VC-OPD: Visual Counterfactual On-Policy Distillation for Grounded Vision-Language Reasoning

Jiahe Feng
College of Intelligence and Computing
Tianjin University
Tianjin, China
fjh3475227047@tju.edu.cn

Abstract

GRPO-style reinforcement learning has advanced vision-language reasoning, but its sequence-level rewards provide limited credit assignment for individual reasoning tokens. On-policy distillation (OPD) addresses this issue by providing dense supervision on student-generated trajectories, yet standard OPD assigns similar importance to all teacher-preferred tokens, including template phrases and visually irrelevant continuations. We introduce Visual Counterfactual On-Policy Distillation (VC-OPD), a grounding-aware distillation method that prioritizes tokens whose correctness depends on image-specific evidence. After a teacher-supervised warm start that stabilizes response formats and reduces noisy rollouts, VC-OPD performs on-policy distillation with two teacher evaluations for each student trajectory: one under the original image and one under a counterfactual image with corrupted instance-specific evidence. The original-image teacher distribution determines the distillation target, while the original–counterfactual teacher gap estimates token-level visual dependence and softly reweights the loss. To avoid unstable response-level update scales, VC-OPD further applies mass-preserving normalization to the token weights. This design preserves the dense supervision of OPD while shifting learning capacity toward visually grounded reasoning tokens. Experiments on multimodal reasoning benchmarks show that VC-OPD improves visual reasoning performance and produces more effective token-level distillation.The corresponding code is available at https://github.com/feng-feng932/VC-OPD.

1 Introduction

Large vision-language models (VLMs) have achieved strong progress in multimodal reasoning, yet post-training them remains substantially harder than post-training text-only language models. Recent reinforcement and preference-based methods, including PPO, GRPO, multimodal RLHF, reinforcement fine-tuning, and verifiable-reward optimization, have improved mathematical and long-chain reasoning [Schulman et al., 2017, Shao et al., 2024, Sun et al., 2023, Yu et al., 2023, Liu et al., 2025a, Shen et al., 2025, Zhang et al., 2025a, Tan et al., 2025]. However, VLM reasoning has an additional prerequisite: the model must first identify and use the relevant visual evidence. This has motivated perception-aware post-training methods, such as corrupted-image objectives, perception-first curricula, self-rewarding decomposition, and grounded visual exploration [Wang et al., 2025a, Chen et al., 2025, Li et al., 2025, Sarch et al., 2025]. These studies indicate that effective VLM post-training should not only optimize final answers, but also assign learning signals to tokens that are genuinely grounded in the image.

On-policy distillation (OPD) is a natural framework for such fine-grained supervision. Instead of relying only on sequence- or step-level rewards, OPD samples trajectories from the student and asks

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

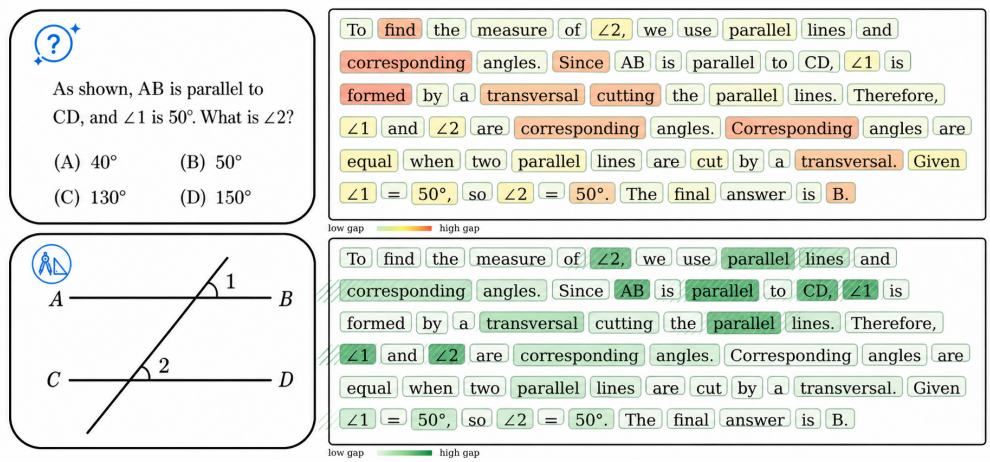


Figure 1: Token-level grounding visualization. Standard student–teacher gaps highlight both visual and generic tokens, whereas factual–counterfactual teacher gaps more selectively emphasize image-dependent tokens.

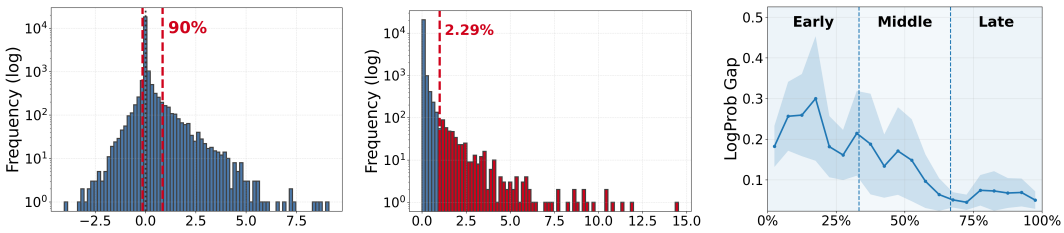
a teacher to score the same trajectories, providing dense token-level feedback under the student's own state distribution [Agarwal et al., 2024]. Recent theoretical analyses further interpret OPD as KL-regularized token-level policy optimization, where the teacher–student likelihood ratio acts as a dense reward-like signal [Yang et al., 2026, Ko et al., 2026]. Nevertheless, token-level density does not imply grounding awareness. Standard OPD treats all teacher-preferred tokens as equally distillable, although many tokens in VLM responses are generic templates, discourse markers, or language-prior continuations rather than image-dependent content.

This limitation is illustrated in Figure 1. The student–teacher gap used by standard OPD highlights both visually meaningful tokens and generic reasoning expressions, revealing where the student differs from the teacher but not whether the discrepancy is grounded in the image. In contrast, a factual–counterfactual teacher gap, obtained by comparing teacher scores under the original image and a visually weakened counterfactual image, more selectively emphasizes tokens tied to visual evidence, such as angle symbols, parallel-line relations, and answer-relevant quantities.

Token-level statistics provide the same message at scale. As shown in Figure 2, student–teacher gaps are heavily concentrated near zero, suggesting that dense OPD supervision often yields weak effective gradients. By contrast, large factual–counterfactual gaps appear on only a sparse subset of tokens, consistent with the structure of VLM reasoning: only some tokens encode objects, quantities, spatial relations, or diagram-specific information, while many others mainly serve as linguistic scaffolding. The position-wise pattern further shows that visual dependence is stronger in early perception-related positions and weaker in later reasoning or answer-formatting stages. These observations motivate a grounding-aware OPD objective that preserves dense supervision while reallocating update strength toward visually dependent tokens.

We propose *Visual Counterfactual On-Policy Distillation* (VC-OPD), a grounding-aware OPD framework for vision-language reasoning. VC-OPD decouples the *direction* and *magnitude* of token-level updates. For each student-generated trajectory, the factual teacher under the original image provides the teacher–student discrepancy that determines the update direction. The same teacher is then evaluated under a counterfactual image where instance-specific visual evidence is weakened. The resulting factual–counterfactual log-probability gap estimates each token's dependence on the original visual input and is converted into a soft grounding weight that modulates the update magnitude. In this way, VC-OPD retains the dense token-level supervision of OPD while concentrating learning on tokens that are more likely to encode visual perception and grounding.

VC-OPD uses a simple two-stage pipeline. First, we warm-start the student with teacher-generated responses to stabilize response format and reduce noisy on-policy rollouts. Second, we perform visual counterfactual on-policy distillation: the student rolls out under the real image, the teacher scores the same trajectory under both factual and counterfactual visual conditions, and the resulting visual-dependence score reweights the token-level OPD objective. We further apply mass-preserving



(a) Student–teacher gap (b) Factual–counterfactual gap (c) Position-wise visual gap

Figure 2: Token-level signal statistics. Standard OPD gaps are often small, while factual–counterfactual gaps reveal sparse and position-dependent visual grounding signals.

normalization to preserve the relative emphasis on visually grounded tokens while keeping the overall token weight stable across responses.

Our contributions are threefold:

- We identify a grounding-related limitation of standard OPD for VLM reasoning: the teacher–student discrepancy captures model disagreement but does not explicitly distinguish image-dependent tokens from generic linguistic continuations.
- We introduce the factual–counterfactual teacher log-probability gap as a token-level proxy for visual dependence, enabling grounding-aware reweighting of student-generated trajectories.
- We formulate VC-OPD with a direction–magnitude separation: the factual teacher–student discrepancy determines the update direction, while the counterfactual signal modulates its magnitude through soft weighting and mass-preserving normalization.

2 Related Work

2.1 Reinforcement learning for vision-language models

Reinforcement learning has become an important post-training paradigm for vision-language models (VLMs), building on the success of policy optimization in language reasoning. Modern VLM RL inherits trust-region and proximal policy methods such as TRPO and PPO, as well as critic-free variants such as GRPO, which has shown strong performance in mathematical reasoning [Schulman et al., 2015, 2017, Shao et al., 2024]. These optimization paradigms have been extended to multimodal alignment and reasoning through RLHF, reward modeling, preference alignment, and reinforcement fine-tuning [Sun et al., 2023, Yu et al., 2023, Zhang et al., 2025b, Ji et al., 2025, Liu et al., 2025a, Shen et al., 2025, Zhang et al., 2025a, Tan et al., 2025, Liu et al., 2025b].

A growing line of work further argues that visual perception is a central bottleneck in VLM RL. PAPO introduces perception-aware policy optimization by comparing model behavior under original and corrupted visual inputs; PeBR proposes a two-stage curriculum that improves perception before reasoning; Vision-SR1 decomposes visual reasoning into perception and language reasoning with self-rewarding signals; ViGoRL grounds intermediate reasoning steps to visual coordinates [Wang et al., 2025a, Chen et al., 2025, Li et al., 2025, Sarch et al., 2025]. These methods show that successful multimodal reasoning requires models to use instance-specific visual evidence rather than relying on language priors or generic reasoning templates. However, most existing VLM RL methods still optimize sequence-level rewards, step-level rewards, or auxiliary perception losses, and do not provide dense teacher-guided token-level credit assignment. Our method addresses this gap by introducing visual counterfactual signals into on-policy distillation.

2.2 On-policy distillation for reasoning models

Knowledge distillation transfers capabilities from large teachers to smaller students, but classical distillation and supervised fine-tuning usually optimize the student on fixed teacher-provided trajectories, which can introduce distribution mismatch [Hinton et al., 2015, Kim and Rush, 2016].

On-policy distillation (OPD) reduces this mismatch by sampling trajectories from the student and querying the teacher on those trajectories, thereby aligning supervision with the student's own rollout distribution [Agarwal et al., 2024, Yang et al., 2026, Ko et al., 2026].

Recent work has provided a stronger theoretical and practical foundation for OPD. GKD formalizes on-policy distillation for autoregressive models [Agarwal et al., 2024]. G-OPD interprets OPD as dense KL-constrained policy optimization and generalizes the reference-policy design [Yang et al., 2026]. REOPOLD further shows that the teacher–student likelihood ratio can be viewed as a token-level reward and improves training efficiency through relaxed token selection and reward clipping [Ko et al., 2026]. Other OPD variants study context distillation, reasoning-prefix supervision, and entropy-aware objectives [Ye et al., 2026, Zhang et al., 2026, Jin et al., 2026]. Compared with sparse reward RL, OPD provides dense token-level supervision, which is attractive for long-chain reasoning.

While effective for text reasoning, existing OPD methods typically treat all generated tokens as distillable. This assumption is suboptimal for VLM reasoning, where many tokens are structural markers, template phrases, or language-prior continuations, and only a subset is truly supported by visual evidence. We extend the generalized OPD perspective to multimodal reasoning by using a counterfactual teacher to estimate token-level visual dependence and reweighting OPD updates toward visually grounded tokens.

3 Method

3.1 Preliminaries: on-policy distillation as token-level RL

We consider a vision-language reasoning setting with textual input x, visual input v, and autoregressive output sequence $y = (y_1, \ldots, y_T)$. The student policy is denoted by $\pi_\theta(y_t \mid x, v, y_{<t})$, and the rollout policy μ is typically the current student or a lagged copy. During training, we sample trajectories on-policy under the real visual input:

$$y \sim \mu(\cdot \mid x, v). \tag{1}$$

On-policy distillation (OPD) differs from conventional offline distillation in one essential aspect: the teacher does not supervise teacher-forced trajectories, but instead evaluates trajectories that are actually generated by the student itself. As a result, the distillation signal is aligned with the student's own state distribution, which reduces the train–test mismatch that commonly arises in offline knowledge distillation. This property is particularly important for reasoning tasks, where small deviations in early tokens may lead the student to entirely different downstream trajectories.

Recent work has shown that OPD can be interpreted through the lens of token-level policy optimization. Let $s_t = (x, v, y_{<t})$ denote the autoregressive state and $a_t = y_t$ the sampled token. If we measure the discrepancy between the student and the teacher with the reverse KL divergence,

$$D_{\mathrm{RKL}}(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_{\mathrm{tea}}(\cdot \mid s_t)) = \mathbb{E}_{a_t \sim \pi_\theta}\left[\log \pi_\theta(a_t \mid s_t) - \log \pi_{\mathrm{tea}}(a_t \mid s_t)\right], \tag{2}$$

then maximizing the negative reverse KL induces the token-level learning signal

$$r_t = \log \pi_{\mathrm{tea}}(a_t \mid s_t) - \log \pi_\theta(a_t \mid s_t). \tag{3}$$

This quantity can be viewed as a dense per-token reward: if the teacher assigns higher probability to the sampled token than the student, the reward is positive, encouraging the student to increase its probability; otherwise the reward becomes small or negative. In this sense, OPD can be understood as a form of dense RL where the reward is not provided by an external environment, but is instead induced by the discrepancy between the teacher and the student on student-generated trajectories.

Under this view, a broad class of OPD-style methods can be written as weighted log-likelihood optimization:

$$\mathcal{L}(\theta) = -\mathbb{E}_{y \sim \mu}\left[\sum_{t=1}^{T} \mathrm{sg}(A_t) \log \pi_\theta(y_t \mid x, v, y_{<t})\right], \tag{4}$$

where A_t is an advantage-like token weight and $\mathrm{sg}(\cdot)$ denotes stop-gradient. Therefore, once OPD is viewed as token-level RL, the central modeling question becomes how to define A_t.

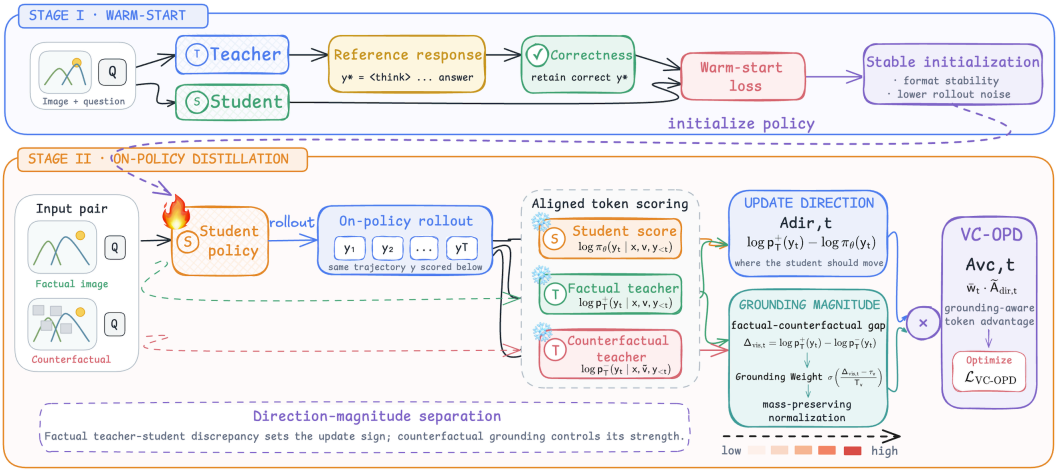


Figure 3: Overview of VC-OPD. The framework first warm-starts the student with filtered teacher-generated reasoning responses. During visual counterfactual on-policy distillation, student rollouts are scored by the teacher under factual and counterfactual visual inputs. The factual teacher provides the update direction, while the factual–counterfactual gap produces grounding-aware token weights that modulate the final token-level advantage.

For text-only reasoning, this weight mainly reflects how strongly the teacher prefers a continuation. In multimodal reasoning, however, teacher preference alone is insufficient: a token may be preferred either because it is genuinely grounded in visual evidence or because it is a generic reasoning phrase, a structural template token, or a continuation supported only by language priors. Our goal is therefore to redesign the token-level weight so that student updates are concentrated on tokens that are truly supported by visual evidence.

3.2 Two-stage training

Our training consists of two stages. The first stage provides a stable initialization, and the second stage performs visual counterfactual on-policy distillation.

Stage I: warm-start. Before entering on-policy training, we first warm-start the student using teacher-generated reference responses $y^\star$. Given (x, v), the teacher produces a structured target response, and the student is optimized with standard autoregressive supervision:

$$\mathcal{L}_{\text{warm}}(\theta) = -\mathbb{E}_{(x,v,y^\star)\sim\mathcal{D}}\left[\sum_{t=1}^{T^\star} \log \pi_\theta(y_t^\star \mid x, v, y_{<t}^\star)\right]. \tag{5}$$

The purpose of this stage is not to solve the final problem by imitation alone, but to place the student in a reasonable response regime before on-policy optimization. In practice, this reduces rollout noise, stabilizes response format, and makes the subsequent teacher feedback more informative.

Stage II: visual counterfactual on-policy distillation. After warm-start, the student generates its own response trajectory under the real image,

$$y \sim \mu(\cdot \mid x, v). \tag{6}$$

We then evaluate the same trajectory with the teacher under two visual conditions: the original image and a counterfactual image in which instance-specific evidence has been weakened. The factual branch determines *where* the student should move, while the counterfactual branch determines *how strongly* each token should be updated.

3.3 Counterfactual teacher scoring and grounding-aware weighting

For each sampled token y_t, we define the factual teacher score and the counterfactual teacher score as

$$p_T^+(y_t) = p_T(y_t \mid x, v, y_{<t}), \qquad p_T^-(y_t) = p_T(y_t \mid x, \tilde{v}, y_{<t}), \tag{7}$$

where $\tilde{v} = c(v)$ is a corrupted visual input produced by a visual perturbation operator $c(\cdot)$. The role of $c(\cdot)$ is not to destroy the entire input, but to weaken instance-level visual evidence while preserving the general input format, so that the teacher remains in a comparable prediction regime.

Based on these two teacher evaluations, we compute two quantities. The first is a direction term,

$$A_t^{\mathrm{dir}} = \log p_T^+(y_t) - \log \mu(y_t \mid x, v, y_{<t}), \tag{8}$$

which compares the factual teacher and the student on the sampled token. This term determines the basic update direction: when the student underestimates a token relative to the factual teacher, the token receives a positive update signal; when the student already overestimates it, the update is weakened or reversed.

The second is a visual dependence score,

$$\Delta_t^{\mathrm{vis}} = \log p_T^+(y_t) - \log p_T^-(y_t), \tag{9}$$

This metric quantifies the shift in teacher token support under visual perturbation. We use this difference as a proxy for token-level visual dependence, not a rigorous causal estimate of image effects. A positive value means the original factual image yields stronger teacher support for the token, while near-zero values show insensitivity to visual perturbations. The score hinges on whether the perturbation removes task-relevant visual evidence while preserving sufficient context.

Rather than hard-selecting tokens with a binary rule, we convert the visual dependence score into a continuous soft weight:

$$w_t = \sigma\left(\frac{\Delta_t^{\mathrm{vis}} - \tau_v}{T_v}\right), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function, τ_v is a visual threshold, and T_v is a temperature parameter. This design preserves continuity during training: highly grounded tokens receive larger weights, while weakly grounded tokens are downweighted but not necessarily discarded.

A practical issue is that the number and strength of visually grounded tokens may vary substantially across samples, which can lead to unstable update scales if w_t is used directly. To stabilize optimization, we adopt a mass-preserving normalization:

$$\bar{w}_t = w_t \cdot \frac{N}{\sum_{j=1}^{T} w_j + \epsilon}, \qquad N = \sum_{t=1}^{T} m_t, \tag{11}$$

where m_t is the response mask and N is the number of valid response tokens. This normalization preserves the relative emphasis across tokens while keeping the total effective weight approximately constant for each response, thereby preventing large variations in update magnitude across batches.

3.4 Final VC-OPD objective

We combine the direction term and the normalized grounding weight to define the final token-level advantage:

$$A_t^{\mathrm{VC}} = \bar{w}_t \cdot \tilde{A}_t^{\mathrm{dir}}, \tag{12}$$

where $\tilde{A}_t^{\mathrm{dir}}$ denotes a normalized version of A_t^{dir} for numerical stability. The key design principle is that the factual teacher–student discrepancy determines the direction of the update, while the counterfactual branch modulates only its magnitude according to visual grounding strength.

The final stage-II objective is

$$\mathcal{L}_{\text{VC-OPD}}(\theta) = -\mathbb{E}_{y \sim \mu(\cdot|x,v)}\left[\sum_{t=1}^{T} \mathrm{sg}\big(A_t^{\mathrm{VC}}\big) \log \pi_\theta(y_t \mid x, v, y_{<t})\right]. \tag{13}$$

This objective retains the on-policy nature of OPD while making the token-level update explicitly grounding-aware. Tokens that are both under-modeled by the student and strongly supported by real visual evidence receive larger updates; tokens that are mainly driven by language priors are correspondingly suppressed.

Table 1: Main results on multimodal reasoning benchmarks. We report exact-match accuracy under the @8 setting with inference temperature 1.0. Avg. denotes the unweighted average over all listed benchmarks.

Model	Geo3K	MVer.	MVist.	WeM.	MMMU	Logic	Count	Avg.
Qwen2.5-VL-3B-Instruct	24.96	38.08	58.50	53.22	24.50	34.72	51.71	40.81
7B teacher → 3B student								
+ GRPO	30.52	40.60	59.04	59.10	25.86	38.34	57.81	44.47
+ PAPO	30.95	41.50	62.38	60.09	27.11	38.67	62.56	46.18
+ RKL	32.12	40.75	62.08	60.23	28.40	39.41	63.50	46.64
+ VC-OPD	**34.35**	**41.90**	**64.10**	**61.46**	**32.20**	**42.40**	**65.68**	**48.87**
Δ_{RKL} (%)	+6.94	+2.82	+3.25	+2.04	+13.38	+7.59	+3.43	+4.78
32B teacher → 3B student								
+ RKL	45.09	43.25	65.50	64.56	31.24	41.52	65.80	50.99
+ VC-OPD	**57.58**	**44.40**	**65.72**	**65.74**	**33.50**	**42.90**	**67.24**	**53.87**
Δ_{RKL} (%)	+27.70	+2.66	+0.34	+1.83	+7.23	+3.32	+2.19	+5.64
Qwen2.5-VL-7B-Instruct	36.80	43.90	66.40	64.80	33.80	43.50	64.90	50.59
32B teacher → 7B student								
+ RKL	48.70	47.40	69.70	66.30	36.20	46.00	69.00	54.76
+ VC-OPD	**55.40**	**48.90**	**70.90**	**68.10**	**38.80**	**48.60**	**71.20**	**57.41**
Δ_{RKL} (%)	+13.76	+3.17	+1.72	+2.71	+7.18	+5.65	+3.19	+4.85

4 Experiments

4.1 Experimental setup

We use Qwen2.5-VL [Bai et al., 2025a] as the primary backbone for our main experiments. Unless otherwise specified, Qwen2.5-VL-3B-Instruct is the student, while larger Qwen2.5-VL models serve as teachers. All methods start from the same student checkpoint and are trained on ViRL39K [Wang et al., 2025b], which contains approximately 39K verifiable visual-language reasoning examples. VC-OPD uses a two-stage procedure consisting of teacher-supervised warm-start and visual counterfactual on-policy distillation. We further evaluate Qwen3-VL and InternVL3.5 to test cross-backbone generalization, with additional details provided in Appendix E and F.

We evaluate seven multimodal reasoning benchmarks covering mathematical, geometry, multidisciplinary, logical, and counting tasks: Geometry3K [Lu et al., 2021], MathVista [Lu et al., 2024], MathVerse [Zhang et al., 2024], We-Math [Qiao et al., 2024], MMMU-Pro [Yue et al., 2024], LogicVista [Xiao et al., 2024], and Super-CLEVR Counting [Li et al., 2023]. All results use exact match after answer extraction, with the average reported as the unweighted mean across benchmarks.

Unless otherwise specified, we use the @8 setting with inference temperature 1.0. All experiments are conducted on four NVIDIA A800 GPUs. Detailed hyperparameters and implementation details are provided in Appendix B.

4.2 Main results

Table 1 reports the main results on multimodal reasoning benchmarks. We adopt Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct as the student models, with Qwen2.5-VL-7B and Qwen2.5-VL-32B serving as teacher models. For the Qwen2.5-VL-3B student, we evaluate both 7B- and 32B-teacher settings, while the Qwen2.5-VL-7B student is distilled from the 32B teacher. In the 7B-teacher setting with the 3B student, we compare VC-OPD with GRPO, PAPO, and reverse-KL on-policy distillation (RKL). For all distillation settings, RKL serves as the direct on-policy distillation baseline, and Δ_{RKL} denotes the relative improvement of VC-OPD over RKL.

Table 2: Ablation study under the 7B-teacher setting. Warm, CF, Soft, and Mass denote warm-start, counterfactual teacher scoring, soft visual reweighting, and mass-preserving normalization, respectively.

Variant	Warm	CF	Soft	Mass	Math Avg.	Vision Avg.	Overall Avg.
Vanilla OPD	–	–	–	–	48.80	43.77	46.64
w/o warm-start	–	✓	✓	✓	49.37	45.55	47.73
w/o counterfactual teacher	✓	–	–	✓	49.42	44.63	47.37
w/ hard mask	✓	✓	–	✓	49.86	45.92	48.17
w/o mass norm.	✓	✓	✓	–	50.05	46.10	48.36
VC-OPD	✓	✓	✓	✓	**50.45**	**46.76**	**48.87**

VC-OPD consistently improves over RKL across all evaluated student–teacher settings. With the 7B teacher and the Qwen2.5-VL-3B student, VC-OPD improves the average accuracy from 46.64 to 48.87, yielding a 2.23-point absolute gain and a 4.78% relative improvement. With the 32B teacher and the same 3B student, VC-OPD further improves the average accuracy from 50.99 to 53.87, corresponding to a 2.87-point absolute gain and a 5.64% relative improvement. For the Qwen2.5-VL-7B student, the base model achieves an average accuracy of 50.59, while distillation from the 32B teacher improves the RKL performance from 54.76 to 57.41, corresponding to a 4.85% relative improvement. The consistent gains across different student and teacher scales demonstrate that VC-OPD remains effective as the student model capacity increases, while preserving the advantages of grounding-aware token-level distillation.

4.3 Ablation study

We conduct ablation studies under the 7B-teacher setting to isolate the effect of each component in VC-OPD. Starting from vanilla on-policy distillation, we progressively examine the contribution of warm-start training, counterfactual teacher scoring, soft visual reweighting, and mass-preserving normalization.

Table 2 shows that each component contributes to the final performance of VC-OPD. Vanilla OPD already provides dense token-level supervision, but its performance remains limited because all generated tokens are treated uniformly. Adding counterfactual visual scoring improves both math-oriented and visually grounded benchmarks, indicating that token-level visual dependence is a useful signal for multimodal distillation.

Warm-start training also matters: removing it lowers the overall average from 48.87 to 47.73, suggesting that stable initial rollouts are important before applying on-policy teacher feedback. Replacing soft reweighting with a hard mask further degrades performance, showing that continuous token weighting is preferable to binary token selection. Finally, removing mass-preserving normalization reduces the overall average to 48.36, confirming that stabilizing the effective token weight is beneficial for robust optimization. Overall, the full VC-OPD configuration achieves the best performance, supporting the design choice of combining counterfactual visual dependence with stable soft token-level reweighting.

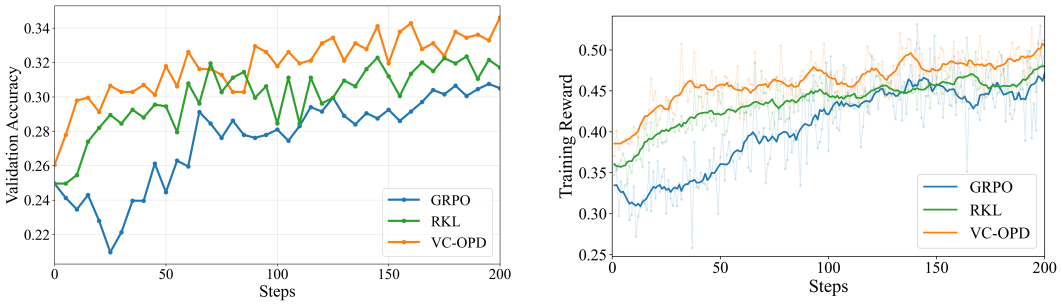
4.4 Training dynamics

We further compare the optimization behavior of GRPO, RKL, and VC-OPD under the same 200-step training budget. As shown in Table 3 and Figure 4, both RKL and VC-OPD improve faster than GRPO, suggesting that dense teacher-guided token supervision is more sample-efficient than outcome-level reward optimization. VC-OPD further achieves a more stable trajectory than RKL and reaches the best final performance on both Geo3K and LogicVista, indicating that visual counterfactual reweighting improves the effectiveness and stability of token-level distillation.

Compared with GRPO, RKL improves more rapidly because it provides dense token-level teacher feedback instead of relying only on sequence-level rewards. However, RKL still applies supervision uniformly across tokens, which may amplify visually irrelevant or template-like continuations. VC-OPD addresses this issue by reweighting token updates according to visual counterfactual de-

Table 3: Step-wise validation accuracy during training. We report exact-match accuracy (%) on Geo3K and LogicVista every 50 optimization steps.

Method	Geo3K				LogicVista			
	50	**100**	**150**	**200**	**50**	**100**	**150**	**200**
GRPO	26.10	27.75	28.90	30.52	34.20	36.30	37.45	38.34
RKL	29.20	30.35	31.25	32.12	37.05	38.20	38.85	39.41
VC-OPD	**31.80**	**32.45**	**33.20**	**34.35**	**39.10**	**40.55**	**41.45**	**42.40**



(a) Geo3K validation accuracy

(b) Training reward

Figure 4: Training dynamics over 200 optimization steps. VC-OPD improves faster than GRPO, is more stable than RKL, and achieves the highest validation accuracy and training reward.

pendence. As a result, it maintains a smoother reward curve and consistently outperforms RKL at later training stages.

4.5 Hyperparameter analysis

We analyze the proportion of visually dependent tokens emphasized by the soft visual weighting function. The goal is to verify that VC-OPD benefits from a balanced grounding signal rather than overly dense or overly sparse token reweighting.

Table 4: Sensitivity analysis of the visual token ratio. Default setting is marked with *.

Visual ratio	Math	Vision	Overall
20%	49.92	46.18	48.32
10%*	**50.45**	**46.76**	**48.87**
5%	49.58	45.90	48.00

Table 4 shows that a moderate visual-token ratio performs best. Since visually grounded tokens form only a sparse subset of the response, using a large ratio such as 20% assigns substantial weight to weakly visual or template-like tokens, reducing the selectivity of VC-OPD. In contrast, using only 5% of tokens makes the supervision too sparse, resulting in insufficient effective updates. The 10% setting provides a favorable trade-off between grounding selectivity and optimization density and is therefore used as the default configuration.

4.6 Alternative Counterfactual Constructions

We further investigate whether VC-OPD depends on a particular counterfactual construction. The counterfactual image is intended to weaken instance-specific visual evidence while preserving a meaningful visual context. We compare the default patch dropout with random dropout, Gaussian blur, contiguous region masking, object-level erasing, saliency-guided corruption, and mismatched-image counterfactuals under identical training and evaluation settings.

Table 5 shows that VC-OPD remains effective across a range of counterfactual constructions. All alternatives except mismatched-image counterfactuals outperform RKL, indicating that the proposed objective is not tied to a specific patch-dropout implementation. Patch dropout at 60% provides a

Table 5: Comparison of alternative counterfactual visual constructions under the Qwen2.5-VL-3B student with the Qwen2.5-VL-7B teacher. The RKL result is shown as the reference baseline.

Counterfactual	Configuration	Math Avg.	Vision Avg.	Overall Avg.
No counterfactual	RKL	48.80	43.77	46.64
Random dropout	–	49.86	46.20	48.29 (+3.54%)
Patch dropout	20%	49.72	45.98	48.12 (+3.17%)
Patch dropout	60%	**50.45**	**46.76**	**48.87 (+4.78%)**
Patch dropout	100%	49.20	45.55	47.64 (+2.14%)
Gaussian blur	Kernel 31×31, $\sigma = 10$	50.06	46.35	48.47 (+3.92%)
Contiguous region masking	60% masked area	50.30	46.55	48.69 (+4.40%)
Object-level erasing	Randomly erase object instances	49.62	45.70	47.94 (+2.79%)
Saliency-guided corruption	Mask high-saliency regions	**50.40**	**46.83**	**48.87 (+4.78%)**
Mismatched image	Random batch permutation	47.22	41.05	44.36 (-4.89%)

favorable balance between weakening instance-specific evidence and preserving a meaningful visual prediction context, and is therefore retained as the default configuration.

The mismatched-image condition performs substantially worse, suggesting that replacing the original visual context with unrelated evidence does not provide an appropriate counterfactual for estimating token-level visual dependence. More targeted constructions, such as object-level erasing and saliency-guided corruption, are promising but may require additional models or computations, whereas patch dropout provides a simple and lightweight alternative.

5 Conclusion

We introduced VC-OPD, a visual counterfactual extension of on-policy distillation for grounded vision-language reasoning. By comparing teacher scores under factual and counterfactual visual inputs, VC-OPD estimates token-level visual dependence and uses it to reweight dense distillation updates. The factual teacher–student discrepancy determines the update direction, while the counterfactual signal adjusts the update magnitude through soft weighting and mass-preserving normalization. Experiments across multiple student scales and the Qwen2.5-VL, Qwen3-VL, and InternVL3.5 families show consistent improvements over reverse-KL on-policy distillation in the evaluated settings. These findings support grounding-aware token-level credit assignment as a useful complement to general multimodal reasoning objectives.

6 Limitations

VC-OPD is primarily designed to improve visually grounded token-level credit assignment rather than generic text-only reasoning. Its gains may therefore be more modest on tasks dominated by linguistic logic, symbolic manipulation, or visually invariant deduction. Moreover, visual dependence is not equivalent to reasoning importance: some visually invariant transition or reasoning-control tokens may still play an important role in organizing the reasoning trajectory and can therefore be underweighted. The factual–counterfactual teacher gap is also based on visual perturbations rather than a perfect causal intervention, and the additional counterfactual teacher evaluation introduces computational overhead. Future work could combine visual-dependence weighting with complementary signals for reasoning importance and develop more precise and efficient counterfactual interventions.

7 Acknowledgments and Declarations

This work received research funding from Tianjin University. The experiments also used computational resources obtained through the AutoDL cloud computing platform, which provided computing services only and did not sponsor the research. The authors declare no competing interests that could reasonably be perceived as influencing this work.

References

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025a.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025a.

Yan Chen, Long Li, Teng Xi, Long Zeng, and Jingdong Wang. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models. *arXiv preprint arXiv:2509.13031*, 2025.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

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. *arXiv preprint arXiv:2306.13649*, 2024.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025b.

Jiaming Ji, Xinyu Chen, Rui Pan, Conghui Zhang, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Chi-Min Chan, Yida Tang, Sirui Han, Yike Guo, and Yaodong Yang. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. *arXiv preprint arXiv:2503.17682*, 2025.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*, 2025b.

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

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes. *arXiv preprint arXiv:2602.15260*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025b.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

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*, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

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

Zhenwen Li, Xuehui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Shuai Bai, Yuxuan Cai, Rui-Zhe Chen, Ke qin Chen, Xiong-Hui Chen, Zesen Cheng, Liang-Hao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025b. URL `https://api.semanticscholar.org/CorpusID:283262018`.

Wei-Yun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Sheng-Long Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv*, abs/2508.18265, 2025c. URL `https://api.semanticscholar.org/CorpusID:280710824`.

Ziwei Zheng, Michael Yang, Jack Hong, Chen-Xiao Zhao, Guo-Hai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *ArXiv*, abs/2505.14362, 2025. URL `https://api.semanticscholar.org/CorpusID:278769859`.

Yi-Fan Zhang, Xingyu Lu, Shu-Kang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kai-Yu Jiang, Chang-Yi Liu, Tian-Ke Zhang, Hao-Nan Fan, Kaibing Chen, Jian-Kang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Ting-Ting Gao, and Guorui Zhou. Thyme: Think beyond images. *ArXiv*, abs/2508.11630, 2025c. URL `https://api.semanticscholar.org/CorpusID:280671726`.

A Additional theoretical analysis

This appendix provides additional theoretical justification for VC-OPD. The main text already shows that OPD can be interpreted as token-level policy optimization induced by the reverse KL divergence. Here we focus on three properties that are specific to VC-OPD: (i) the factual–counterfactual teacher gap is a pointwise visual-dependence signal, (ii) mass-preserving normalization stabilizes the total update scale while preserving relative token importance, and (iii) the counterfactual branch modulates update magnitude without changing the update direction.

A.1 Counterfactual teacher gap as visual likelihood ratio

Let $s_t = (x, v, y_{<t})$ and $a_t = y_t$. For the same sampled token, VC-OPD evaluates the teacher under the factual visual input and the counterfactual visual input:

$$p_T^+(a_t) = p_T(a_t \mid x, v, y_{<t}), \qquad p_T^-(a_t) = p_T(a_t \mid x, \tilde{v}, y_{<t}). \tag{14}$$

The visual dependence score used in VC-OPD is

$$\Delta_t^{\text{vis}} = \log p_T^+(a_t) - \log p_T^-(a_t) = \log \frac{p_T(a_t \mid x, v, y_{<t})}{p_T(a_t \mid x, \tilde{v}, y_{<t})}. \tag{15}$$

Thus, Δ_t^{vis} is a pointwise log-likelihood ratio between the factual and counterfactual teacher distributions. A positive value means that the real image increases the teacher's support for the token relative to the counterfactual image, while a value close to zero means that the token is largely invariant to the removal of instance-specific visual evidence. This justifies using Δ_t^{vis} as a token-level grounding signal.

VC-OPD converts this score into a non-negative soft weight:

$$w_t = \sigma\left(\frac{\Delta_t^{\text{vis}} - \tau_v}{T_v}\right). \tag{16}$$

Since the sigmoid function is monotonic, if $\Delta_i^{\text{vis}} > \Delta_j^{\text{vis}}$, then $w_i > w_j$. Therefore, the soft weighting function preserves the ranking induced by the visual likelihood ratio: tokens more dependent on real visual evidence receive larger update weights.

A.2 Mass-preserving normalization

Directly using w_t may make the total update scale vary across responses, because different examples contain different numbers of visually grounded tokens. VC-OPD therefore uses the normalized weight

$$\bar{w}_t = m_t w_t \cdot \frac{N}{\sum_{j=1}^{T} m_j w_j + \epsilon}, \qquad N = \sum_{t=1}^{T} m_t. \tag{17}$$

where m_t is the response mask. Ignoring the small numerical constant ϵ, the normalized weights satisfy

$$\sum_{t=1}^{T} \bar{w}_t = N. \tag{18}$$

Thus, the total token weight is approximately equal to the number of valid response tokens. At the same time, for any two valid tokens i and j,

$$\frac{\bar{w}_i}{\bar{w}_j} = \frac{w_i}{w_j}. \tag{19}$$

Therefore, mass-preserving normalization keeps the relative visual emphasis unchanged while stabilizing the overall update magnitude. This property explains why VC-OPD can emphasize visually grounded tokens without changing the effective learning scale across batches.

A.3 Direction–magnitude separation

The final VC-OPD advantage is

$$A_t^{\mathrm{VC}} = \bar{w}_t \cdot \tilde{A}_t^{\mathrm{dir}}, \qquad A_t^{\mathrm{dir}} = \log p_T^+(a_t) - \log \mu(a_t \mid s_t), \tag{20}$$

where $\tilde{A}_t^{\mathrm{dir}}$ is a normalized version of the factual teacher–student direction term. The resulting objective is

$$\mathcal{L}_{\text{VC-OPD}}(\theta) = -\mathbb{E}_{y\sim\mu}\left[\sum_{t=1}^{T} \mathrm{sg}\big(A_t^{\mathrm{VC}}\big) \log \pi_\theta(a_t \mid s_t)\right]. \tag{21}$$

Taking the gradient gives

$$\nabla_\theta \mathcal{L}_{\text{VC-OPD}} = -\mathbb{E}_{y\sim\mu}\left[\sum_{t=1}^{T} \mathrm{sg}\Big(\bar{w}_t \tilde{A}_t^{\mathrm{dir}}\Big) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]. \tag{22}$$

This is still an advantage-weighted policy-gradient estimator, but the advantage is reweighted by token-level visual dependence.

Since $\bar{w}_t \geq 0$, the counterfactual branch cannot reverse the update direction:

$$\mathrm{sign}\big(\bar{w}_t A_t^{\mathrm{dir}}\big) = \mathrm{sign}\big(A_t^{\mathrm{dir}}\big) \quad \text{whenever } \bar{w}_t > 0. \tag{23}$$

Therefore, the factual teacher–student discrepancy determines whether the student should increase or decrease the probability of a sampled token, while the counterfactual teacher only determines how strongly this update should be applied. This separation is important for stability: the corrupted visual input is used to estimate grounding strength, but it does not directly decide the sign of the optimization signal.

A.4 View as selective token-level distillation

VC-OPD can also be interpreted as performing OPD under a reweighted token distribution. For a fixed response, define

$$q(t \mid y) = \frac{m_t w_t}{\sum_{j=1}^{T} m_j w_j}. \tag{24}$$

Then the VC-OPD objective is proportional to

$$-\mathbb{E}_{y\sim\mu}\left[N \cdot \mathbb{E}_{t\sim q(\cdot|y)}\left[\mathrm{sg}\Big(\tilde{A}_t^{\mathrm{dir}}\Big) \log \pi_\theta(a_t \mid s_t)\right]\right]. \tag{25}$$

Thus, VC-OPD can be viewed as standard token-level OPD applied to a token distribution biased toward visually grounded tokens, while preserving the total update mass. This interpretation formalizes the main design goal of VC-OPD: retain the dense supervision advantage of OPD, but reallocate learning capacity toward tokens that are supported by real visual evidence.

B Training Details

B.1 Stage-I warm-start

Before visual counterfactual on-policy distillation, we first warm-start the student with teacher-generated responses. For each training example, the teacher is prompted to produce a reasoning response followed by a boxed final answer. The prompt template is:

```
{{ content | trim }}
You FIRST think about the reasoning process as an internal monologue
and then provide the final answer.  The reasoning process MUST BE
enclosed within <think> </think> tags.  The final answer MUST BE put
in \boxed{}.
```

After generation, we extract the final boxed answer and compare it with the ground-truth answer using the same answer-extraction and exact-match evaluation protocol as in the main experiments. Only examples for which the teacher produces a correct final answer are retained for warm-start

Table 6: Warm-start training hyperparameters.

Hyperparameter	Value
Student model	Qwen2.5-VL-3B-Instruct
Teacher model	Qwen2.5-VL-32B-Instruct
Optimizer	AdamW
Learning rate	1×10^{-6}
Batch size	256
Response format	`<think>` reasoning + `\boxed{}` answer

Table 7: Stage-II training hyperparameters for GRPO and VC-OPD.

Hyperparameter	GRPO	VC-OPD
Optimizer	AdamW	AdamW
Learning rate	1×10^{-6}	5×10^{-6}
Batch size (rollouts)	5	1
Max response length	8192	8192
Max prompt length	4096	4096
Generation temperature	1.0	1.0
Top-p	1.0	1.0
Tensor parallel size	2	2
Visual token ratio	–	10%
Counterfactual corruption	–	Patch dropout 60%

training. This filtering removes noisy teacher trajectories and ensures that the student is initialized with reliable response formats and valid reasoning paths before entering on-policy optimization.

The warm-start stage uses standard autoregressive supervision on the filtered teacher responses. Table 6 summarizes the warm-start hyperparameters.

B.2 Stage-II on-policy training

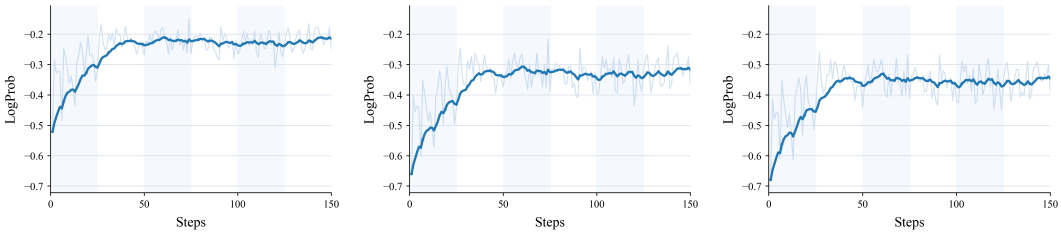
After warm-start, we perform on-policy optimization using the same training prompts. GRPO is trained with verifiable outcome-level rewards, while VC-OPD uses the student-generated trajectories for teacher scoring under both factual and counterfactual visual inputs. The stage-II implementation is based on the `verl` training framework. For VC-OPD, the factual teacher branch provides the token-level distillation direction, while the counterfactual teacher branch estimates visual dependence and reweights the token-level objective.

Table 7 reports the main hyperparameters used for GRPO and VC-OPD.

C Log-probability Diagnostics

We further visualize the log-probability traces during VC-OPD training to verify that the factual and counterfactual teacher scoring remains numerically stable. At each optimization step, we record the average response-token log probability for the student model, the factual teacher, and the counterfactual teacher. The light curve denotes the raw step-wise value, and the darker curve denotes the smoothed trend.

As shown in Figure 5, the student log probability increases rapidly in the early stage and then gradually stabilizes, reflecting the effect of warm-starting followed by on-policy adaptation. The factual and counterfactual teacher branches remain stable throughout training, which is important because VC-OPD relies on their log-probability difference to estimate token-level visual dependence. The counterfactual branch does not serve as an independent training target; instead, it provides a relative grounding signal by measuring how teacher support changes when instance-specific visual evidence is weakened. These diagnostics suggest that the counterfactual scoring procedure does not introduce unstable probability estimates and can be used reliably for grounding-aware token reweighting.



(a) Student model (b) Factual teacher (c) Counterfactual teacher

Figure 5: Log-probability diagnostics during VC-OPD training. We plot the average response-token log probability of the student, factual teacher, and counterfactual teacher over training steps. The curves stabilize after the early optimization phase, indicating that both teacher-scoring branches remain in a comparable and numerically stable prediction regime.

D Token-level Grounding Verification

The qualitative visualization in Figure 1 suggests that the factual–counterfactual teacher gap more selectively highlights image-dependent tokens than the standard student–teacher gap. To quantitatively verify this mechanism, we conduct a token-level grounding verification study on a subset of validation examples.

We sample model responses from the validation set and annotate response tokens as visually grounded if their correctness depends on image-specific evidence, such as objects, quantities, spatial relations, diagram symbols, chart elements, or answer-relevant visual attributes. Tokens that only serve as discourse markers, generic reasoning templates, or language-prior continuations are annotated as non-visual. Given these annotations, we compare four token scoring signals:

- **Random score**: a random token ranking baseline.
- **Token log-probability**: the factual teacher log-probability of each token.
- **Student–teacher gap**: the standard OPD score based on the teacher–student log-probability difference.
- **Factual–counterfactual gap**: the proposed visual-dependence score used by VC-OPD.

For each scoring signal, we rank all response tokens within each example and evaluate how well the top-ranked tokens match the human visual-grounding annotations. We report Precision@5%, Precision@10%, Recall@10%, and AUPRC. These metrics measure whether a scoring signal can retrieve visually grounded tokens from the full response.

As shown in Table 8, the factual–counterfactual gap provides the strongest alignment with human-annotated visually grounded tokens. The random baseline roughly reflects the sparsity of visually grounded tokens, while token log-probability performs only slightly better because high-probability tokens often correspond to common function words or response templates. The student–teacher gap improves over these baselines, but it still ranks many generic reasoning tokens highly because it measures teacher–student disagreement rather than visual dependence. In contrast, the factual–counterfactual gap achieves the highest precision and recall, indicating that it more accurately prioritizes tokens whose correctness depends on image-specific evidence. This result supports the central mechanism of VC-OPD: the counterfactual teacher branch does not merely identify high-discrepancy tokens, but provides a token-level signal that is more closely aligned with visual grounding.

E Generalization to Qwen3-VL

To evaluate whether VC-OPD transfers to newer VLM backbones, we further conduct experiments with Qwen3-VL-2B Bai et al. [2025b] and Qwen3-VL-4B students using Qwen3-VL-8B as the teacher. RKL and VC-OPD use the same student initialization, training data, rollout budget, and evaluation protocol. We additionally report the performance of the original base models to provide a direct reference for the effectiveness of post-training.

Table 8: Token-level grounding verification. We compare different token scoring signals for identifying human-annotated visually grounded tokens.

Scoring signal	P@5%	P@10%	Recall@10%	AUPRC
Random score	11.8	11.6	10.1	12.0
Token log-probability	18.7	17.4	15.2	18.6
Student–teacher gap	36.5	31.8	27.7	34.2
Factual–counterfactual gap	**58.9**	**52.1**	**45.4**	**57.6**

Table 9: Generalization results on the Qwen3-VL family. We report exact-match accuracy (%) under the same evaluation protocol as the main experiments.

Model	Geo3K	MVer.	MVist.	WeM.	MMMU	Logic	Count	Avg.
Qwen3-VL-2B								
+ Base	28.20	39.50	59.80	56.60	27.50	36.80	55.10	43.36
+ RKL	39.50	42.80	63.00	60.50	31.40	40.90	61.20	48.47
+ VC-OPD	**42.40**	**44.20**	**64.50**	**61.90**	**33.80**	**42.80**	**63.80**	**50.49**
Δ_{RKL} (%)	+7.34	+3.27	+2.38	+2.31	+7.64	+4.65	+4.25	+4.16
Qwen3-VL-4B								
+ Base	35.20	42.80	64.00	61.80	32.20	41.50	62.00	48.50
+ RKL	43.20	45.50	66.70	64.50	35.60	44.80	67.30	52.51
+ VC-OPD	**46.70**	**47.80**	**68.30**	**66.40**	**38.40**	**47.50**	**70.40**	**55.07**
Δ_{RKL} (%)	+8.10	+5.05	+2.40	+2.95	+7.87	+6.03	+4.61	+4.87

VC-OPD consistently improves over RKL for both Qwen3-VL student scales. For the 2B and 4B students, VC-OPD achieves average accuracies of 50.49 and 55.07, corresponding to relative improvements of 4.16% and 4.87% over RKL, respectively. Compared with the corresponding base models, the gains are also substantial, with absolute improvements of 7.13 and 6.57 points in average accuracy. These results indicate that the proposed visual-dependent token reweighting remains effective when transferred to the newer Qwen3-VL family.

F Generalization to InternVL3.5

To assess cross-family generalization, we further implement VC-OPD on the VERLInternVL framework and evaluate InternVL3.5-2B Wang et al. [2025c] and InternVL3.5-4B students with an InternVL3.5-8B teacher. RKL and VC-OPD use the same student initialization, training data, rollout budget, and evaluation protocol. We also include the corresponding base-model performance as a reference for evaluating the effect of post-training.

VC-OPD consistently outperforms RKL for both InternVL3.5 student scales, achieving relative improvements of 4.70% and 3.76% in average accuracy for the 2B and 4B students, respectively. Relative to the corresponding base models, VC-OPD further improves average accuracy by 6.03 and 6.37 points, respectively. The consistent gains across both student scales and a different VLM family demonstrate that VC-OPD is not restricted to the Qwen architecture used in the main experiments, and supports its cross-family generalization.

G Training Efficiency

VC-OPD evaluates the teacher under both factual and counterfactual visual conditions. The two conditions share the same student-generated trajectory and are concatenated along the batch dimension for a single batched teacher invocation. In our implementation, the additional counterfactual branch

Table 10: Generalization results on the InternVL3.5 family. We report exact-match accuracy (%) under the same evaluation protocol as the main experiments.

Model	Geo3K	MVer.	MVist.	WeM.	MMMU	Logic	Count	Avg.
InternVL3.5-2B								
+ Base	30.80	39.40	60.50	28.40	31.20	41.00	56.20	41.07
+ RKL	37.90	42.10	64.90	35.20	34.00	39.80	61.00	44.99
+ VC-OPD	**40.80**	**44.20**	**66.10**	**37.40**	**35.80**	**42.30**	**63.10**	**47.10**
Δ_{RKL} (%)	+7.65	+4.99	+1.85	+6.25	+5.29	+6.28	+3.44	+4.70
InternVL3.5-4B								
+ Base	39.20	49.80	71.10	35.80	38.80	53.20	64.80	50.39
+ RKL	44.70	53.10	73.50	47.00	42.00	54.10	68.50	54.70
+ VC-OPD	**47.30**	**55.00**	**74.40**	**49.10**	**44.30**	**56.00**	**71.20**	**56.76**
Δ_{RKL} (%)	+5.82	+3.58	+1.22	+4.47	+5.48	+3.51	+3.94	+3.76

Table 11: Wall-clock time (hours) required to reach target Geo3K accuracies for Qwen2.5-VL-3B.

Target Accuracy	**29%**	**30%**	**31%**	**32%**	**33%**	**34%**
RKL	2.5	4.0	7.0	14.5	–	–
VC-OPD	**1.5**	**2.8**	**5.0**	**7.2**	**14.4**	**17.4**

increases the wall-clock time per optimization step by approximately 20%, rather than doubling the end-to-end training cost.

Table 11 compares the wall-clock time required by RKL and VC-OPD to reach the same target Geo3K accuracy. Although VC-OPD introduces additional teacher computation at each optimization step, it consistently reaches the reported accuracy targets earlier than RKL. The difference becomes particularly evident at higher accuracy levels: VC-OPD reaches 32% accuracy in 7.2 hours compared with 14.5 hours for RKL, and further reaches 33% and 34% in 14.4 and 17.4 hours, respectively, whereas RKL does not reach these targets within the reported training budget. These results indicate that the extra per-step computation can be offset by faster optimization progress, leading to improved time-to-accuracy.

H Comparison with Recent Visual Reasoning Methods

We additionally compare the Qwen2.5-VL-7B student distilled from the Qwen2.5-VL-32B teacher with recent visual reasoning post-training methods on four shared benchmarks. Since DeepEyes Zheng et al. [2025] and Thyme-VL Zhang et al. [2025c] report results obtained under potentially different training data, evaluation splits, and tool-augmented inference procedures, this comparison should be interpreted as an external reference rather than a strictly controlled comparison.

Table 12 reports the results on MathVista, MathVerse, We-Math, and LogicVista. VC-OPD achieves an average accuracy of 59.13% across the four benchmarks. Compared with the reported results of DeepEyes and Thyme-VL, VC-OPD shows particularly large gains on We-Math, while the differences on the other benchmarks are more modest. Because the compared methods may differ in training resources, inference settings, and evaluation protocols, these results are provided primarily to contextualize the empirical performance of VC-OPD rather than to establish a controlled state-of-the-art comparison.

Table 12: Comparison with recent visual reasoning methods on shared benchmarks.

Method	MathVista	MathVerse	We-Math	LogicVista	Avg.
DeepEyes	70.10	47.30	38.90	47.70	51.00
Thyme-VL	70.00	39.10	39.30	**49.00**	49.35
VC-OPD	**70.90**	**48.90**	**68.10**	48.60	**59.13**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction present VC-OPD as a method for using factual–counterfactual visual dependence to reweight token-level on-policy distillation. The experiments in Sections 4.1–4.6 and Appendices E–F support the reported improvements within the evaluated benchmarks, model scales, and VLM families.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the method's scope, the distinction between visual dependence and reasoning importance, the limitations of visual perturbations as counterfactual interventions, and the additional computation required by counterfactual scoring.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix A provides algebraic interpretations of the visual-dependence score, mass-preserving normalization, and direction–magnitude separation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 4.1 and Appendices B–H report the datasets, benchmarks, training stages, several hyperparameters, and selected efficiency measurements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The manuscript provides a public code repository at `https://github.com/feng-feng932/VC-OPD`, containing full implementation, training/evaluation scripts and instructions to reproduce our main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix B describe the main models, benchmarks, evaluation metric, optimizer, learning rates, batch sizes, sequence limits, and several training settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The reported tables and figures provide point estimates, but the manuscript does not report repeated-run variability, error bars, confidence intervals, or statistical significance tests for the main comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper states that experiments use four NVIDIA A800 GPUs . It does not provide GPU memory requirements, CPU and storage details, per-run resource accounting for every experiment, or the total compute used across the full project.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research studies post-training and evaluation of vision-language reasoning models using cited existing datasets, benchmarks, and model families. The paper does not describe human-subject experiments, deceptive procedures, privacy-invasive data collection, or other activities that would require deviation from the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 of this manuscript discusses the technical limitations as well as associated broader impacts of our method.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new high-risk pretrained model, image generator, or scraped dataset. The work proposes a training objective and evaluates it on cited existing models and benchmarks, so responsible-release safeguards for newly released high-risk assets are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the existing model family, training dataset, evaluation benchmarks, and training framework used in the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The manuscript does not indicate that a new dataset, model checkpoint, or asset package is being released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not report any crowdsourcing experiments or new research with human subjects. It uses existing datasets and benchmarks for model training and evaluation, so participant instructions, screenshots, and compensation details are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subjects research conducted by the authors. Therefore, participant-risk disclosure and IRB or equivalent institutional approval are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Vision-language models are central to the method as student and teacher models, and their roles are described in Sections 3–4 and Appendices B, E, and F. The manuscript identifies the model families and the factual and counterfactual teacher-scoring procedure used by the proposed method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.