theta_0$
for each adaptation epoch **do**
  for all $x = (q, I) \in \mathcal{D}_{\text{adapt}}$ **do**
    Set rollout policy $\pi_{\text{old}} \leftarrow \pi_\theta$
    Construct teacher supervision

$$q_T(\cdot \mid x) \leftarrow \textsc{TeacherDistribution}(x, \pi_{\text{old}}, p_T, K_T)$$

    Sample original-view student rollouts

$$\mathcal{Y}_S \leftarrow \left\{ y_i^S \sim \pi_{\text{old}}(\cdot \mid p_S, q, I) \right\}_{i=1}^{N_S}$$

    for all $y_i^S \in \mathcal{Y}_S$ **do**
      Assign trajectory reward

$$r_i \leftarrow q_T(e(y_i^S) \mid x) - \lambda_H \bar{H}(q_T)$$

      Construct visual-token mask

$$M_i^\rho \leftarrow \textsc{VisualTokenSelection}(q, I, y_i^S, \pi_\theta, \rho)$$

    end for
    Compute group-relative advantages

$$\{A_i\}_{i=1}^{N_S} \leftarrow \text{GRPOAdvantage}\left(\{r_i\}_{i=1}^{N_S}\right)$$

    Compute the visually masked objective

$$\mathcal{L}_{\text{TTRSD}} \leftarrow -\frac{\sum_{i,t} M_{i,t}^\rho \ell_{\text{clip}}(\theta; A_i)}{\sum_{i,t} M_{i,t}^\rho} + \beta \mathcal{L}_{\text{KL}}$$

    Update the policy

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{TTRSD}}$$

  end for
end for
return π_θ