TRIP-ADV: TOKEN-LEVEL CREDIT ASSIGNMENT FOR MULTIMODAL REINFORCEMENT LEARNING VIA CONTROLLED VISUAL INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards improves multimodal reasoning, but most methods assign the same sequence advantage to every response token. This uniform credit treats grounded decisions, language priors, and accidental guesses alike, weakening supervision for tokens that truly depend on visual evidence. We propose TRIP-Adv, a framework that identifies these tokens through controlled visual interventions and assigns credit. An exponential moving average teacher scores each trajectory under the original image, target enhancement, target corruption, and matched random corruption. These comparisons estimate whether target evidence is sufficient to support a token and necessary for its prediction, while reducing sensitivity to generic image changes. Agreement with the outcome verifier and consensus across interventions calibrate the evidence. The calibrated signal is sparsely routed to reshape token advantages and concentrate learning on visually grounded reasoning. Across seven benchmarks, TRIP-Adv achieves the best overall performance among models of comparable scale. The code is available in the appendix.

1 INTRODUCTION

General-purpose multimodal large language models (MLLMs) build strong visual–language interfaces through pretrained vision and language components (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Chen et al., 2023b). Fine-grained reasoning, however, requires locating small regions for text recognition, attribute judgment, spatial analysis, or multi-step reasoning. Because decisive evidence is often surrounded by irrelevant content, models may instead rely on language priors and produce plausible but visually ungrounded answers (Li et al., 2023b; Yu et al., 2024). High-resolution encoding, dynamic tiling, and adaptive cropping preserve more local detail (Li et al., 2024; Ye et al., 2023b;a; Guo et al., 2024; Wang et al., 2024b), but access to detail does not ensure that it determines the response.

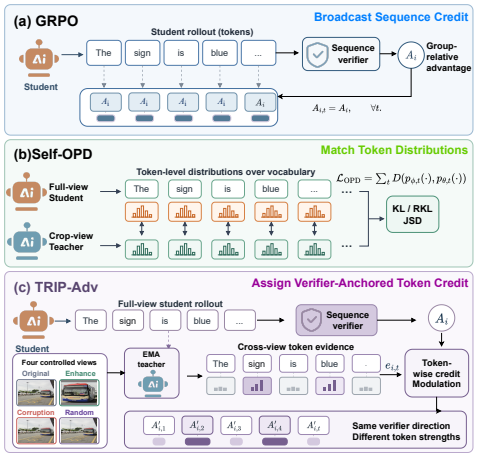


Figure 1: Credit-assignment comparison. GRPO broadcasts one sequence-level scalar, Self-OPD densely matches cross-view distributions, and TRIP-Adv uses controlled views to allocate token-level credit.

Recent work has pursued several ways to strengthen fine-grained visual reasoning. Region-grounded models associate language with localized image content (Peng et al., 2024; Chen et al., 2023a; You et al., 2024), while visual-search systems actively crop, zoom, or annotate relevant regions during reasoning (Wu & Xie, 2024; Hu et al., 2024; Cao et al., 2025). These methods can recover evidence missed in the full image but often

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.

require repeated encoding or tool use. Training-only approaches instead transfer privileged regional information into a single-pass policy through supervision or on-policy distillation (Wei et al., 2026; Yuan et al., 2026; Li et al., 2026a; Sun et al., 2026). In parallel, GRPO reinforces answer-correct trajectories with verifiable rewards (Shao et al., 2024), and multimodal variants add outcome or perception rewards (Liu et al., 2025; Xiao et al., 2025).

These approaches still leave a central credit-assignment problem: GRPO broadcasts one final-answer advantage to visually grounded decisions, generic language templates, and accidental guesses alike. Recent methods address this mismatch by estimating token-level visual dependence. VPPO reweights trajectories and filters token gradients according to perceptual dependence (Huang et al., 2026); PGPO uses a threshold-gated, mass-conserving redistribution of advantage toward visually dependent tokens (Ye et al., 2026); and PRPO identifies perturbation-stable perceptual tokens before dynamically reshaping their advantages (Li et al., 2026b). However, visual-dependence estimates can still be confounded by missing context, transformation artifacts, or generic sensitivity to view changes (Pan et al., 2026). Without controlled comparisons and outcome-aware calibration, a large cross-view response need not indicate task-relevant evidence, especially when the sampled answer is incorrect.

These observations motivate the following question: *Can controlled training-time visual interventions identify which tokens depend on task-relevant visual evidence and thereby provide finer-grained credit assignment for multimodal RL?*

TRIP-Adv addresses this question by pairing outcome supervision with a controlled visual evidence critic (Figure 1). An EMA teacher replays each trajectory under the original view, target enhancement, target corruption, and operation-matched random corruption while holding the prefix and sampled token fixed. Enhancement estimates evidence sufficiency, whereas the matched corruption contrast estimates target-specific necessity after subtracting generic perturbation sensitivity. Because automatically constructed views may remain imperfect, TRIP-Adv calibrates these residuals by answer-correctness agreement and cross-intervention consensus, then sparsely modulates GRPO advantages only for strongly supported tokens. The resulting policy preserves efficient single-pass inference.

Our contributions are as follows:

- We introduce token-level credit assignment that combines outcome supervision with controlled visual evidence.
- We derive calibrated visual evidence from target enhancement and matched corruption using verifier agreement and cross-intervention consensus.
- We propose sparse advantage modulation that improves GRPO while preserving efficient single-pass inference.

2 Related Work

Fine-grained multimodal perception. Foundational MLLMs connect pretrained vision and language components through cross-attention, query transformers, or lightweight projection layers (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Chen et al., 2023b). Tiling, adaptive cropping, and dynamic-resolution systems retain more high-resolution or text-rich detail (Li et al., 2024; Ye et al., 2023b;a; Guo et al., 2024; Wang et al., 2024b). Region-grounded models further expose points, boxes, and free-form regions to the language interface (Peng et al., 2024; Chen et al., 2023a; You et al., 2024). Visual-search and tool-augmented systems actively inspect these local regions (Wu & Xie, 2024; Hu et al., 2024; Cao et al., 2025), but require multi-step inference. Zooming without Zooming transfers crop-grounded supervision to a full-image model and introduces ZoomBench (Wei et al., 2026). Vision-OPD aligns a crop-conditioned teacher and full-image student on student rollouts (Yuan et al., 2026); Visual-OPSD distills privileged visual-thought traces (Li et al., 2026a); and V-Zero gates answer-label-free OPD with positive and negative visual views (Sun et al., 2026). Unlike behavioral imitation, TRIP-Adv treats the teacher as an evidence critic and uses only sampled-token residuals to allocate RL credit.

RLVR and on-policy distillation. GRPO estimates an advantage from within-group relative rewards without a value model (Shao et al., 2024). Visual-RFT applies task-specific verifiable rewards

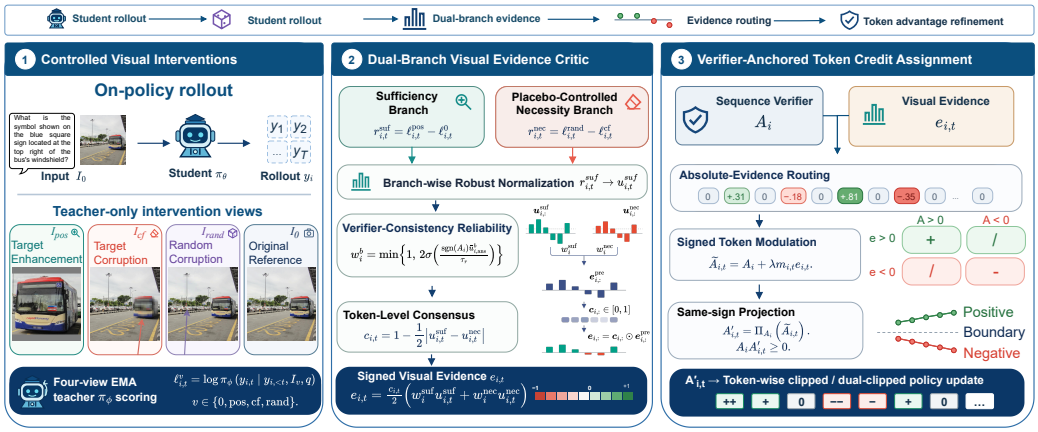


Figure 2: TRIP-Adv overview. The student generates one on-policy rollout from I_0. The EMA teacher replays the same tokens under I_0, target enhancement I_{pos}, target corruption I_{cf}, and matched random corruption I_{rand}. Dual residuals are calibrated and routed to modulate token advantages under a same-sign constraint.

to visual perception and grounding, and Perception-R1 introduces a separate reward for perceptual content (Liu et al., 2025; Xiao et al., 2025). Multimodal preference alignment, including MDPO and fine-grained correctional RLHF-V, strengthens image-conditioned behavior and reduces hallucination (Wang et al., 2024a; Yu et al., 2024), but does not assign on-policy token credit from controlled evidence. Generalized knowledge distillation reduces exposure mismatch by querying a teacher on student-generated sequences (Agarwal et al., 2024); on-policy self-distillation conditions the same model on privileged and ordinary contexts (Zhao et al., 2026). RLCSD contrasts correct and incorrect privileged contexts to cancel shared style shifts before modulating RL advantages (Pan et al., 2026). TRIP-Adv shares the verifier-anchored motivation, but obtains evidence from paired visual interventions and explicitly controls generic image-corruption sensitivity.

3 Method

3.1 Overview

Given an image I_0, question q, and answer $a^\star$, the student samples on-policy responses, and an EMA teacher scores each fixed trajectory under the original, target-enhanced, target-corrupted, and matched random-corruption views. As shown in Figure 2, the resulting visual evidence is calibrated and sparsely routed to modulate token advantages, with a projection enforcing consistency with the original advantage sign. The additional views and teacher are used only during training; inference requires one standard generation from I_0.

3.2 Controlled Visual Interventions

We use fine-grained VQA samples from ZwZ-RL-VQA (Wei et al., 2026), each containing an original image, a question, a reference answer, and a target bounding box. Around the same target region, we construct

$$\mathcal{I} = \{I_0, I_{\text{pos}}, I_{\text{cf}}, I_{\text{rand}}\}. \tag{1}$$

Here I_0 is the unmarked original image, I_{pos} highlights task-relevant evidence, I_{cf} suppresses that evidence, and I_{rand} controls for generic sensitivity to local image perturbations. All views retain the same question and reference answer.

Target-region enhancement. We expand target box B by 5% along both width and height, crop the expanded region, and letterbox-resize it to the original canvas. The resulting I_{pos} increases the

effective resolution of the target while retaining a small amount of neighboring context. It measures how the response changes when relevant evidence becomes easier to perceive.

Target-evidence corruption. To form I_{cf}, we modify only the interior of B and leave the remainder unchanged. The operation is sampled from strong blur, pixelation, local block permutation, and texture filling, with parameters determined by target scale and a per-sample random seed. These operators suppress complementary cues such as fine detail, text, color, texture, and object structure, reducing the risk that a fixed occlusion pattern becomes a shortcut.

Operation-matched random corruption. The response to I_{cf} can combine the effect of losing target evidence with the model's general sensitivity to local damage. We therefore sample a box B^{-} outside the target with the same width and height and require $\mathrm{IoU}(B, B^{-}) < 0.05$. When other annotated objects exist, we prefer candidates with minimal aggregate overlap. We then apply exactly the same operator, strength, and random parameters used for I_{cf}, producing I_{rand}. The principal difference between I_{cf} and I_{rand} is thus whether the intervention covers the task target. Matching location, geometry, and corruption parameters reduces generic perturbation effects in the necessity estimate.

Quality control. We first verify that all images are readable, have identical canvas sizes, and use target and random boxes of matching shapes and areas. We then use Qwen3.5-397B-A17B as an external visual judge f and retain samples satisfying

$$f(I_0, q) = f(I_{\mathrm{pos}}, q) = f(I_{\mathrm{rand}}, q) = a^{\star}, \qquad f(I_{\mathrm{cf}}, q) \neq a^{\star}. \tag{2}$$

These conditions jointly test label preservation and intervention effectiveness: the judge predicts the reference answer from the original, enhanced, and random-corruption views, but changes its prediction under target corruption. The four views are bound to the same question–answer pair through a shared sample identifier. A sample missing any view or failing any consistency check does not enter the four-view TRIP-Adv path.

3.3 Dual-Branch Visual Evidence

Our next goal is to determine which response tokens actually depend on the target region. Under a single view, a token probability entangles visual information, language priors, the textual prefix, and token frequency. The student therefore generates only under I_0, after which the EMA teacher scores the same sampled sequence under each visual condition:

$$\ell_{i,t}^{v} = \log \pi_{\phi}(y_{i,t} \mid y_{i,<t}, I_v, q), \quad v \in \{0, \mathrm{pos}, \mathrm{cf}, \mathrm{rand}\}. \tag{3}$$

Because the question, response prefix, and target token are fixed, cross-view log-probability differences primarily reflect the visual condition. The EMA teacher serves only as a stable evidence critic; the student is not required to imitate its full vocabulary distribution.

Sufficiency branch. We measure the additional support obtained by making the target easier to see:

$$r_{i,t}^{\mathrm{suf}} = \ell_{i,t}^{\mathrm{pos}} - \ell_{i,t}^{0}. \tag{4}$$

When $r_{i,t}^{\mathrm{suf}} > 0$, target enhancement raises the probability of the sampled token. Since language context and the token itself are shared, their common contribution is largely canceled. The residual is therefore an approximate measure of whether the target evidence is sufficient to support that token.

Necessity branch. A direct original-versus-corrupted contrast would mix target-evidence removal with generic corruption artifacts. Instead, we compare two views using matched operations:

$$r_{i,t}^{\mathrm{nec}} = \ell_{i,t}^{\mathrm{rand}} - \ell_{i,t}^{\mathrm{cf}}. \tag{5}$$

If $r_{i,t}^{\mathrm{nec}} > 0$, corrupting the target hurts the token more than corrupting a matched non-target region, indicating target-specific dependence. Matching the operator, strength, and region scale cancels part of the model's general response to local image damage.

Robust branch-wise normalization. Enhancement and corruption may induce different log-probability scales, which also vary with target size, response length, and sample difficulty. We normalize each branch separately over valid, non-special response tokens $\mathcal{V}_i$:

$$s_i^b = \frac{1}{|\mathcal{V}_i|} \sum_{t \in \mathcal{V}_i} |r_{i,t}^b|, \quad b \in \{\text{suf}, \text{nec}\}, \tag{6}$$

$$u_{i,t}^b = \tanh\left(\frac{r_{i,t}^b}{s_i^b + \tau_b} \right) \tag{7}$$

We use $\tau_{\text{suf}} = \tau_{\text{nec}} = 0.02$. Independent scaling makes the branches more comparable across trajectories, while $\tanh$ suppresses extreme residuals and bounds evidence within $[-1, 1]$. The sign is deliberately retained: positive values support the sampled token and negative values contradict it. Negative values are meaningful counter-evidence, rather than noise to be discarded.

3.4 Verifier-Consistent Calibration

Sensitivity to a visual intervention does not by itself make a reliable learning signal. A crop can omit context, a corruption can be incomplete, and a random box can include unannotated evidence; incorrect answers can also be visually sensitive. We therefore calibrate evidence first at the trajectory–branch level against the outcome verifier, and then at the token level across the two branches.

Let Ω_i^{ans} be the final-answer tokens, located from the task output structure or an answer mask. Let R_i^{ver} denote the signed outcome-correctness reward, with positive values for a correct final answer and negative values for an incorrect one. We define the verifier direction from this outcome reward,

$$d_i = \text{sgn}(R_i^{\text{ver}}), \tag{8}$$

rather than from the sign of the composite group-relative advantage

$$A_i = R_i - \frac{1}{G} \sum_{j=1}^{G} R_j. \tag{9}$$

For each branch, we average evidence on the answer span and compare its direction with d_i:

$$\bar{u}_{i,\text{ans}}^b = \frac{\sum_{t \in \Omega_i^{\text{ans}}} u_{i,t}^b}{|\Omega_i^{\text{ans}}|}, \quad q_i^b = d_i \bar{u}_{i,\text{ans}}^b, \tag{10}$$

$$w_i^b = \min\left\{1, 2\sigma(q_i^b / \tau_r)\right\} \tag{11}$$

with $\tau_r = 0.1$. A reliable branch supports the sampled answer when the correctness verifier marks it correct and contradicts it when the verifier marks it incorrect. Both cases give $q_i^b > 0$. This one-sided gate leaves consistent branches at unit weight and continuously downweights conflicting branches. If the final answer span cannot be located reliably, the rollout falls back to its original sequence advantage.

After trajectory-level gating, we form preliminary fused evidence

$$e_{i,t}^{\text{pre}} = \frac{1}{2} \left(w_i^{\text{suf}} u_{i,t}^{\text{suf}} + w_i^{\text{nec}} u_{i,t}^{\text{nec}} \right). \tag{12}$$

The token-level consensus between the two branches is

$$c_{i,t} = 1 - \frac{1}{2} \left| u_{i,t}^{\text{suf}} - u_{i,t}^{\text{nec}} \right|, \qquad c_{i,t} \in [0, 1], \tag{13}$$

and the final evidence is

$$e_{i,t} = c_{i,t} e_{i,t}^{\text{pre}}. \tag{14}$$

Agreement produces $c_{i,t} \approx 1$, whereas opposing and strongly divergent estimates drive it toward zero. Thus w_i^b filters intervention branches that conflict with the answer-level outcome, and $c_{i,t}$ suppresses token-wise evidence not supported by both complementary interventions. Only evidence compatible with sequence supervision and stable across the dual branches receives high weight.

3.5 VERIFIER-ANCHORED TOKEN CREDIT

The calibrated $e_{i,t}$ describes the direction and strength of token dependence on target evidence, but cannot determine whether the whole response deserves reward or penalty. TRIP-Adv therefore gives the two signals different roles: A_i decides the global optimization direction, while $e_{i,t}$ redistributes update magnitude inside that trajectory.

Absolute-evidence routing. Many connectives, formatting symbols, and generic reasoning tokens do not directly depend on the image. Applying visual modulation densely would dilute the relevant signal and inject small estimation errors into policy optimization. We use thresholding followed by Top-ρ routing:

$$\begin{aligned} \mathcal{E}_i &= \{t \in \mathcal{V}_i : |e_{i,t}| \geq \delta\}, \qquad k_i = \lceil \rho |\mathcal{E}_i| \rceil, \\ \mathcal{K}_i &= \mathrm{TopK}_{k_i}(\mathcal{E}_i; |e_{i,t}|), \qquad m_{i,t} = \mathbf{1}[t \in \mathcal{K}_i]. \end{aligned} \tag{15}$$

When $\mathcal{E}_i = \varnothing$, $k_i = 0$ and every $m_{i,t} = 0$. Because Top-ρ acts only on tokens that pass the threshold, the actual routed fraction is usually below ρ. We rank by $|e_{i,t}|$, not only positive evidence, because both strong support and strong contradiction identify visual dependence. Magnitude determines whether to route a token; sign determines whether its advantage should be strengthened or attenuated.

Sign-safe modulation. For a routed token, we add visual evidence to the trajectory advantage and project the result to the original sign interval:

$$\begin{aligned} \widetilde{A}_{i,t} &= A_i + \lambda m_{i,t} e_{i,t}, \qquad A'_{i,t} = \Pi_{A_i}(\widetilde{A}_{i,t}), \\ \Pi_A(z) &= \mathrm{sgn}(A)\big[\mathrm{sgn}(A) z\big]_+, \qquad \mathrm{sgn}(0) = 0. \end{aligned} \tag{16}$$

For a positive advantage, positive evidence increases a token's reward and negative evidence reduces it; for a negative advantage, negative evidence increases the penalty and positive evidence mitigates it. The projection implements these magnitude adjustments within the original advantage-sign interval.

Policy optimization and EMA update. The sequence reward combines answer correctness, output-format compliance, and thinking-format quality:

$$\begin{aligned} R_i &= \alpha_{\mathrm{ans}} R_i^{\mathrm{ans}} + \alpha_{\mathrm{fmt}} R_i^{\mathrm{fmt}} + \alpha_{\mathrm{think}} R_i^{\mathrm{think}}, \\ A_i &= R_i - \frac{1}{G} \sum_{j=1}^{G} R_j. \end{aligned} \tag{17}$$

To distinguish the policy importance ratio from evidence residuals $r^b_{i,t}$, we denote it by $\kappa_{i,t}(\theta)$:

$$\begin{aligned} \kappa_{i,t}(\theta) &= \frac{\pi_\theta(y_{i,t} \mid y_{i,<t}, I_0, q)}{\pi_{\mathrm{old}}(y_{i,t} \mid y_{i,<t}, I_0, q)}, \\ \bar{\kappa}_{i,t}(\theta) &= \mathrm{clip}(\kappa_{i,t}(\theta), 1 - \epsilon_{\mathrm{low}}, 1 + \epsilon_{\mathrm{high}}) \,. \end{aligned} \tag{18}$$

The clipped surrogate is

$$S_{i,t}(\theta) = \min\big[\kappa_{i,t}(\theta) A'_{i,t}, \bar{\kappa}_{i,t}(\theta) A'_{i,t}\big] \,. \tag{19}$$

For negative-advantage tokens, we additionally use dual clipping (Yu et al., 2026) to restrict updates caused by abnormal ratios:

$$\mathcal{D}_{i,t}(\theta) = \begin{cases} S_{i,t}(\theta), & A'_{i,t} \geq 0, \\ \max\{S_{i,t}(\theta), c_{\mathrm{dual}} A'_{i,t}\}, & A'_{i,t} < 0, \end{cases} \tag{20}$$

$$\mathcal{L}_{\text{TRIP-Adv}} = -\frac{\sum_i \sum_{t \in \mathcal{V}_i} \mathcal{D}_{i,t}(\theta)}{\sum_i |\mathcal{V}_i|}. \tag{21}$$

TRIP-Adv introduces no additional visual-distillation loss. Teacher scoring is stop-gradient, and interventions influence the student only through $A'_{i,t}$. After every actor update,

$$\phi \leftarrow (1 - \eta)\phi + \eta\theta. \tag{22}$$

This lets the evidence critic evolve smoothly with the student, limiting mismatch with the changing on-policy distribution. At inference, only π_θ remains and generates once from I_0.

Algorithm 1 TRIP-Adv training

```
Sample {y_i}_{i=1}^G ~ π_θ(· | I_0, q)
Compute verifier rewards and A_i using Eq. (9)
for each rollout y_i do
    Replay y_i with π_φ on all four views
    Compute dual residuals, normalization, and e_{i,t}
    Route by |e_{i,t}| and compute A'_{i,t}
    Fall back to A_i if views or answer span are invalid
end for
Update θ with token advantages A'_{i,t}
Update teacher φ ← (1 − η)φ + ηθ
```

Table 1: Comparison on seven benchmarks (accuracy, %). Avg. denotes the equal-weight macro average within each task group. Results for visually grounded reasoning models are taken from their original papers. Large-scale models are context only and excluded from best/second-best formatting.

Method	Fine-Grained Visual Tasks						Holdout Tasks			Overall Avg.
	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-CN	Avg.	CV-Bench	MMStar	Avg.	
General Large Vision-Language Models										
Qwen3-VL-4B (Bai et al., 2025)	82.7	44.9	79.8	74.5	63.2	69.0	84.7	72.9	78.8	71.8
Qwen3-VL-8B (Bai et al., 2025)	83.8	43.3	78.4	75.6	64.4	69.1	86.2	76.5	81.4	72.6
Qwen3.5-4B	83.8	47.5	84.8	77.8	65.4	71.9	86.2	72.7	79.5	74.0
Qwen3.5-9B	87.4	50.1	84.1	79.9	66.8	73.7	86.1	72.9	79.5	75.3
MiMo-VL-7B-RL (Yue et al., 2025)	81.7	46.6	74.1	69.6	54.4	65.3	79.9	71.7	75.8	68.3
Large-Scale Models										
Qwen3.6-27B	89.5	54.9	86.0	83.1	68.2	76.3	87.2	76.2	81.7	77.9
Qwen3-VL-235B-A22B (Bai et al., 2025)	88.0	49.1	84.5	81.6	65.3	73.7	86.7	76.3	81.5	75.9
Kimi-K2.5 (1T) (Team et al., 2026)	85.9	56.3	81.9	75.4	68.4	73.6	89.2	81.8	85.5	77.0
Visually Grounded Reasoning Models										
DeepEyesV2-7B (Hong et al., 2025)	81.7	45.0	77.9	73.8	65.1	68.7	–	–	–	–
Thyme-7B (Zhang et al., 2025a)	82.2	45.1	77.0	72.0	64.6	68.2	–	–	–	–
Vision-OPD-4B (Yuan et al., 2026)	<u>92.2</u>	<u>59.8</u>	84.5	80.4	70.8	<u>77.5</u>	87.3	**79.6**	**83.5**	<u>79.2</u>
ZwZ-4B (Wei et al., 2026)	**92.7**	55.7	81.8	79.5	68.1	75.6	<u>87.9</u>	71.1	79.5	76.7
ZwZ-8B (Wei et al., 2026)	91.1	58.1	84.4	<u>82.0</u>	70.6	77.2	87.4	73.1	80.3	78.1
V-Zero-4B (Sun et al., 2026)	89.0	57.8	<u>87.8</u>	**82.6**	–	–	–	74.4	–	–
Our Model										
TRIP-Adv-4B (Qwen3 based)	90.2	57.6	83.7	80.2	<u>71.1</u>	76.6	87.2	<u>78.2</u>	<u>82.7</u>	78.3
TRIP-Adv-4B (Qwen3.5 based)	92.1	**61.2**	**88.5**	**82.6**	**72.8**	**79.4**	**89.3**	77.6	**83.5**	**80.6**

4 EXPERIMENTS

4.1 SETUP AND BENCHMARKS

We validate TRIP-Adv on two 4B base models, Qwen3-VL-4B and Qwen3.5-4B. Both variants are trained for one epoch on the four-view data with the same optimization setup. We evaluate accuracy on five fine-grained benchmarks—ZoomBench (Wei et al., 2026), V* Bench (Wu & Xie, 2024), HR-Bench-4K/8K (Wang et al., 2025), and MME-RealWorld-CN (Zhang et al., 2025b)—plus the held-out general benchmarks MMStar (Chen et al., 2024) and CV-Bench (Tong et al., 2024). We report equal-weight macro averages for the fine-grained, holdout, and overall groups. Detailed training configurations and evaluation protocols are available in the appendix.

4.2 COMPARISON WITH EXISTING MODELS

Table 1 compares TRIP-Adv with general-purpose MLLMs and visually grounded reasoning models. TRIP-Adv-4B reaches 79.4 on the fine-grained average, 83.5 on the holdout average, and 80.6 overall, corresponding to relative gains of 10.4%, 5.0%, and 8.9% over its base model. The Qwen3-based variant also reaches 78.3 overall and an 82.7 holdout average, showing that the gains are not tied to a single initialization. Among comparable-scale models, TRIP-Adv obtains the best overall average and leads on ZoomBench, both HR-Bench tracks, MME-RealWorld-CN, and CV-Bench.

Table 2: Training method comparison (accuracy, %). The final column reports mean wall-clock seconds per training step.

Method	V*	Zoom	HR-4K	HR-8K	MME-CN	CV	MMStar	Overall	Mean s/step
GRPO	86.4	56.9	81.1	77.6	72.0	87.0	69.8	75.8	275.7
OPD	91.6	58.7	80.0	78.8	71.3	86.8	61.5	75.5	269.0
TRIP-Adv (Full)	**92.1**	**61.2**	**88.5**	**82.6**	**72.8**	**89.3**	**77.6**	**80.6**	407.1

Table 3: Visual evidence branches (accuracy, %).

Variant	V*	Zoom	H4K	H8K	MME	CV	MM*	All
Positive only	89.5	55.3	82.5	78.6	68.3	84.7	68.9	75.4
Unmatched removal	89.0	56.0	84.0	80.6	71.6	88.1	72.2	77.4
Matched removal	91.1	57.6	85.1	81.9	72.6	88.2	73.1	78.5
Full (Qwen3.5)	**92.1**	**61.2**	**88.5**	**82.6**	**72.8**	**89.3**	**77.6**	**80.6**

Table 4: Evidence calibration (accuracy, %).

Variant	V*	Zoom	H4K	H8K	MME	CV	MM*	All
No calibration	90.1	60.7	84.3	79.9	71.9	83.3	70.2	77.2
No reliability	91.6	59.9	86.1	**83.6**	71.8	88.1	71.5	78.9
No consensus	90.6	58.1	85.5	82.4	71.2	86.9	72.3	78.1
Full (Qwen3.5)	**92.1**	**61.2**	**88.5**	82.6	**72.8**	**89.3**	**77.6**	**80.6**

4.3 Training Method Comparison

Unless otherwise stated, Table 2 and all subsequent experiments use Qwen3.5-4B. GRPO applies a sequence-level composite reward, OPD distills a privileged-view teacher on policy, and TRIP-Adv adds controlled four-view evidence to token credit. TRIP-Adv reaches 80.6 overall, yielding relative improvements of 6.3% over GRPO and 6.8% over OPD. OPD improves three benchmarks but remains weaker on MMStar, whereas TRIP-Adv improves all seven, suggesting that calibrated token credit strengthens fine-grained perception without sacrificing general capability. The timing results show only a modest efficiency gap among the three methods, indicating that controlled evidence introduces limited training overhead relative to its accuracy gains.

4.4 Ablations

We analyze the key components of TRIP-Adv at three levels: the source of visual evidence, evidence calibration, and token credit assignment. Except for the component under study, each ablation uses the full model's configuration.

Table 3 isolates the source of visual evidence. Positive Only uses the enhancement residual $I_{\mathrm{pos}} - I_0$. Unmatched Removal estimates necessity by contrasting the original and target-corrupted views, $I_0 - I_{\mathrm{cf}}$, whereas Matched Removal contrasts random-region and target-region corruption, $I_{\mathrm{rand}} - I_{\mathrm{cf}}$, to separate target-specific evidence loss from generic perturbation effects. Operation matching yields a 1.4% relative improvement over unmatched removal, and combining matched necessity with sufficiency provides a further 2.7% gain.

Table 4 evaluates the two calibration stages. No Calibration removes both answer-consistency reliability and token consensus; No Reliability keeps only consensus; and No Consensus keeps only the answer-level gate. Joint calibration improves overall accuracy over the uncalibrated model by 4.4%. No Reliability is the strongest partial variant, while the complete calibration performs best, indicating that trajectory-level reliability and token-level consensus provide complementary filtering.

Table 5 compares token-allocation strategies. All-token Routing modulates every valid response token; Random Routing selects the same number of tokens as evidence routing but chooses them uniformly; No Projection keeps evidence-based sparse routing while allowing token advantages to cross zero. The full method improves overall accuracy over Random Routing, All-token Routing, and No Projection by 11.6%, 3.7%, and 4.7%, respectively, showing that evidence-aware sparsity and sign-safe projection each contribute to performance.

4.5 Visual Evidence Analysis

Figure 3(a–b) shows systematic responses to the controlled views: enhancement supports the sampled tokens, matched random corruption has little effect, and removing target evidence changes them most strongly. The separation between target and random corruption indicates that the critic captures task-specific evidence rather than generic image damage. In Figure 3(c), correct rollouts have somewhat stronger fused evidence, but the overlapping distributions confirm that visual dependence

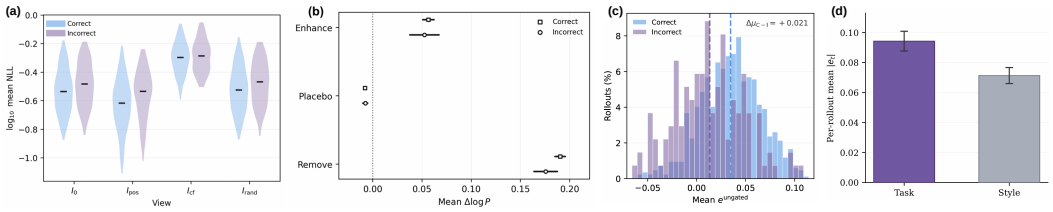


Figure 3: Visual-evidence analysis after training. (a) Four-view uncertainty and (b) paired intervention effects reproduce the view-level analysis: target enhancement is supportive, matched random corruption is near zero, and target removal has the strongest effect. (c) Fused-evidence distributions for correct and incorrect rollouts. (d) Evidence after a large model annotates task-relevant tokens and task-independent style tokens.

Table 5: Token credit assignment (accuracy, %).

Variant	V*	Zoom	H4K	H8K	MME	CV	MM*	All
All-token routing	91.1	56.8	86.9	80.8	71.3	85.4	71.6	77.7
Random routing	87.4	54.1	76.9	72.9	68.5	84.7	60.9	72.2
No projection	91.1	58.5	83.1	79.1	71.1	86.2	70.2	77.0
Full (Qwen3.5)	**92.1**	**61.2**	**88.5**	**82.6**	**72.8**	**89.3**	**77.6**	**80.6**

alone is not a correctness signal and motivate verifier-conditioned calibration. Figure 3(d) further shows that task tokens annotated by a large model receive stronger evidence than task-independent style tokens, supporting sparse routing toward semantically decisive content. Concrete annotated spans and evidence values are reported in Appendix Table 6.

4.6 Qualitative Attention Analysis

Figure 4 provides a qualitative view of the behavior encouraged by evidence-aware credit assignment. Under the same input question, the base model distributes attention more broadly over the scene, whereas TRIP-Adv places a stronger response around the annotated key region. This example is consistent with the intended role of TRIP-Adv: assigning stronger learning credit to tokens whose prediction depends on localized task evidence, rather than changing the inference interface. Additional qualitative examples are provided in the appendix.

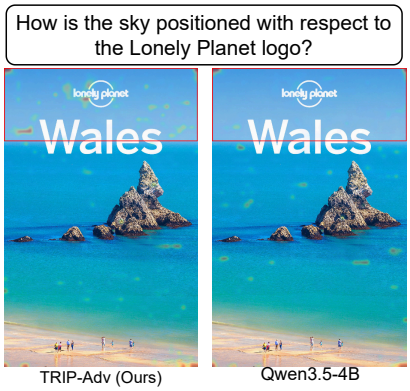


Figure 4: Attention visualization for the same question and image. Compared with Qwen3.5-4B, TRIP-Adv focuses more strongly on the task-critical region.

5 Conclusion

We presented TRIP-Adv, a token-level credit-assignment framework for multimodal RLVR. Controlled target enhancement and matched corruption estimate complementary visual evidence on fixed student rollouts; outcome consistency and cross-intervention consensus calibrate that evidence; and sparse modulation allocates the resulting signal to visually dependent tokens. TRIP-Adv-4B improves fine-grained and holdout performance over GRPO while retaining ordinary single-image inference. Future work could extend box-based interventions with learned region proposals and evaluate variance across independent training runs.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, support retrieval and discovery such as finding related work, assist with research ideation and execution, draft sections of the paper, generate synthetic datasets, and prove mathematical claims.

ETHICS STATEMENT

This work studies post-training and evaluation of multimodal language models using existing benchmarks and does not involve the collection of human-subject data. As with other multimodal models, deployment may inherit biases, privacy risks, and failure modes from the underlying models and data. The constructed visual interventions and external visual judge may also introduce systematic biases; we therefore use them only during training and evaluate the resulting model on multiple held-out benchmarks.

REPRODUCIBILITY STATEMENT

The main paper specifies the intervention construction, evidence scores, calibration, routing, and policy objective. The appendix provides expanded derivations, validity and fallback conditions, complete optimization and sampling settings, the per-step training procedure, and evaluation protocols. These sections, together with the supplementary qualitative examples, document the information needed to reproduce the method and reported evaluations.

REFERENCES

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

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.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023b.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: An lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pp. 390–406, 2024.

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

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. volume 37, pp. 139348–139379, 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*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

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

Qiming Li, Tianlun Li, Xiaolong Cheng, Hangyu Li, Ruiyan Gong, Kangning Niu, Kaitao Jiang, and Mu Xu. PRPO: Perception-reinforced policy optimization via token-level dynamic advantage reshaping. *arXiv preprint arXiv:2606.08708*, 2026b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26763–26773, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, 2024.

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. 2024.

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning. *arXiv preprint arXiv:2606.25319*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, Adithya Jairam Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

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

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

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward. *arXiv e-prints*, pp. arXiv–2506, 2025.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023a.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2841–2858, 2023b.

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.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 5, 2025.

YiFan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. In *NeurIPS 2025 Fourth Workshop on Deep Learning for Code*, 2025a.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025b.

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.

SUPPLEMENTARY MATERIAL

This appendix provides expanded derivations and properties of the method, the complete methodological and optimization setup, and qualitative examples. All notation follows the main paper. The additional visual views, bounding boxes, and EMA teacher described below are used only during training.

A ADDITIONAL METHOD DETAILS

A.1 NOTATION AND SEQUENCE-LEVEL SIGNAL

A training item is

$$x = (I_0, I_{\text{pos}}, I_{\text{cf}}, I_{\text{rand}}, q, a^\star, B, B^-), \tag{23}$$

where I_0 is the unmarked original image, q is the question, $a^\star$ is the reference answer, B is the task-relevant target box, and B^- is an operation-matched non-target box. The remaining three images are constructed interventions. For each (I_0, q), the student samples a group of G sibling responses

$$y_i = (y_{i,1}, \ldots, y_{i,T_i}) \sim \pi_\theta(\cdot \mid I_0, q), \qquad i = 1, \ldots, G. \tag{24}$$

Table 6: Representative large-model annotations of task-relevant and task-independent style spans, with their mean absolute evidence.

Token class	Annotated response span	Mean $\lvert e_t \rvert$
Task-relevant	*black shape*	0.9724
	circular symbol	0.9439
	umbrella canopy	0.8505
	4 visible panels	0.8459
Task-independent style	*Check choices*	0.7517
	Let's check	0.6719
	So the	0.6471
	There are	0.6257

Let $\mathcal{V}_i \subseteq \{1, \ldots, T_i\}$ contain valid, non-padding, non-special response positions. A composite sequence reward R_i yields the non-standardized group-relative advantage

$$A_i = R_i - \frac{1}{G} \sum_{j=1}^{G} R_j. \tag{25}$$

Thus $\sum_i A_i = 0$ within a complete rollout group. Unlike the standard broadcast implementation, TRIP-Adv converts A_i into token-dependent $A'_{i,t}$. The intervention critic reallocates update magnitude within the sign interval defined by A_i.

A.2 Why the Four Scores Are Paired

After the student samples y_i only from I_0, the EMA teacher π_ϕ is teacher-forced on the *same* question, response prefix, and sampled next token under every visual condition:

$$\ell^v_{i,t} = \log \pi_\phi(y_{i,t} \mid y_{i,<t}, I_v, q), \quad v \in \{0, \mathrm{pos}, \mathrm{cf}, \mathrm{rand}\}. \tag{26}$$

No response is re-sampled from an intervention view. Consequently, a cross-view difference does not compare different decoded trajectories: the language prefix $y_{i,<t}$ and evaluated token $y_{i,t}$ are held fixed, and only the visual condition changes. This pairing removes variation caused by decoding different prefixes and makes the residuals local contrasts on the sampled trajectory.

A crop can change global context and a corruption can introduce artifacts. TRIP-Adv treats the paired differences as evidence proxies and uses matched random corruption, verifier consistency, and dual-branch consensus to reduce these effects.

A.3 Formalization of the Controlled Views

Let the target box be $B = (x_1, y_1, x_2, y_2)$ on a canvas of width W and height H. We expand it by $\gamma = 0.05$ of its width and height, clip the result to the image boundary, crop it, and letterbox-resize the crop to the original canvas:

$$\begin{aligned} B^+ &= \mathrm{Clip}\big(\mathrm{Expand}(B, \gamma), [0, W] \times [0, H]\big), \\ I_{\mathrm{pos}} &= \mathrm{Letterbox}\big(\mathrm{Crop}(I_0, B^+); W, H\big). \end{aligned} \tag{27}$$

The target-corrupted view is

$$I_{\mathrm{cf}} = \mathcal{C}(I_0; B, \kappa, \xi), \quad \kappa \in \{\text{blur}, \text{pixelate}, \text{shuffle}, \text{texture}\}, \tag{28}$$

where ξ contains the strength and sampled random parameters. Only pixels inside B are modified.

For the control, we choose a box B^- with the same width and height as B, require $\mathrm{IoU}(B, B^-) < 0.05$, and prefer candidates with little overlap with other annotated objects. We then reuse the *same* operator and parameters:

$$I_{\mathrm{rand}} = \mathcal{C}(I_0; B^-, \kappa, \xi). \tag{29}$$

Matching (κ, ξ) and the box geometry makes intervention location the principal systematic difference between I_{cf} and I_{rand}. The contrast remains an approximate placebo-controlled region-specific effect because an unannotated relevant object can occur inside B^-.

A.4 DUAL-BRANCH RESIDUALS

The sufficiency-oriented branch is

$$r_{i,t}^{\mathrm{suf}} = \ell_{i,t}^{\mathrm{pos}} - \ell_{i,t}^{0} = \log \frac{\pi_\phi(y_{i,t} \mid y_{i,<t}, I_{\mathrm{pos}}, q)}{\pi_\phi(y_{i,t} \mid y_{i,<t}, I_0, q)}. \tag{30}$$

Hence $r_{i,t}^{\mathrm{suf}} > 0$ means that making the target easier to inspect increases teacher support for the sampled token; a negative value means that the enhanced target contradicts it.

The necessity-oriented branch is

$$r_{i,t}^{\mathrm{nec}} = \ell_{i,t}^{\mathrm{rand}} - \ell_{i,t}^{\mathrm{cf}} = \log \frac{\pi_\phi(y_{i,t} \mid y_{i,<t}, I_{\mathrm{rand}}, q)}{\pi_\phi(y_{i,t} \mid y_{i,<t}, I_{\mathrm{cf}}, q)}. \tag{31}$$

A positive value means that corrupting the target lowers support more than applying the same corruption to a matched non-target location. In contrast, the seemingly simpler residual $\ell_{i,t}^{0} - \ell_{i,t}^{\mathrm{cf}}$ mixes removal of target evidence with generic sensitivity to local damage. Equation (31) subtracts part of the latter response.

The two branches are intentionally asymmetric. I_{pos} changes both scale and visible context, so I_0 is its natural reference. In the removal branch, the operation-matched random view is the appropriate reference for isolating location-specific damage.

A.5 BRANCH-WISE ROBUST NORMALIZATION

For $b \in \{\mathrm{suf}, \mathrm{nec}\}$, define

$$s_i^b = \frac{1}{|\mathcal{V}_i|} \sum_{t \in \mathcal{V}_i} |r_{i,t}^b|, \qquad u_{i,t}^b = \tanh\left(\frac{r_{i,t}^b}{s_i^b + \tau_b} \right). \tag{32}$$

We use $\tau_{\mathrm{suf}} = \tau_{\mathrm{nec}} = 0.02$. The mean absolute residual is used rather than a signed mean because positive and negative evidence could otherwise cancel. Normalizing each branch independently prevents a numerically higher-variance intervention from dominating fusion merely because its residual scale is larger.

Boundedness and sign preservation. For every valid token,

$$-1 < u_{i,t}^b < 1, \qquad \mathrm{sgn}(u_{i,t}^b) = \mathrm{sgn}(r_{i,t}^b). \tag{33}$$

This follows because $s_i^b + \tau_b > 0$ and $\tanh(\cdot)$ is odd, strictly increasing, and maps finite inputs to $(-1, 1)$. Thus the normalizer clips extreme magnitudes smoothly without discarding counter-evidence.

Degenerate residuals. If $r_{i,t}^b = 0$ for all $t \in \mathcal{V}_i$, then $s_i^b = 0$ and $u_{i,t}^b = 0$. The positive τ_b also avoids division by zero and reduces amplification when all raw residuals are very small.

A.6 VERIFIER-CONSISTENT RELIABILITY GATE

Let $\Omega_i^{\mathrm{ans}} \subseteq \mathcal{V}_i$ be the content tokens inside the first valid final-answer span, and let R_i^{ver} be the signed outcome-correctness reward. We compute

$$\bar{u}_{i,\mathrm{ans}}^b = \frac{1}{|\Omega_i^{\mathrm{ans}}|} \sum_{t \in \Omega_i^{\mathrm{ans}}} u_{i,t}^b, \qquad q_i^b = \mathrm{sgn}(R_i^{\mathrm{ver}}) \bar{u}_{i,\mathrm{ans}}^b, \tag{34}$$

and use the one-sided gate

$$w_i^b = \min\left\{1,\ 2\sigma(q_i^b/\tau_r)\right\}, \qquad \tau_r = 0.1. \tag{35}$$

The gate has the following explicit piecewise behavior:

$$w_i^b = \begin{cases} 1, & q_i^b \geq 0, \\ 2\sigma(q_i^b/\tau_r), & q_i^b < 0. \end{cases} \tag{36}$$

Therefore $0 < w_i^b \leq 1$. Agreement never amplifies a branch; only conflict is downweighted. For $R_i^{\text{ver}} > 0$, agreement means that the branch supports the sampled answer. For $R_i^{\text{ver}} < 0$, it means that the branch contradicts the sampled answer. This is why the product in Equation (34) treats both cases uniformly.

When $A_i = 0$, the same-sign projection later returns zero, so visual evidence cannot create an update for that rollout. When Ω_i^{ans} cannot be located, the complete rollout falls back to $A'_{i,t} = A_i$ rather than estimating an answer mask heuristically.

A.7 Fusion and Cross-Intervention Consensus

The gated branches are first averaged:

$$e_{i,t}^{\text{pre}} = \frac{1}{2}\left(w_i^{\text{suf}} u_{i,t}^{\text{suf}} + w_i^{\text{nec}} u_{i,t}^{\text{nec}}\right). \tag{37}$$

The consensus and final evidence are

$$\begin{aligned} c_{i,t} &= 1 - \frac{1}{2}\left|u_{i,t}^{\text{suf}} - u_{i,t}^{\text{nec}}\right|, \\ e_{i,t} &= c_{i,t} e_{i,t}^{\text{pre}}. \end{aligned} \tag{38}$$

Because both normalized branches lie in $[-1, 1]$, their absolute difference is at most two; hence

$$0 \leq c_{i,t} \leq 1, \qquad |e_{i,t}| \leq 1. \tag{39}$$

Equal branch estimates give $c_{i,t} = 1$. Maximally opposed estimates, $u_{i,t}^{\text{suf}} = 1$ and $u_{i,t}^{\text{nec}} = -1$ (or vice versa), give $c_{i,t} = 0$. Reliability and consensus address different failure modes: w_i^b evaluates a whole intervention branch against the answer-level verifier direction, while $c_{i,t}$ detects disagreement at a particular token.

A.8 Sparse Absolute-Evidence Routing

Routing first applies a magnitude threshold and then selects a fraction of only the eligible tokens:

$$\begin{aligned} \mathcal{E}_i &= \{t \in \mathcal{V}_i : |e_{i,t}| \geq \delta\}, \\ k_i &= \lceil \rho |\mathcal{E}_i| \rceil, \\ \mathcal{K}_i &= \text{TopK}_{k_i}(\mathcal{E}_i; |e_{i,t}|), \\ m_{i,t} &= \mathbf{1}[t \in \mathcal{K}_i]. \end{aligned} \tag{40}$$

We use $\delta = 0.02$ and $\rho = 0.2$. If $\mathcal{E}_i = \varnothing$, we explicitly set $k_i = 0$ and $m_{i,t} = 0$ for all t. Otherwise the routed fraction among all valid tokens is

$$\frac{|\mathcal{K}_i|}{|\mathcal{V}_i|} = \frac{\lceil \rho |\mathcal{E}_i| \rceil}{|\mathcal{V}_i|}, \tag{41}$$

which is generally lower than ρ because $|\mathcal{E}_i| \leq |\mathcal{V}_i|$. Ties at the Top-$K$ boundary are resolved deterministically by the implementation. Using $|e_{i,t}|$ for ranking retains both strong visual support and strong visual contradiction; the sign is used only in the subsequent modulation.

A.9 Same-Sign Projection and Its Guarantee

For a routed token, the unconstrained candidate and final advantage are

$$\widetilde{A}_{i,t} = A_i + \lambda m_{i,t} e_{i,t}, \qquad A'_{i,t} = \Pi_{A_i}(\widetilde{A}_{i,t}), \tag{42}$$

where

$$\Pi_A(z) = \begin{cases} \max(z, 0), & A > 0, \\ \min(z, 0), & A < 0, \\ 0, & A = 0. \end{cases} \tag{43}$$

Proposition 1 (advantage-sign preservation). For every rollout i and valid token t,

$$A_i A'_{i,t} \geq 0, \qquad A_i = 0 \Longrightarrow A'_{i,t} = 0. \tag{44}$$

Table 7: Effect of signed evidence before the projection boundary is reached. "Increase" refers to advantage magnitude.

A_i	$e_{i,t}$	Effect on token credit
> 0	> 0	increase reward
> 0	< 0	attenuate reward
< 0	< 0	increase penalty
< 0	> 0	attenuate penalty

Proof. If $A_i > 0$, Equation (43) gives $A'_{i,t} = \max(\widetilde{A}_{i,t}, 0) \geq 0$, and therefore $A_i A'_{i,t} \geq 0$. If $A_i < 0$, it gives $A'_{i,t} = \min(\widetilde{A}_{i,t}, 0) \leq 0$, so the product is again non-negative. If $A_i = 0$, the third branch sets $A'_{i,t} = 0$. □

The projection preserves the sign interval while allowing evidence to reduce a token update to zero. For $m_{i,t} = 0$, it is an identity on A_i, so unrouted tokens retain $A'_{i,t} = A_i$.

A.10 Policy Objective and Gradient Boundary

Let

$$\kappa_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid y_{i,<t}, I_0, q)}{\pi_{\text{old}}(y_{i,t} \mid y_{i,<t}, I_0, q)} \tag{45}$$

be the policy importance ratio. With asymmetric clipping,

$$\bar{\kappa}_{i,t} = \text{clip}\left(\kappa_{i,t}, 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}\right), \tag{46}$$

the clipped surrogate is

$$S_{i,t} = \min\{\kappa_{i,t} A'_{i,t}, \bar{\kappa}_{i,t} A'_{i,t}\}. \tag{47}$$

For negative advantages, dual clipping gives

$$\mathcal{D}_{i,t} = \begin{cases} S_{i,t}, & A'_{i,t} \geq 0, \\ \max\{S_{i,t}, c_{\text{dual}} A'_{i,t}\}, & A'_{i,t} < 0, \end{cases} \tag{48}$$

and the token-mean loss is

$$\mathcal{L}_{TRIP-Adv} = -\frac{\sum_i \sum_{t \in \mathcal{V}_i} \mathcal{D}_{i,t}}{\sum_i |\mathcal{V}_i|}. \tag{49}$$

TRIP-Adv adds no KL, JSD, or full-vocabulary distillation objective. All teacher scores and the derived $A'_{i,t}$ are stop-gradient with respect to the actor update. Thus intervention views affect the student only by changing the scalar token credit multiplying the on-policy policy-gradient term.

A.11 Validity Checks and Exact Fallback Behavior

A rollout uses TRIP-Adv modulation only when its original and three intervention views are present, teacher scoring succeeds on the aligned sampled tokens, and a non-empty final-answer span can be identified. Otherwise,

$$A'_{i,t} = A_i, \qquad t \in \mathcal{V}_i, \tag{50}$$

so the rollout reduces exactly to the underlying GRPO update. The same fallback occurs naturally when no token passes the evidence threshold. This design prevents missing metadata or an uncertain answer mask from silently introducing fabricated token supervision.

B Detailed Methodological Setup

B.1 Four-View Data Construction

The seed samples come from ZwZ-RL-VQA (Wei et al., 2026) and contain an image, question, reference answer, and target bounding box. Human box markings are not provided to the student at training or inference. Before constructing interventions, we remove explicit textual references to a red box and reject malformed image paths, answers, or bounding boxes. A valid target box must

have ordered in-bound coordinates, sufficient crop area and short-side length, and a reasonable area ratio and aspect ratio. Samples whose question becomes ambiguous after removal of the artificial box reference are excluded.

The final training file contains 11,130 rows, corresponding to 11,130 unique VQA instances (both the source identifier and sample identifier are unique). Every row contains one original image and exactly three intervention views—target enhancement, target corruption, and matched random corruption—for a total of 44,520 unique image-view paths. The dataset size is therefore 11,130 question–answer instances, rather than 44,520 independent questions.

For each retained seed sample, the construction procedure emits exactly one record containing $(I_0, I_{\mathrm{pos}}, I_{\mathrm{cf}}, I_{\mathrm{rand}})$, a shared sample identifier, the unchanged question and answer, the target and random boxes, and the matched corruption metadata. Structural checks require readable images, identical canvas sizes, matching target/control box dimensions, and complete view paths. The semantic validation described in the main paper retains samples satisfying

$$\begin{aligned} &f(I_0, q) = f(I_{\mathrm{pos}}, q) = f(I_{\mathrm{rand}}, q) = a^\star, \\ &f(I_{\mathrm{cf}}, q) \neq a^\star, \end{aligned} \tag{51}$$

where f is the independent visual judge. No view is borrowed from a different sample to repair an incomplete record.

B.2 Sequence Reward

The rollout reward is

$$R_i = 0.80R_i^{\mathrm{ans}} + 0.15R_i^{\mathrm{fmt}} + 0.05R_i^{\mathrm{think}}. \tag{52}$$

R_i^{ans} evaluates answer correctness, R_i^{fmt} checks the required output structure, and R_i^{think} is a DAPO-style soft-overlong term. Let L_i be the number of valid response tokens, $L_{\max} = 2048$, and $L_{\mathrm{safe}} = 1536$. The length term is

$$R_i^{\mathrm{think}} = \begin{cases} 0, & L_i \le L_{\mathrm{safe}}, \\ -\dfrac{L_i - L_{\mathrm{safe}}}{L_{\max} - L_{\mathrm{safe}}}, & L_{\mathrm{safe}} < L_i < L_{\max}, \\ -1, & L_i \ge L_{\max}. \end{cases} \tag{53}$$

It does not reward a preferred reasoning length and imposes no minimum length. It only increases the penalty as a response approaches the hard rollout limit. Group advantages use Equation (25) without division by the within-group standard deviation.

B.3 Optimization and Sampling Configuration

We initialize both the student and teacher from Qwen3.5-4B and train for one data epoch on eight NVIDIA H200 GPUs. Table 8 summarizes the principal configuration used for the main experiments.

Table 8: Training, rollout, and TRIP-Adv hyperparameters.

Component	Hyperparameter	Value
Model	initialization	Qwen3.5-4B
Data	training epochs	1
GRPO	sibling rollouts per question G	8
GRPO/PPO	global / mini / micro batch size	$32/32/2$
PPO	epochs; lower / upper clip	$1;\ 0.2/0.3$
Optimization	peak LR; schedule; warm-up	2×10^{-6}; cosine; 10 steps
Context	maximum prompt / response length	$25{,}000/2{,}048$ tokens
Vision	maximum visual tokens per image	16,384
Rollout	temperature / top-p / top-k	$1.0/0.95/16$
TRIP-Adv	τ_{suf} / τ_{nec} / τ_r	$0.02/0.02/0.1$
TRIP-Adv	routing threshold / ratio (δ, ρ)	$0.02/0.2$
Teacher	EMA update coefficient η	0.05

B.4 Modulation Warm-Up and Teacher Update

To avoid applying intervention-derived credit before the on-policy teacher scores are sufficiently stable, the effective modulation coefficient follows

$$\lambda_s = \begin{cases} 0, & s < 30, \\ \lambda \dfrac{s-30}{30}, & 30 \le s < 60, \\ \lambda, & s \ge 60, \end{cases} \qquad \lambda = 0.01, \tag{54}$$

where s is the actor update step. Hence the first 30 steps use the unmodified GRPO advantage, followed by a 30-step linear ramp.

The teacher is initialized from the student and updated after each actor update:

$$\phi \leftarrow (1-\eta)\phi + \eta\theta, \qquad \eta = 0.05. \tag{55}$$

It is used only for sampled-token scoring under the four views. No teacher gradient is computed, and the teacher is removed at inference.

B.5 Per-Step Training Procedure

Algorithm 2 Expanded TRIP-Adv training step

Require: Complete four-view batch, student π_θ, EMA teacher π_ϕ
Sample $G = 8$ responses per question from (I_0, q)
Compute composite rewards and non-standardized group advantages
for each sampled rollout y_i **do**
  if all views and a valid answer span are available **then**
    Score the fixed sampled tokens with π_ϕ on four views
    Form r^{suf} and r^{nec} and normalize each branch
    Apply answer-consistency gates and token consensus
    Threshold by δ and route Top-ρ eligible tokens
    Add $\lambda_s e_{i,t}$ and apply same-sign projection
  else
    Set $A'_{i,t} \leftarrow A_i$ for every valid response token
  end if
end for
Update θ once with the dual-clipped token-mean objective
Stop actor gradients before all teacher quantities
Update $\phi \leftarrow (1-\eta)\phi + \eta\theta$

B.6 Evaluation Benchmarks and Reporting

We evaluate accuracy (%) on seven benchmarks using their official protocols. Five tasks form the fine-grained group:

- **ZoomBench** (Wei et al., 2026) contains 845 human-verified fine-grained VQA examples covering counting, OCR, color, structure, material, and object recognition.
- **V* Bench** (Wu & Xie, 2024) tests localization and recognition of small targets in complex high-resolution scenes.
- **HR-Bench-4K and HR-Bench-8K** (Wang et al., 2025) test high-resolution perception and fine-grained reasoning at their respective resolution tracks.
- **MME-RealWorld-CN** (Zhang et al., 2025b) uses Chinese questions on high-resolution, real-world scenes.

The two holdout benchmarks are **MMStar** (Chen et al., 2024), which contains 1,500 visually necessary questions spanning six core capabilities and 18 subdimensions, and **CV-Bench** (Tong et al., 2024), which evaluates 2D and 3D visual-spatial understanding. Neither holdout benchmark is used to construct the fine-grained four-view training data.

Fine-grained average is the unweighted mean of V*, ZoomBench, HR-Bench-4K, HR-Bench-8K, and MME-RealWorld-CN. Holdout average is the unweighted mean of CV-Bench and MMStar. Overall average is the unweighted mean of all seven benchmark accuracies; it is not an average of the two group averages. Missing public benchmark values are left unreported rather than imputed.

B.7 Training and Inference Interfaces

During training, each student rollout requires four teacher-forced sampled-token scoring passes. During inference, the deployed interface is simply

$$(I_0, q) \longrightarrow y \sim \pi_\theta(\cdot \mid I_0, q). \tag{56}$$

No bounding box, crop, corruption, external visual judge, extra visual tool, or teacher model is required. Therefore TRIP-Adv changes post-training compute and supervision, but not the base model's single-image inference interface.

C Qualitative Examples

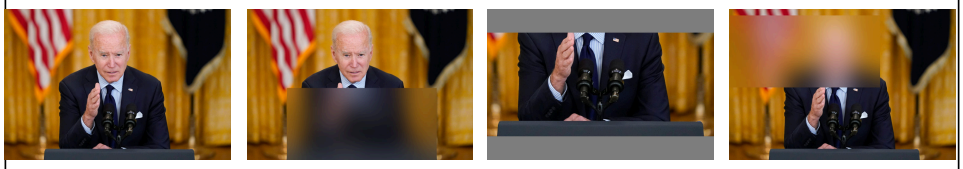
Figures 5–8 jointly present four representative examples, each combining the model question, the four-view training record, heatmaps from TRIP-Adv and Qwen3.5-4B, and their complete responses. The first two examples emphasize answer correctness. The green text marks the visually grounded reasoning used by TRIP-Adv, whereas the red text marks the incorrect spatial inference made by Qwen3.5-4B. In both cases, TRIP-Adv identifies the relevant object pair and resolves the requested relation correctly, while the base model attends to plausible scene content but confuses the direction or depth relation.

The last two examples emphasize reasoning efficiency and visual localization. Both models reach the correct final answer, but TRIP-Adv gives a shorter and more direct justification, highlighted in green. Its heatmaps are also more concentrated on the task-relevant logo–background or elephant–net regions, whereas the Qwen3.5-4B response is more verbose and its visual response is comparatively diffuse. Taken together, these examples illustrate two complementary advantages of TRIP-Adv: controlled intervention training improves the reliability of fine-grained spatial judgments, and token-level credit assignment encourages concise reasoning that remains anchored to the decisive visual evidence.

Model Question

In what position is the podium relative to the lapel pin?
Please identify their relative position: left of / right of / above / below / overlap / inside / contain / in front of / behind / next to.

Four Intervention Views (Training Only)



I_0 I_{cf} I_{pos} I_{rand}

Heatmap



TRIP-Adv (Ours) Qwen3.5-4B

Model Response

TRIP-Adv (Ours) : <think> The user wants me to determine the relative position of the podium with respect to the lapel pin.
1. Identify the lapel pin: It is on the suit jacket, near the top center of the image.
2. Identify the podium: It is at the bottom of the image.
3. Determine relative position: Since the pin is at the top and the podium is at the bottom, the podium is below the pin? No, the question is \"podium relative to the lapel pin\". So if the pin is above, the podium is below the pin. Therefore, the podium is \"below\" the pin? Wait.
- Pin is at the top.
- Podium is at the bottom.
- So relative to the pin, the podium is below.
Let's check options: \"below\" is an option.
So the podium is below the lapel pin.
Thus, relative to the lapel pin, the podium is below.
</think>
<answer>below</answer>

Qwen3.5-4B : <think> The user wants to know the relative position of the podium compared to the lapel pin.
1. **Identify the lapel pin:** It's a small American flag pin on the man's left lapel (viewer's right).
2. **Identify the podium:** It's the dark surface in the foreground where the microphones are attached.
3. **Determine the relationship:** The podium is in the foreground, closer to the camera. The man (and thus the lapel pin) is behind the podium. Therefore, the podium is *in front of* the lapel pin. Alternatively, the lapel pin is *behind* the podium. The question asks \"In what position is the podium relative to the lapel pin?\". So, the podium is in front of the pin.
</think>
The podium is in front of the lapel pin.<answer>in front of</answer>

Figure 5: Qualitative example 1: the podium relative to the lapel pin. The top portion shows the four training views, followed by the model heatmaps and full responses. TRIP-Adv correctly concludes that the podium is *below* the pin; Qwen3.5-4B instead predicts *in front of*. Green and red text highlight the corresponding correct and incorrect reasoning steps.

Model Question

In what position is the mirror relative to the flowers?
Please identify their relative position: left of / right of / above / below / overlap / inside / contain / in front of / behind / next to.

Four Intervention Views (Training Only)



I_0 I_{cf} I_{pos} I_{rand}

Heatmap



TRIP-Adv (Ours) Qwen3.5-4B

Model Response

TRIP-Adv (Ours)： <think> The user wants me to determine the relative position of the mirror with respect to the flowers.
In the image, the flowers are inside the mirror (reflected) or in front of it, but typically the mirror is the object behind the flowers in the reflection, or more simply, the mirror is the frame around the flowers. Looking at the image, the flowers are inside the mirror, so the mirror is behind the flowers? Wait, usually the question is \"where is the mirror relative to the flowers\". In the image, the flowers are in front of the mirror (reflected), so the mirror is behind the flowers.
So relative to the flowers, the mirror is behind.
Let's check options: behind is an option.
So the mirror is behind the flowers.
</think>
<answer>behind</answer>

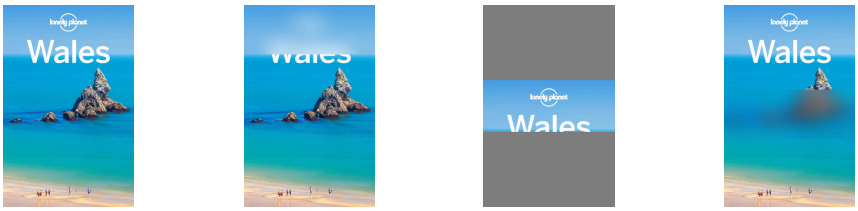
Qwen3.5-4B： <think> The user wants to know the relative position of the mirror and the flowers.
1. **Locate the mirror:** The mirror is visible on the right side of the image, mounted on the wall. It's an oval-shaped mirror.
2. **Locate the flowers:** There is a vase with yellow flowers sitting on a shelf below the mirror.
3. **Determine the relationship:** The mirror is positioned directly above the shelf where the flowers are placed. Therefore, the mirror is above the flowers.
</think>
The mirror is above the flowers.<answer>above</answer>

Figure 6: Qualitative example 2: the mirror relative to the flowers. TRIP-Adv uses the reflected flowers as the decisive cue and correctly predicts that the mirror is *behind* them. Qwen3.5-4B mistakes the relation for *above*. The colored text identifies the correct visual inference from TRIP-Adv and the erroneous inference from the base model.

Model Question

How is the sky positioned with respect to the Lonely Planet logo?
Please identify their relative position: left of / right of / above / below / overlap / inside / contain / in front of / behind / next to.

Four Intervention Views (Training Only)



I_0 I_{cf} I_{pos} I_{rand}

Heatmap



TRIP-Adv (Ours) Qwen3.5-4B

Model Response

TRIP-Adv (Ours)： <think> The user wants me to determine the relative position of the sky with respect to the Lonely Planet logo.
The logo is in the image.
So, the logo is in front of the sky, meaning the sky is behind the logo.
Therefore, with respect to the logo, the sky is behind it.
Check options: behind is an option.\nSo the answer should be \"behind\".
</think>
<answer>behind</answer>

Qwen3.5-4B： <think> The user wants to know the position of the sky relative to the Lonely Planet logo.
1. **Locate the Lonely Planet logo:** It's at the very top center of the image, consisting of the text \"lonely planet\" inside a circle.
2. **Locate the sky:** The sky is the blue area at the top of the image, above the horizon line where the sea meets the sky.
3. **Determine the relationship:** The logo is superimposed on the sky. Therefore, the sky is *behind* the logo. Alternatively, the logo is *in front of* the sky. The question asks \"How is the sky positioned with respect to the Lonely Planet logo?\".
* If the logo is in front of the sky, the sky is behind the logo.
* Let's check the options: left of, right of, above, below, overlap, inside, contain, in front of, behind, next to.
* \"Behind\" fits perfectly. The sky serves as the background for the logo.
</think>
The sky is behind the Lonely Planet logo.
<answer>behind</answer>

Figure 7: Qualitative example 3: the sky relative to the Lonely Planet logo. Both models output *behind*, but TRIP-Adv reaches the answer with a concise evidence-based explanation. Its heatmap is more concentrated on the logo and its surrounding sky, the regions needed to establish the foreground–background relation.

Model Question	Where is the elephant standing in relation to the net? Please identify their relative position: left of / right of / above / below / overlap / inside / contain / in front of / behind / next to.
Four Intervention Views (Training Only)	 I_0 I_{cf} I_{pos} I_{rand}
Heatmap	 TRIP-Adv (Ours) Qwen3.5-4B
Model Response	**TRIP-Adv (Ours)：** <think> The user wants me to determine the relative position of the elephant with respect to the net. In the image, the net is on the right side. The elephant is on the left side. Therefore, the elephant is to the left of the net. So the option should be \"left of\". </think> <answer>left of</answer> **Qwen3.5-4B：** <think> The user wants to know the position of the elephant relative to the net. 1. **Identify the elephant:** The elephant is the large grey animal in the center foreground, with a rider on its back. 2. **Identify the net:** There is a white net structure, likely a basketball hoop or similar game, positioned to the right of the elephant. 3. **Determine relative position:** The elephant is positioned to the left side of the image, and the net is to the right side of the image. Therefore, the elephant is to the left of the net. </think> The elephant is standing to the left of the net. <answer>left of</answer>

Figure 8: Qualitative example 4: the elephant relative to the net. Both models predict *left of*; however, TRIP-Adv provides a shorter, direct comparison and exhibits a more focused heatmap over the elephant–net pair. This example illustrates improved localization without unnecessary reasoning steps.