SEEING WHILE REASONING: ADVANCING VISUAL FAITHFULNESS IN SMALL MULTIMODAL LANGUAGE MODELS WITH VERIFIABLE PERCEPTION-AWARE INTERMEDIATE PROCESS OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has advanced multimodal chain-of-thought (MCoT) reasoning, yet prevailing objectives primarily optimize answer correctness and textual consistency, providing limited supervision for maintaining visual faithfulness throughout intermediate reasoning. This work analyzes direct text-to-visual attention using **Visual Attention Mass** (**VAM**) and identifies two systematic behaviors. ***Visual Attenuation*** characterizes the abrupt reduction of visual attention when generation transitions from captioning to reasoning. ***Perception Sensitivity*** captures the positive association between stage-level VAM and the correctness of the corresponding intermediate process. Motivated by these findings, we formulate the ***Seeing while Reasoning*** principle and construct **MCoT-ViG**, a structured MCoT dataset pairing four-stage rationales with controlled visual variants and object grounding annotations. We further introduce **Perception-aware Intermediate Process Optimization** (**PaIPO**), which integrates process supervision with stage-specific perception-aware rewards. Specifically, **Caption Perception Reward** (**CPR**) measures caption sensitivity to controlled visual variants, while **Reasoning Grounding Reward** (**RGR**) verifies object references against annotated target regions. Experiments across five benchmarks and three Qwen3-VL model scales demonstrate that **PaIPO** consistently improves multimodal reasoning and visual faithfulness. Compared with the backbones, PaIPO achieves absolute improvements of $+3.53$ to $+6.14$ on VRC-Bench Step Score and increases edited-image accuracy on VFaith-Bench by $+3.96\%$ to $+5.06\%$. Ablations confirm the complementary contributions of CPR and RGR, while analyses of reward dynamics and post-training VAM demonstrate faster reward convergence and more persistent visual attention throughout intermediate reasoning. Code and dataset will be available to facilitate reproducible research.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have demonstrated substantial advances in visual understanding, multimodal reasoning, and complex problem-solving across diverse domains (Lu et al., 2024; Yue et al., 2024; Chen et al., 2024). Multimodal chain-of-thought (MCoT) further extends these capabilities by decomposing visual interpretation and logical deduction into explicit intermediate processes (Xu et al., 2025; Zhang et al., 2024). Reinforcement learning with verifiable rewards (RLVR) has emerged as an effective post-training paradigm for strengthening multi-step reasoning with explicit and verifiable signals. Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) and DAPO (Yu et al., 2025b) improve the efficiency and stability of RLVR, while multimodal extensions adapt this paradigm to vision-language reasoning (Zhang et al., 2025; Liu et al., 2025). However, prevailing objectives prioritize answer correctness or other text-centric criteria, providing limited supervision regarding whether intermediate reasoning remains grounded in visual evidence. Perception-aware approaches have begun to address visual reliance (Ni et al., 2025), but explicit verification of both perceptual descriptions and their subsequent use in reasoning remains underexplored. Consequently, MLLMs sometimes generate plausible rationales that lack sufficient visual supportiveness, producing convincing yet visually unfaithful predictions (Uppaal et al., 2026).

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.

We address this challenge by incorporating verifiable perception and grounding rewards into intermediate process optimization. To support stage-specific analyses and supervision, we adopt a four-stage MCoT structure comprising *Summary*, *Caption*, *Reasoning*, and *Conclusion* (Xu et al., 2025). Within this structure, we define **Visual Attention Mass (VAM)** as a token-level diagnostic that aggregates the attention allocated by each generated text token to all visual tokens. Our analyses reveal two systematic patterns (Figures 1 and 2). **(1)** ***Visual Attenuation*** denotes the abrupt reduction in visual attention when generation transitions from *Caption* to *Reasoning*. **(2)** ***Perception Sensitivity*** captures the positive association between stage-level VAM and the correctness of the corresponding intermediate stage. Together, these observations expose a critical limitation: *models continue reasoning while attending less to the original visual evidence* (Jia et al., 2026). This motivates the formulation of the ***Seeing while Reasoning*** principle, which calls for continued grounding throughout rationale generation rather than reliance on a single initial perceptual representation.

To operationalize this principle, we construct **MCoT-ViG**, a structured MCoT dataset that pairs four-stage rationales with controlled visual variants and object grounding annotations. Each sample comprises a question, an original image, a controlled visual variant, and a four-stage rationale. The rationale is constructed from the original question and image, while the controlled variant serves as a counterfactual reference for perception supervision. MCoT-ViG concentrates supervision on the *Caption* and *Reasoning* stages, where visual evidence is first described and then propagated into multi-step inference. During *Caption*, paired original and controlled images enable the measurement of caption sensitivity to localized visual transformations. During *Reasoning*, annotated bounding boxes provide explicit grounding targets for objects referenced throughout multi-step inference.

Building on MCoT-ViG, we propose **Perception-aware Intermediate Process Optimization** (**PaIPO**), a reinforcement learning framework that transforms intermediate visual faithfulness into verifiable process-level supervision. **Caption Perception Reward** (**CPR**) quantifies the sensitivity of generated captions to visual evidence by measuring their semantic consistency across the original image and controlled visual variants. **Reasoning Grounding Reward** (**RGR**) evaluates object grounding by matching referenced objects to annotated regions and measuring spatial agreement through the intersection over union (IoU) between predicted and annotated bounding boxes. A complementary **Step Reward** (**SR**) evaluates multi-step inference quality through semantic alignment, ordering consistency, and terminal verification. PaIPO combines these process-level signals with conventional answer and format rewards and can be instantiated with either GRPO or DAPO, thereby separating intermediate supervision from the underlying policy optimization algorithm.

We evaluate **PaIPO** on five benchmarks covering two categories: general-purpose multimodal reasoning (MMMU-Pro, MMStar, and VRC-Bench) and visual faithfulness (VFaith-Bench and SPD-Faith). Experiments are conducted on Qwen3-VL-Instruct models at the 2B, 4B, and 8B scales. Under both GRPO and DAPO optimization, PaIPO_{G} and PaIPO_{D} achieve consistent improvements in multimodal reasoning and visual faithfulness. Both PaIPO variants further outperform the corresponding PAPO variants on VRC-Bench and VFaith-Bench. Ablations confirm the complementary contributions of **CPR** and **RGR**, while analyses of training dynamics and post-training **VAM** indicate faster reward convergence and more sustained visual attention during intermediate reasoning.

Our contributions are summarized as follows:

1. Through **Visual Attention Mass** analysis, we identify two systematic phenomena: ***Visual Attenuation*** and ***Perception Sensitivity***, exposing a progressive decline in direct text-to-visual attention and a positive association between visual attention and intermediate reasoning.
2. We introduce the ***Seeing while Reasoning*** principle and construct **MCoT-ViG**, a structured MCoT dataset that enables stage-specific supervision of visual faithfulness through controlled visual variants and object-level grounding annotations.
3. We propose **PaIPO**, a perception-aware RLVR framework that integrates verifiable process-level visual supervision, including **CPR**, **RGR**, and **SR**, to optimize intermediate visual faithfulness while remaining compatible with different policy optimization algorithms.

2 Visual Attention Mass Analysis

Reliable MCoT reasoning requires models to remain grounded in the original visual evidence as generation unfolds. When later inferences rely increasingly on preceding language tokens rather than

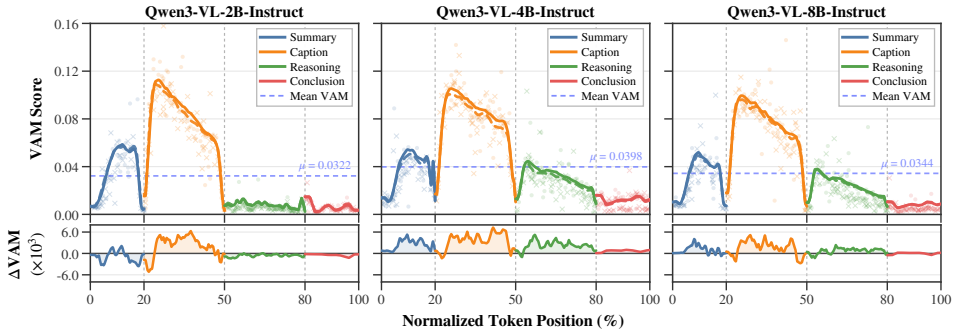


Figure 1: **Direct visual attention concentrates in *Caption* and attenuates thereafter.** The upper panels show token-level VAM observations and smoothed trajectories over normalized completion positions. The lower panels report the position-wise difference, where positive values indicate greater direct visual attention among generations with correct final answers.

re-engaging with the visual input, early perceptual errors may propagate along the reasoning trajectory. To examine how MLLMs allocate direct attention to visual evidence throughout structured rationale generation, we define **Visual Attention Mass** as a token-level diagnostic that quantifies the post-softmax attention mass assigned by each generated text token to visual tokens. Sec. 2.1 formalizes the token-level measure, while Sec. 2.2 examines the stage-wise evolution and association with intermediate reasoning quality. Appendix A defines the corresponding instance-level stage aggregation and provides normalized token-wise trajectories and correctness-conditioned analyses.

2.1 Definition of Visual Attention Mass

Visual Attention Mass (**VAM**) quantifies the proportion of direct attention that each generated text token allocates to the visual input sequence. Let t_i denote the i-th generated text token and $\{v_j\}_{j=1}^{N_v}$ represent the sequence of N_v visual tokens. For Transformer layer ℓ and attention head h, let $\mathbf{A}^{(\ell,h)}$ denote the post-softmax attention matrix. We first average the attention weights across H attention heads, aggregate over all visual tokens, and then average across the L Transformer layers:

$$\mathrm{VAM}(t_i) = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{N_v}\left(\frac{1}{H}\sum_{h=1}^{H} A_{i,j}^{(\ell,h)}\right). \tag{1}$$

The attention matrix $\mathbf{A}^{(\ell,h)}$ is normalized over all unmasked keys. Therefore, for layer ℓ and head h, the sum $\sum_{j=1}^{N_v} A_{i,j}^{(\ell,h)}$ represents the fraction of direct attention allocated by token t_i to visual tokens. Larger values indicate greater direct visual attention allocation. However, attention weights do not establish causal influence. We therefore interpret VAM as an attention-based diagnostic of direct visual-attention allocation rather than as a causal measure of visual grounding.

2.2 Empirical Analysis

We analyze Qwen3-VL-Instruct models under the same four-stage generation protocol. The analyses are organized around two closely related questions: ***Q.1.*** *Do the models sustain attention to visual evidence throughout rationale generation?* ***Q.2.*** *Is visual-attention allocation associated with intermediate reasoning quality?*

Visual Attenuation. Fig. 1 reveals a consistent pattern: direct visual attention concentrates during *Caption* and attenuates thereafter. The transition to *Reasoning* produces another pronounced reduction. Attention then decreases further toward *Conclusion*, where it remains consistently low. We term this systematic post-caption reduction ***Visual Attenuation***. The phenomenon describes a stage-level redistribution of direct attention rather than a strictly monotonic token-wise decay or the disappearance of visual attention. Appendix A.1 provides the corresponding stage-level means and normalized within-stage trajectories, further characterizing this redistribution. The recurrence across scales indicates that increasing model capacity alone does not eliminate the pattern within the evaluated models. **The lower panels** further show predominantly positive VAM differences for correct answers, with the clearest separation during *Caption*. Overall, generations yielding correct final

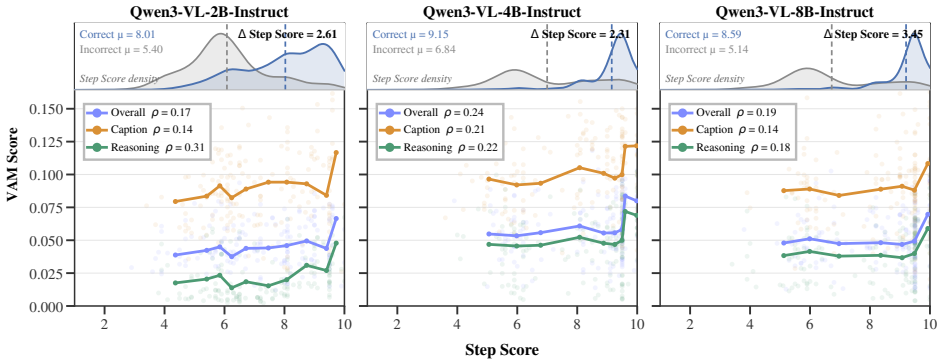


Figure 2: **Direct visual attention is positively associated with intermediate reasoning.** The upper insets compare Step Score distributions between the correct and incorrect answer groups. The lower panels relate Step Score to overall, *Caption*, and *Reasoning* VAM.

answers tend to maintain stronger direct visual attention. We next examine whether this relationship extends beyond answer correctness to the quality of the intermediate reasoning process.

Perception Sensitivity. We next examine whether direct visual attention covaries with intermediate reasoning quality. Following the VRC-Bench protocol, Step Score averages ten rationale-quality criteria across four sampled completions, with higher values indicating more faithful and complete reasoning. **The upper insets** of Figure 2 show a consistent separation between the correct and incorrect answer groups, indicating that final-answer correctness is consistently associated with more faithful and complete intermediate reasoning. **The lower panels** extend this analysis to continuous variation in Step Score. Every Spearman correlation is positive, with coefficients ranging from 0.14 to 0.31. Notably, the *Reasoning*-stage coefficient exceeds its *Caption*-stage counterpart at every model scale. We refer to this consistently positive association between visual-attention allocation and reasoning quality as ***Perception Sensitivity***. The stage-correctness-conditioned trajectories in Appendix A.2 further localize this association across different combinations of *Caption* and *Reasoning* correctness. Although these results are correlational, their consistency motivates explicit visual supervision throughout intermediate reasoning.

3 Visual Grounding MCoT Dataset

The empirical analyses motivate stage-specific supervision that preserves visual grounding beyond the initial perceptual description. Following the *Seeing while Reasoning* principle, we construct **MCoT-ViG**, which pairs four-stage rationales with controlled visual variants and object-level grounding annotations. Fig. 3 summarizes the construction and quality-verification pipeline. Appendix B.1 and B.2 detail data provenance, leakage control, and dataset construction protocols. Appendix B.3 and B.4 present verification criteria, dataset statistics, and quality assessments.

3.1 Dataset Construction

MCoT-ViG provides complementary supervision for the two visually intensive stages of MCoT. Controlled visual variants establish perceptual conditions for measuring sensitivity to image content during *Caption*, while object grounding annotations connect textual references to their supporting visual regions during *Reasoning*. Fig. 3(a) illustrates the complete construction pipeline.

Source Samples. We begin with 50,000 records drawn from five subsets of Visual-CoT (Shao et al., 2024a). Each record is represented as $(q, y, I, \mathcal{B})$, where q denotes the question, y the verified reference answer, I the original image, and $\mathcal{B} = \{(o_m, b_m)\}_{m=1}^{M}$ contains the identities and bounding boxes of the annotated target objects. We preserve records requiring visual evidence, containing valid annotations, and non-overlapping with evaluation benchmarks at both the image and question levels. Source provenance, preprocessing, and benchmark-overlap screening are detailed in Appendix B.1, while the corresponding record-flow statistics are reported in Appendix B.4.1.

Four-stage MCoT Generation. Given a source record x, GPT-4o generates a structured rationale comprising *Summary*, *Caption*, *Reasoning*, and *Conclusion*. *Summary* identifies the task require-

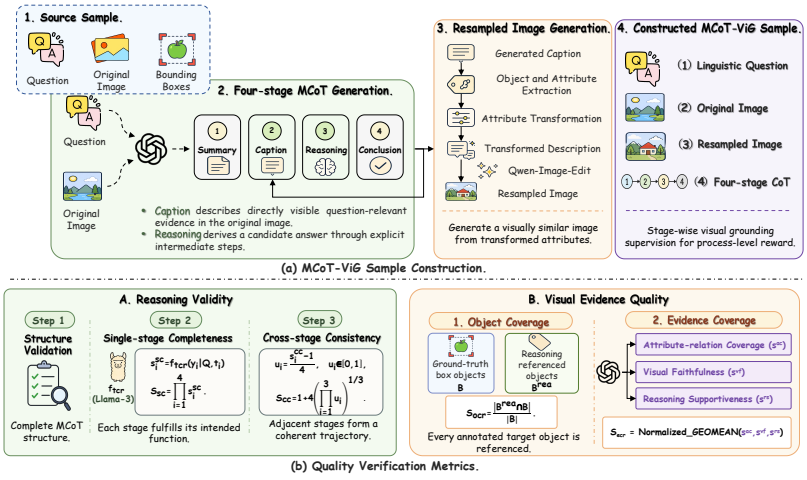


Figure 3: **Overview of the MCoT-ViG construction and quality-verification pipeline. (a)** Source records are converted into four-stage rationales with object grounding annotations and controlled visual variants. **(b)** Candidate records are verified for reasoning validity and visual evidence quality.

ments and outlines the reasoning plan. *Caption* describes question-relevant visual evidence. *Reasoning* derives a candidate answer through explicit intermediate steps, and *Conclusion* reports the final answer in a concise format. Visual references in *Reasoning* are aligned with the source annotations, producing a grounding map $\mathcal{G} = \{(u_k, o_k, b_k)\}_{k=1}^{K}$, where u_k denotes a textual reference to object o_k with bounding box b_k. The first occurrence of an object includes its identifier and complete coordinates, whereas subsequent mentions keep only the identifier to preserve cross-step co-reference. The complete generation prompt and output schema are provided in Appendix B.2.1.

Resampled Image Construction. Using the generated *Caption* and grounding map $\mathcal{G}$, GPT-4o identifies explicitly stated visual attributes and constructs a localized edit. For each annotated target o_k, the model replaces exactly one eligible attribute a_k with a perceptually distinct yet plausible alternative a'_k, defining the localized edit $e_k = (o_k, a_k \rightarrow a'_k, b_k)$. For records containing multiple annotated targets, this procedure produces target-specific edits for all targets. For each transformable record, Qwen-Image-Edit receives the original image I and localized editing instructions constructed from the attribute-transformation outputs. The model modifies only the specified attributes within the designated target regions while preserving the unselected target attributes and non-target content. The resulting image I^{rs} is paired with I as its controlled visual variant. The attribute transformation and image generation protocols are detailed in Appendices B.2.2 and B.2.3.

3.2 Multi-Criteria Quality Verification

Each candidate record is evaluated with a non-compensatory quality verification procedure covering two complementary dimensions: *reasoning validity* and *visual evidence quality*, as summarized in Fig. 3(b). Reasoning validity comprises structural validity, single-stage completeness, and cross-stage consistency. Visual evidence quality comprises complete coverage of annotated target objects and the relevance, faithfulness, and reasoning supportiveness of the reported visual evidence. Structural validity and object coverage are assessed using deterministic validators. A frozen Llama-3 (Grattafiori et al., 2024) verifier evaluates stage-role fulfillment and consistency across the three adjacent stage transitions. GPT-5.1 (OpenAI, 2025) separately evaluates evidence coverage from the original image, question, target-object annotations, and complete generated rationale. The verified reference answer is withheld from this multimodal verifier to reduce outcome-conditioned assessment. For both continuous criteria, the component scores are normalized to $[0, 1]$. Let $\mathcal{H} = \{\mathrm{str}, \mathrm{sc}, \mathrm{ocr}\}$ denote the mandatory criteria and $\mathcal{K} = \{\mathrm{cc}, \mathrm{ecr}\}$ the continuous criteria. The final acceptance rule is

$$\mathrm{Accept}(x_n) = \left[\bigwedge_{h \in \mathcal{H}} S_{h,n} = 1\right] \wedge \left[\bigwedge_{k \in \mathcal{K}} S_{k,n} \geq \tau_k\right]. \tag{2}$$

The continuous thresholds are estimated only from records that satisfy all mandatory criteria. We define this candidate set as $\mathcal{D}_{\mathcal{H}} = \{x_n : \bigwedge_{h\in\mathcal{H}} S_{h,n} = 1\}$. For each continuous criterion, the threshold is $\tau_k = Q_{q^\star}(\{S_{k,n} : x_n \in \mathcal{D}_{\mathcal{H}}\})$, where $Q_{q^\star}(\cdot)$ denotes the empirical $q^\star$-quantile. Complete criterion definitions and calibration are detailed in Appendix B.3. Sample retention statistics and verifier–human alignment analyses are reported in Appendix B.4.

4 Perception-aware Intermediate Process Optimization

Motivated by *Visual Attenuation* and *Perception Sensitivity*, we introduce **Perception-aware Intermediate Process Optimization** (**PaIPO**), which decomposes intermediate process supervision into three complementary rewards. The **Caption Perception Reward** (**CPR**) evaluates the semantic fidelity and counterfactual visual sensitivity of the generated *Caption* using paired original and controlled images. The **Reasoning Grounding Reward** (**RGR**) evaluates whether object references in *Reasoning* remain spatially aligned with annotated target regions and consistent across reasoning steps. The complementary **Step Reward** (**SR**) assesses semantic trajectory alignment, sequential order consistency, and terminal answer entailment. Complete reward definitions and optimizer-specific objectives are provided in Appendix C.

4.1 Problem Formulation and Method Overview

Each training record is represented as $x = (q, y, I, I^{\mathrm{rs}}, \mathcal{E}, \mathcal{G}, \mathbf{o}^\star)$, where q and y denote the question and reference answer, I and I^{rs} are the original image and its counterfactual variant, $\mathcal{E}$ contains the target-specific visual interventions, $\mathcal{G}$ contains the object-level grounding annotations, and $\mathbf{o}^\star$ denotes the reference four-stage rationale. Given a question and image pair (q, I), the rollout policy samples a group of G four-stage MCoT completions:

$$\mathbf{o}_i = (o_i^{\mathrm{sum}}, o_i^{\mathrm{cap}}, o_i^{\mathrm{rea}}, o_i^{\mathrm{con}}) \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid q, I), \qquad i = 1, \ldots, G. \tag{3}$$

Each completion receives three normalized rewards. CPR and RGR provide stage-specific perception supervision, while SR evaluates the semantic validity of the intermediate steps in *Reasoning*.

4.2 Caption Perception Reward

A reliable caption reward should capture both semantic correctness and visual specificity. Reference alignment alone may reward descriptions supported primarily by linguistic priors, while visual discrimination alone may favor distinctive but incomplete statements. *Caption perception reward* therefore combines linguistic reference alignment with counterfactual visual discrimination.

Linguistic Reference Alignment. Let $c_i = o_i^{\mathrm{cap}}$ denote the generated *Caption*, and let $c^\star = o^{\star,\mathrm{cap}}$ denote the reference counterpart. A frozen atomic extractor decomposes both captions into visual propositions, and a frozen textual entailment verifier measures their bidirectional semantic agreement. We define $r_i^{\mathrm{ref}} \in [0, 1]$ as the harmonic mean of atom-level candidate precision and reference coverage. This construction penalizes both unsupported content and omitted reference evidence while remaining tolerant of valid paraphrases. Complete formulations are detailed in Appendix C.1.

Counterfactual Visual Discrimination. Reference alignment cannot determine whether the generated caption is grounded in the original visual evidence. We therefore evaluate the identical generated caption under the original image I and controlled visual variant I^{rs}. Let $\mathcal{T}_i^{\mathrm{cap}}$ denote the token positions of the generated *Caption*. With a frozen rollout-policy $\pi_{\bar{\theta}}$, we teacher-force the sampled completion under both visual conditions while keeping the textual sequence unchanged. For each caption token, we define

$$\delta_{i,t} = \log \frac{\pi_{\bar{\theta}}(o_{i,t} \mid q, I, o_{i,<t})}{\pi_{\bar{\theta}}(o_{i,t} \mid q, I^{\mathrm{rs}}, o_{i,<t})}, \qquad e_{i,t} = \tanh\left(\frac{\delta_{i,t}}{2\tau_{\mathrm{cf}}}\right), \qquad t \in \mathcal{T}_i^{\mathrm{cap}}, \tag{4}$$

where $\tau_{\mathrm{cf}} > 0$ controls sensitivity to the counterfactual likelihood difference. Positive $e_{i,t}$ indicates stronger support from the original image, while values near zero indicate visual invariance. Each image pair differs through localized target-specific interventions and most caption tokens should remain invariant. We emphasize visually discriminative tokens through the weighted aggregation:

$$D_i^{\mathrm{cf}} = \frac{\sum_{t\in\mathcal{T}_i^{\mathrm{cap}}} |e_{i,t}|^\gamma e_{i,t}}{\sum_{t\in\mathcal{T}_i^{\mathrm{cap}}} |e_{i,t}|^\gamma + \epsilon}, \qquad r_i^{\mathrm{vis}} = \left[\frac{D_i^{\mathrm{cf}} - m_{\mathrm{cf}}}{1 - m_{\mathrm{cf}}}\right]_{[0,1]}, \tag{5}$$

where $\gamma \geq 1$ emphasizes tokens with stronger counterfactual evidence and $\epsilon > 0$ ensures numerical stability. Here, $[z]_{[0,1]} = \min\{1, \max\{0, z\}\}$, and $m_{\text{cf}} \in [0, 1)$ is a calibrated indifference margin. The margin discounts the intrinsic preference for I induced by sampling the rollout under the original visual condition. Full computational details are provided in Appendix C.1.

Caption Perception Reward. We treat semantic fidelity and counterfactual visual sensitivity as complementary requirements and combine them using the harmonic mean:

$$r_i^{\text{CPR}} = \frac{2 r_i^{\text{ref}} r_i^{\text{vis}}}{r_i^{\text{ref}} + r_i^{\text{vis}} + \epsilon}. \tag{6}$$

A high CPR therefore requires the generated caption to align with the reference semantics without unsupported content while remaining preferentially supported by the original image.

4.3 Reasoning Grounding Reward

CPR encourages visually sensitive descriptions but does not guarantee that subsequent *Reasoning* remains grounded in the relevant image regions. We therefore introduce *Reasoning grounding reward* (RGR) to evaluate object-level spatial grounding and cross-step reference consistency. Let $\widehat{\mathcal{G}}_i = \left\{(\widehat{u}_{i,j}, \widehat{o}_{i,j}, \widehat{b}_{i,j})\right\}_{j=1}^{\widehat{K}_i}$ denote the grounding records parsed from completion $\mathbf{o}_i$. For predicted reference j and annotated target k, we define the grounding score through their spatial overlap:

$$\omega_{i,j,k} = \text{IoU}\left(\widehat{b}_{i,j}, b_k\right) = \frac{\left|\widehat{b}_{i,j} \cap b_k\right|}{\left|\widehat{b}_{i,j} \cup b_k\right|}. \tag{7}$$

Therefore, $\omega_{i,j,k} \in [0, 1]$ measures spatial agreement between the reference and annotated boxes. Maximum-weight one-to-one matching between predicted and annotated objects produces a localization weighted grounding F-score $F_i^{\text{grd}} \in [0, 1]$. Missing targets reduce recall, unsupported predictions reduce precision, and inaccurate regions reduce both through their IoU values. Let $C_i^{\text{id}} \in [0, 1]$ denote the consistency of subsequent references with the object identifiers established at their first coordinate-bearing occurrences. The final reward is

$$r_i^{\text{RGR}} = F_i^{\text{grd}} \left[(1 - \lambda_{\text{id}}) + \lambda_{\text{id}} C_i^{\text{id}}\right], \qquad \lambda_{\text{id}} \in [0, 1]. \tag{8}$$

This multiplicative formulation allows identifier consistency to refine successful grounding while preventing it from producing a positive reward when spatial grounding fails. The matching algorithm, weighted F-score, and identifier-consistency calculation are detailed in Appendix C.2.

4.4 Unified Reward and Policy Optimization

The **Step Reward** (SR) complements CPR and RGR by evaluating semantic trajectory alignment, sequential order consistency, and terminal answer entailment. A frozen textual entailment verifier aligns the generated and reference reasoning steps. A smoothed weighted geometric aggregation produces $r_i^{\text{SR}} \in [0, 1]$. The complete formulation is provided in Appendix C.3.

Unified Process Reward. Since $r_i^{\text{CPR}}, r_i^{\text{RGR}}, r_i^{\text{SR}} \in [0, 1]$ capture complementary dimensions of intermediate-process quality, with $\mathcal{D}_{\text{proc}} = \{\text{CPR}, \text{RGR}, \text{SR}\}$, the unified process reward is

$$\widetilde{r}_i^d = \delta_{\text{proc}} + (1 - \delta_{\text{proc}}) r_i^d, \quad r_i^{\text{proc}} = \frac{\prod_{d \in \mathcal{D}_{\text{proc}}} \left(\widetilde{r}_i^d\right)^{\alpha_d} - \delta_{\text{proc}}}{1 - \delta_{\text{proc}}}, \quad \alpha_d > 0, \quad \sum_{d \in \mathcal{D}_{\text{proc}}} \alpha_d = 1. \tag{9}$$

Here, $0 < \delta_{\text{proc}} \ll 1$ prevents a single zero-valued component from collapsing the reward. The geometric aggregation favors consistently strong intermediate processes while limiting compensation among the three dimensions. Let $r_i^{\text{ans}} \in \{0, 1\}$ denote answer correctness[1] and $r_i^{\text{fmt}} \in \{0, 1\}$ denote four-stage format compliance. For optimizer $\mathcal{B} \in \{\text{G}, \text{D}\}$, the rationale reward is

$$R_i^{(\mathcal{B})} = \lambda_{\text{ans}} r_i^{\text{ans}} + \lambda_{\text{fmt}} r_i^{\text{fmt}} + \lambda_{\text{proc}} r_i^{\text{proc}} + r_{i,\text{aux}}^{(\mathcal{B})}, \tag{10}$$

[1] Full formulation of answer correctness is detailed in Appendix C.4

Table 1: **Main results on general-purpose multimodal reasoning and visual faithfulness across Qwen3-VL model scales.** Each result is reported as $\mathrm{mean}^{\mathrm{std}}$. $\Delta_{\mathrm{pp}} = \mathrm{Edit\ Acc.} - \mathrm{Raw\ Acc.}$. †: reimplementation of PAPO under the matched training and evaluation settings. Implementation details are provided in Appendix D.1. **Bold** and underlined indicate the best and second-best results.

Method	General-purpose Multi-modal Reasoning					Visual Faithfulness Evaluation						
	MMMU^{Pro}	MMStar	VRC-Bench			VFaith-Bench			SPD-Faith			
	10 Opts. Acc.	*Acc.*	*Ans. Acc.*	*Step Score*	*Δ Step ↑*	*Raw Acc.*	*Edit Acc.*	Δ_{pp}	*CR*	*DRF*	*Avg.*	*Δ Avg. ↑*
Qwen3-VL-2B	$29.11^{0.68}$	$55.98^{0.19}$	$56.16^{0.52}$	$69.14^{1.53}$	–	$57.15^{0.37}$	$51.22^{0.26}$	$-5.93^{0.14}$	$42.30^{1.31}$	$13.25^{1.57}$	$27.78^{1.02}$	–
w/ Step GRPO	$30.75^{0.84}$	$56.28^{0.82}$	$56.67^{0.96}$	$70.87^{1.28}$	+ 1.73	$57.50^{0.62}$	$51.42^{1.20}$	$-6.08^{0.76}$	$43.82^{1.83}$	$14.91^{0.83}$	$29.36^{1.21}$	+ 1.59
w/ Step DAPO	$31.28^{0.65}$	$57.08^{0.79}$	$57.76^{0.72}$	$71.74^{1.01}$	+ 2.60	$57.79^{0.70}$	$52.35^{0.73}$	$-5.43^{0.97}$	$44.79^{1.29}$	$14.74^{1.18}$	$29.76^{1.17}$	+ 1.99
w/ PAPO$^{\dagger}_{G}$	$31.42^{0.78}$	$57.33^{0.76}$	$58.34^{0.86}$	$72.47^{0.93}$	+ 3.33	$58.62^{0.42}$	$54.15^{0.74}$	$-4.47^{0.58}$	$45.39^{1.24}$	$15.70^{0.87}$	$30.55^{0.68}$	+2.77
w/ PAPO$^{\dagger}_{D}$	$31.61^{0.85}$	$58.71^{0.81}$	$58.67^{0.91}$	$73.09^{1.02}$	+ 3.95	$59.10^{0.73}$	$54.41^{0.96}$	$-4.69^{0.71}$	$46.33^{1.34}$	$16.66^{1.22}$	$31.49^{0.89}$	+ 3.72
w/ PaIPO$_G$	$32.45^{0.72}$	$58.50^{0.76}$	$\mathbf{59.36}^{0.82}$	$74.86^{1.05}$	+5.72	$59.73^{0.93}$	$55.92^{1.18}$	$-3.81^{0.85}$	$\mathbf{48.86}^{1.65}$	$17.50^{1.13}$	$33.18^{1.17}$	+5.41
w/ PaIPO$_D$	$\mathbf{32.82}^{0.68}$	$\mathbf{60.07}^{0.85}$	$59.24^{0.88}$	$\mathbf{75.28}^{0.98}$	**+ 6.14**	$\mathbf{59.93}^{0.62}$	$\mathbf{56.28}^{0.95}$	$\mathbf{-3.65}^{0.64}$	$48.80^{1.21}$	$\mathbf{17.66}^{0.57}$	$\mathbf{33.23}^{0.79}$	**+ 5.46**
Qwen3-VL-4B	$39.67^{0.49}$	$67.96^{0.51}$	$63.37^{0.67}$	$81.18^{0.82}$	–	$65.79^{0.49}$	$61.37^{0.77}$	$-4.42^{0.32}$	$62.41^{1.07}$	$17.02^{1.24}$	$39.72^{1.14}$	–
w/ Step GRPO	$41.22^{0.66}$	$68.03^{0.64}$	$64.02^{0.72}$	$81.41^{0.89}$	+ 0.23	$66.08^{0.46}$	$61.90^{0.93}$	$-4.18^{0.68}$	$63.43^{0.94}$	$17.54^{0.98}$	$40.48^{0.82}$	+ 0.77
w/ Step DAPO	$41.87^{0.63}$	$68.47^{0.57}$	$64.43^{0.64}$	$82.03^{0.73}$	+ 0.85	$66.77^{0.57}$	$62.49^{0.89}$	$-4.29^{0.62}$	$64.56^{0.80}$	$18.35^{0.76}$	$41.45^{0.70}$	+ 1.74
w/ PAPO$^{\dagger}_{G}$	$41.74^{0.69}$	$69.86^{0.64}$	$64.57^{0.69}$	$83.16^{0.81}$	+ 1.98	$67.04^{0.67}$	$63.73^{0.75}$	$-3.31^{0.57}$	$65.46^{1.12}$	$19.29^{1.01}$	$42.37^{0.98}$	+ 2.66
w/ PAPO$^{\dagger}_{D}$	$41.96^{0.66}$	$70.15^{0.65}$	$64.91^{0.74}$	$83.52^{0.85}$	+ 2.34	$67.34^{0.69}$	$64.07^{0.78}$	$-3.27^{0.55}$	$65.40^{0.89}$	$19.26^{0.99}$	$42.33^{0.77}$	+ 2.62
w/ PaIPO$_G$	$42.14^{0.66}$	$71.07^{0.57}$	$66.53^{0.66}$	$85.47^{0.79}$	+4.29	$68.21^{0.45}$	$65.33^{0.21}$	$-2.88^{0.31}$	$67.08^{0.74}$	$20.84^{0.82}$	$43.96^{0.57}$	+4.24
w/ PaIPO$_D$	$\mathbf{42.31}^{0.63}$	$\mathbf{71.27}^{0.63}$	$\mathbf{66.72}^{0.71}$	$\mathbf{85.71}^{0.77}$	**+ 4.53**	$\mathbf{68.47}^{0.44}$	$\mathbf{65.69}^{0.56}$	$\mathbf{-2.78}^{0.64}$	$\mathbf{67.89}^{1.04}$	$\mathbf{21.16}^{0.85}$	$\mathbf{44.53}^{0.78}$	**+ 4.79**
Qwen3-VL-8B	$46.94^{0.21}$	$67.18^{0.23}$	$64.50^{0.82}$	$83.22^{0.61}$	–	$66.38^{0.38}$	$60.10^{0.17}$	$-6.29^{0.27}$	$65.28^{1.14}$	$16.63^{1.07}$	$40.96^{1.04}$	–
w/ Step GRPO	$48.14^{0.53}$	$67.42^{0.56}$	$65.83^{0.62}$	$84.93^{0.74}$	+ 1.71	$66.70^{0.56}$	$60.27^{0.78}$	$-6.43^{0.63}$	$66.17^{1.13}$	$17.48^{0.65}$	$41.83^{0.74}$	+ 0.87
w/ Step DAPO	$48.72^{0.53}$	$68.13^{0.47}$	$66.21^{0.54}$	$85.43^{0.65}$	+ 2.21	$66.92^{0.36}$	$60.58^{0.73}$	$-6.34^{0.49}$	$66.15^{0.81}$	$17.24^{0.47}$	$41.70^{0.52}$	+ 0.74
w/ PAPO$^{\dagger}_{G}$	$48.66^{0.57}$	$69.01^{0.55}$	$66.29^{0.59}$	$85.36^{0.70}$	+ 2.14	$67.65^{0.44}$	$62.45^{0.69}$	$-5.21^{0.84}$	$67.35^{0.83}$	$18.28^{0.62}$	$42.81^{0.52}$	+ 1.86
w/ PAPO$^{\dagger}_{D}$	$49.03^{0.57}$	$69.18^{0.59}$	$66.47^{0.64}$	$85.66^{0.73}$	+ 2.44	$67.73^{0.53}$	$62.60^{0.71}$	$-5.13^{0.52}$	$68.13^{1.01}$	$19.04^{0.91}$	$43.58^{0.75}$	+ 2.63
w/ PaIPO$_G$	$49.48^{0.54}$	$70.27^{0.53}$	$\mathbf{67.37}^{0.58}$	$86.75^{0.68}$	+3.53	$\mathbf{68.43}^{0.41}$	$64.41^{0.72}$	$-4.02^{0.44}$	$69.72^{0.75}$	$20.31^{1.03}$	$45.02^{0.67}$	+4.06
w/ PaIPO$_D$	$\mathbf{49.87}^{0.56}$	$\mathbf{71.35}^{0.57}$	$67.19^{0.61}$	$\mathbf{87.02}^{0.72}$	**+ 3.80**	$68.13^{0.49}$	$\mathbf{64.44}^{0.54}$	$\mathbf{-3.69}^{0.35}$	$\mathbf{69.77}^{0.88}$	$\mathbf{20.44}^{0.84}$	$\mathbf{45.11}^{0.71}$	**+ 4.15**

where $r^{(\mathrm{G})}_{i,\mathrm{aux}} = 0$, and $r^{(\mathrm{D})}_{i,\mathrm{aux}} = \lambda_{\mathrm{len}} r^{\mathrm{len}}_i$ denotes the DAPO overlong-response penalty. An incorrect answer forfeits the answer reward but retains process reward according to the intermediate quality. Terminal answer entailment within SR penalizes reasoning that fails to support the verified answer.

Policy Optimization. PaIPO uses $R^{(\mathcal{B})}_i$ to estimate group-relative advantages while retaining the objective of the selected base optimizer. Accordingly, $\mathrm{PaIPO_G}$ follows GRPO, whereas $\mathrm{PaIPO_D}$ follows DAPO, including optimizer-specific clipping, sampling, and loss-reduction strategies. PaIPO therefore determines *which* intermediate perceptual and reasoning behaviors receive credit, while the underlying optimizer determines *how* these preferences are converted into policy updates. Complete advantage estimation and optimization objectives are provided in Appendix C.4.

5 Experiments

We evaluate PaIPO across five benchmarks and three Qwen3-VL model scales to comprehensively assess the effectiveness in multimodal reasoning and visual faithfulness. Sec. 5.1 describes the experimental setup, Sec. 5.2 reports the main results, Sec. 5.3 analyzes the individual and complementary contributions of the proposed rewards, and Sec. 5.4 examines reward dynamics and post-training visual attention. Complete configurations and auxiliary analyses are provided in Appendix D.

5.1 Experimental Setup

Models and Training. We evaluate PaIPO on Qwen3-VL-Instruct models at the 2B, 4B, and 8B scales. We instantiate PaIPO with both GRPO and DAPO, denoted as $\mathrm{PaIPO_G}$ and $\mathrm{PaIPO_D}$. The baselines comprise the original backbones, Step GRPO, Step DAPO, and the corresponding PAPO variants (Wang et al., 2026). All post-training methods use MCoT-ViG, $G = 4$ rollouts per prompt, LoRA adaptation (Hu et al., 2022), and matched training and generation budgets. We evaluate general-purpose reasoning on MMMU-Pro, MMStar, and VRC-Bench, and visual faithfulness on VFaith-Bench and SPD-Faith. Full experimental setup details are provided in Appendix D.1.

5.2 Main Results

General-purpose Multimodal Reasoning. Table 1 demonstrates the consistent improvements in general multimodal reasoning across all three model scales and both optimizers. On VRC-Bench,

Table 2: **Ablation study of the proposed reward components.** Results are reported as $\mathrm{mean}^{\mathrm{std}}$.

Baseline	Variants			VRC-Bench			VFaith-Bench					
	SR	*CPR*	*RGR*	*Ans. Acc.* ↑	*Step Score* ↑	Δ *Step* ↑	*Raw Acc.* ↑	*Edit Acc.* ↑	Δ_{pp} ↑	*Repeat* ↓	*Perception* ↑	Δ *Perc.* ↑
Qwen3-VL-2B	–	–	–	$56.16^{0.52}$	$69.14^{1.53}$	–	$57.15^{0.37}$	$51.22^{0.26}$	$-5.93^{0.14}$	$54.60^{0.70}$	$14.30^{1.10}$	–
	✓	–	–	$57.76^{0.72}$	$71.74^{1.01}$	+ 2.60	$57.79^{0.70}$	$52.35^{0.73}$	$-5.43^{0.97}$	$54.02^{1.03}$	$14.91^{1.20}$	+ 0.61
	✓	✓	–	$58.32^{0.84}$	$73.02^{1.03}$	+ 3.88	$58.62^{0.68}$	$54.13^{0.85}$	$-4.49^{0.62}$	$51.16^{0.94}$	$17.48^{1.24}$	+ 3.18
	✓	–	✓	$58.46^{0.86}$	$73.19^{1.08}$	+ 4.05	$58.71^{0.67}$	$54.06^{0.90}$	$-4.65^{0.67}$	$51.68^{1.02}$	$17.02^{1.15}$	+ 2.72
	–	✓	✓	$58.81^{0.87}$	$74.16^{1.01}$	$\underline{+5.02}$	$59.54^{0.60}$	$55.66^{0.89}$	$-3.88^{0.59}$	$48.31^{0.92}$	$19.02^{1.18}$	$\underline{+4.72}$
	✓	✓	✓	$59.24^{0.88}$	$75.28^{0.98}$	**+ 6.14**	$59.93^{0.62}$	$56.28^{0.95}$	$-3.65^{0.64}$	$47.58^{0.82}$	$19.18^{1.07}$	**+ 4.88**
Qwen3-VL-4B	–	–	–	$63.37^{0.67}$	$81.18^{0.82}$	–	$65.79^{0.49}$	$61.37^{0.77}$	$-4.42^{0.32}$	$65.40^{0.20}$	$26.40^{0.20}$	–
	✓	–	–	$64.43^{0.64}$	$82.03^{0.73}$	+ 0.85	$66.77^{0.57}$	$62.49^{0.89}$	$-4.29^{0.62}$	$64.51^{0.81}$	$26.72^{0.87}$	+ 0.32
	✓	✓	–	$65.67^{0.72}$	$83.52^{0.85}$	+ 2.34	$67.43^{0.55}$	$63.91^{0.72}$	$-3.52^{0.50}$	$60.92^{0.83}$	$29.04^{0.91}$	+ 2.64
	✓	–	✓	$65.84^{0.75}$	$83.68^{0.88}$	+ 2.50	$67.56^{0.52}$	$63.82^{0.70}$	$-3.74^{0.56}$	$61.28^{0.86}$	$28.73^{0.97}$	+ 2.33
	–	✓	✓	$66.35^{0.74}$	$85.03^{0.81}$	$\underline{+3.85}$	$68.12^{0.48}$	$65.16^{0.63}$	$-2.96^{0.46}$	$57.02^{0.76}$	$31.27^{0.86}$	$\underline{+4.87}$
	✓	✓	✓	$66.72^{0.71}$	$85.71^{0.77}$	**+ 4.53**	$68.47^{0.44}$	$65.69^{0.56}$	$-2.78^{0.64}$	$56.07^{0.58}$	$31.65^{0.73}$	**+ 5.25**

$\mathrm{PaIPO_G}$ and $\mathrm{PaIPO_D}$ improve Step Score over their backbones by 3.53 to 5.72 and 3.80 to 6.14. Against the matched PAPO variants, the corresponding improvements are 1.39 to 2.39 under GRPO and 1.36 to 2.19 under DAPO. Concurrent improvements in answer accuracy indicate that PaIPO improves intermediate reasoning while simultaneously improving final-answer performance.

Visual Faithfulness. Compared with Step DAPO, $\mathrm{PaIPO_D}$ improves edited-image accuracy by 3.20 to 3.93, while exceeding $\mathrm{PAPO_D}$ by 1.62 to 1.87. It also reduces the raw-to-edited accuracy distance from 4.42 to 6.29 for the backbones to 2.78 to 3.69. On SPD-Faith, $\mathrm{PaIPO_D}$ improves the average score over the backbones by 4.15 to 5.46. These results show that PaIPO improves robustness to controlled visual transitions and strengthens the faithfulness of multimodal reasoning.

5.3 Ablation Study

Contribution of Individual Rewards. Table 2 quantifies the contribution of each reward component on the 2B and 4B models. SR alone improves Step Score by $+2.60$ and $+0.85$ over the backbones, while producing modest gains of $+0.61$ and $+0.32$ in VFaith-Bench Perception. With SR constant, introducing either perception reward leads to substantially larger improvements in both reasoning quality and visual faithfulness. Overall, intermediate reasoning quality obtains more benefit from RGR, while CPR contributes more strongly to visual-faithfulness measures.

Complementarity of Perception Rewards. Without SR, combining CPR and RGR increases Step Score by $+5.02$ and $+3.85$ and improves VFaith-Bench Perception by $+4.72$ and $+4.87$. This combination outperforms both SR+CPR and SR+RGR, while the full configuration performs best. The controlled comparison indicates that CPR and RGR provide complementary perceptual supervision, while SR supplies an additional reasoning-oriented benefit.

5.4 Further Analysis

Reward Dynamics. Fig. 7 in Appendix D.2 compares $\mathrm{PaIPO_G}$ and $\mathrm{PaIPO_D}$ with the matched Step GRPO and Step DAPO baselines. The *Caption* and *Reasoning* rewards increase more rapidly and stabilize earlier with PaIPO, indicating faster empirical convergence under both optimizers.

Post-training Visual Attention. Fig. 8 in Appendix D.3 compares the post-training VAM dynamics of PaIPO with the baselines. PaIPO maintains higher visual attention during *Caption* and *Reasoning* and exhibits a smaller attenuation between these stages, indicating alleviated *Visual Attenuation*.

6 Conclusion

We present **PaIPO**, a perception-aware process optimization approach to improve visual faithfulness in structured MCoT rationale generation. Motivated by the ***Visual Attenuation*** and ***Perception Sensitivity***, we formulate the ***Seeing while Reasoning*** paradigm and construct **MCoT-ViG**, which pairs four-stage rationales with controlled visual variants and object grounding annotations. Experiments across reasoning and faithfulness benchmarks demonstrate that **PaIPO** consistently improves both multimodal reasoning and the visual faithfulness of generated rationales. However, VAM is an attention-based correlational diagnostic rather than a causal measure, and latent factors may jointly influence attention allocation and reasoning quality. Identifying the potential causal relationships between visual attention and intermediate reasoning remains an important direction for future work.

AI USE STATEMENT

Generative AI tools were used to support the construction and verification of MCoT-ViG. GPT-4o generated candidate four-stage rationales, object-level grounding records, and attribute transformations. Qwen-Image-Edit generated controlled visual variants, and frozen Llama-3 and GPT-5.1 models served as quality verifiers. Candidate records were evaluated using the pipeline described in Sections 3.1 and 3.2 and Appendix B. A stratified sample was independently assessed by three domain experts to evaluate verifier–human alignment. Generative AI was not used to design PaIPO or the experimental protocol, analyze the primary results, or formulate the conclusions. The authors reviewed all AI-assisted work and take full responsibility for the manuscript, claims, data artifacts, and reported results.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the main text provides formal definitions of VAM and PaIPO and summarizes the MCoT-ViG construction and evaluation protocols. Appendix A provides additional direct visual attention analyses. Appendix B documents data provenance, construction prompts, and quality verification, while Appendix C provides the complete reward derivations and optimization objectives. Experimental settings and hyperparameters are reported in Appendix D.1. We will publicly release the implementations for data construction, training, evaluation, and VAM analysis, together with MCoT-ViG and its grounding annotations.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37 of *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2024. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Hongrui Jia, Chaoya Jiang, Shikun Zhang, and Wei Ye. Mitigating visual context degradation in large multimodal models: A training-free decoupled agentic framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9574–9585, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Jia_Mitigating_Visual_Context_Degradation_in_Large_Multimodal_Models_A_Training-Free_CVPRF_2026_paper.html.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981, 2020. doi: 10.1007/s11263-020-01316-z.

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. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Visual-RFT_Visual_Reinforcement_Fine-Tuning_ICCV_2025_paper.html.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen, and Bo Chen. SPD-faith bench: Diagnosing and improving faithfulness in chain-of-thought for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19875–19927, 2026.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-RFT: Improving multimodal reasoning with visually grounded reinforcement finetuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23334–23355, 2025. doi: 10.52202/085713-0692. URL https://papers.nips.cc/paper_files/paper/2025/hash/1d83ad88759cef8192451543e5d59bf6-Abstract-Conference.html.

OpenAI. GPT-4o System Card. Technical report, OpenAI, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum. Technical report, OpenAI, November 2025. URL https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37 of *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

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*, 2024b. URL https://arxiv.org/abs/2402.03300.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, 2025.

Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. Journey before destination: On the importance of visual faithfulness in slow thinking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4147–4168. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.eacl-long.194/.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=izbBqTL8vb.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098. IEEE/CVF, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_LLaVA-CoT_Let_Vision_Language_Models_Reason_Step-by-Step_ICCV_2025_paper.html.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Jiachen Yu, Yufei Zhan, Ziheng Wu, Yousong Zhu, Jinqiao Wang, and Minghui Qiu. Vfaith: Do large multimodal models really reason on seen images rather than previous memories? *arXiv preprint arXiv:2506.11571*, 2025a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=2a36EMSSTp.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567. IEEE/CVF, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE/CVF, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Zhang_R1-VL_Learning_to_Reason_with_Multimodal_Large_Language_Models_via_ICCV_2025_paper.pdf.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=y1pPWFVfvR.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4995–5004, 2016.

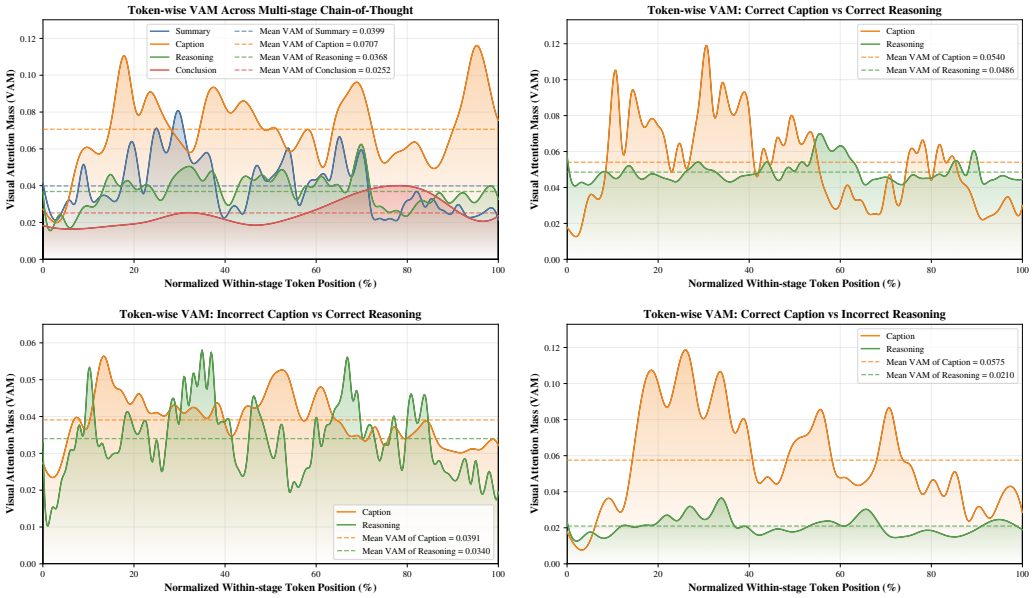


Figure 4: **Token-wise VAM profiles across MCoT stages and under different *Caption* and *Reasoning* correctness conditions.** All trajectories are plotted against normalized within-stage token positions. **Top-left:** the average trajectory over all evaluated records. **Top-right:** records with correct *Caption* and correct *Reasoning*. **Bottom-left:** records with incorrect *Caption* but correct *Reasoning*. **Bottom-right:** records with correct *Caption* but incorrect *Reasoning*. Records with correct reasoning maintain comparatively stronger reasoning-stage VAM, while the correct-caption/incorrect-reasoning group exhibits a pronounced reduction after entering *Reasoning*.

A Visual Attention Mass Analysis Details

This appendix presents additional empirical evidence supporting the *Visual Attenuation* and *Perception Sensitivity* introduced in Section 2.2. Figure 1 characterizes the stage-wise evolution of direct visual attention across model scales, while Figure 2 quantifies its association with intermediate reasoning quality. Figure 4 complements these results with token-wise VAM profiles under three *Caption* and *Reasoning* correctness conditions.

Formally, for a rationale instance x_n, let $\mathcal{T}_{n,s}$ denote the set of generated tokens belonging to stage $s \in \{\text{sum}, \text{cap}, \text{rea}, \text{con}\}$, corresponding to *Summary*, *Caption*, *Reasoning*, and *Conclusion*, respectively. We define the instance-level VAM for stage s as

$$\overline{V}_{n,s} = \frac{1}{|\mathcal{T}_{n,s}|} \sum_{t \in \mathcal{T}_{n,s}} \text{VAM}(t), \tag{11}$$

where $\text{VAM}(t)$ denotes the token-level VAM of token t in instance x_n, as defined in Equation 1. In each panel of Figure 4, the horizontal dashed lines report the mean of $\overline{V}_{n,s}$ over all records in the corresponding correctness group. Figure 4 normalizes token positions separately within each stage. Each stage is mapped to a common interval from 0% to 100%, enabling comparisons across stages and rationales of different lengths. The token-wise trajectories are smoothed only for visualization, while all reported stage means are computed from the original VAM values.

A.1 Visual Attenuation

Figure 1 shows that direct visual attention concentrates in *Caption* and decreases during the subsequent stages across all three model scales. **The top-left panel** of Figure 4 provides a complementary

within-stage view of this pattern. The mean VAM is 0.0707 during *Caption*, compared with 0.0399 during *Summary*, 0.0368 during *Reasoning*, and 0.0252 during *Conclusion*. The *Caption* trajectory also contains the largest and most frequent VAM peaks, whereas *Reasoning* remains within a lower range and *Conclusion* maintains the lowest mean. Visual Attenuation is not strictly monotonic at the individual-token level. Several localized peaks remain during *Reasoning*, indicating that direct attention to visual tokens does not disappear after caption generation. These fluctuations nevertheless occur around a substantially lower stage mean and do not reverse the overall post-caption reduction. Thus, *Visual Attenuation* describes a systematic stage-level redistribution of direct visual attention.

A.2 Perception Sensitivity

Figure 2 shows consistently positive associations between VAM value and intermediate reasoning quality. For the 2B, 4B, and 8B models, the Spearman correlations between Step Score and overall VAM are 0.17, 0.24, and 0.19, respectively. The corresponding coefficients are 0.14, 0.21, and 0.14 for *Caption*-stage VAM, and 0.31, 0.22, and 0.18 for *Reasoning*-stage VAM. All nine coefficients are positive, and the *Reasoning*-stage coefficient exceeds its *Caption*-stage counterpart at every model scale. This consistent positive relationship between visual-attention allocation and intermediate reasoning quality constitutes the *Perception Sensitivity* observed in the main analysis.

The correctness-conditioned panels in Figure 4 further localize this association. When both *Caption* and *Reasoning* are correct, their mean VAM values are 0.0540 and 0.0486, respectively. The comparatively small difference is consistent with relatively stable direct visual attention across the transition from captioning to reasoning. A similar transition is observed when *Caption* is incorrect but *Reasoning* is correct. In this group, mean VAM is 0.0391 during *Caption* and 0.0340 during *Reasoning*. Although the absolute attention level is lower, the reasoning-stage trajectory retains several localized increases, and the reduction between the two stages remains comparatively limited. Thus, correct reasoning can remain associated with continued direct visual attention even when the preceding caption is inaccurate. The reverse condition exhibits a substantially different profile. When *Caption* is correct but *Reasoning* is incorrect, mean VAM decreases from 0.0575 during captioning to 0.0210 during reasoning. Strong direct visual attention during *Caption* alone is therefore insufficient to characterize a correct subsequent reasoning process. Instead, reasoning correctness is associated with retaining direct visual attention after the initial visual description has been generated. Together, Figures 2 and 4 provide complementary evidence for *Perception Sensitivity*. The former quantifies the positive association between VAM and Step Score across model scales, while the latter localizes this association to the transition from visual description to reasoning under different stage-correctness conditions. These results remain correlational because attention weights do not establish causal visual influence. We therefore interpret VAM as a diagnostic of direct visual-attention allocation rather than a causal measure of visual grounding. Nevertheless, the consistency of the stage-level, token-wise, and correctness-conditioned observations motivates explicit visual supervision throughout intermediate reasoning.

B Visual Grounding MCoT Dataset

This appendix provides detailed construction, verification, and characterization protocols for **MCoT-ViG**. We first describe the source data, preprocessing criteria, bounding-box validation and normalization, and checks for overlap with the evaluation benchmarks in Section B.1. We then provide the complete prompts and output schemas for four-stage rationale generation, object-level grounding annotation, attribute transformation, and controlled visual-variant construction in Section B.2. Next, we specify the mandatory and continuous verification criteria, threshold-calibration procedure, and final acceptance rule in Section B.3. Finally, we report construction statistics and sequence-length distributions for the resulting $10{,}240$ records and examine the alignment between human judgements and verifier scores (Section B.4.3). We also visually assess the controlled visual variants through the original–resampled image pairs in Figure 6 and present representative examples in Section B.4.5.

B.1 Dataset Provenance and Leakage Control

We assemble an initial pool of $50{,}000$ records from five publicly available subsets of Visual-CoT (Shao et al., 2024a). The candidate pool comprises $6{,}000$ CUB (Wah et al., 2011) records

(12%), 12,000 Flickr30k (Young et al., 2014) records (24%), 15,000 GQA (Hudson & Manning, 2019) records (30%), 9,000 OpenImages (Kuznetsova et al., 2020) records (18%), and 8,000 Visual7W (Zhu et al., 2016) records (16%). We exclude records with missing images, invalid reference answers, invalid target bounding boxes, or questions that could be answered without consulting the associated image. Before coordinate normalization, we remove annotations with non-positive box areas or coordinates outside the corresponding image boundaries. We then convert all valid boxes to integer xyxy_{1000} coordinates. Following preprocessing and source-stratified selection, 21,751 records remained for leakage control. We check these records for overlap with MMMU-Pro, MMStar, VRC-Bench, VFaith-Bench, and SPD-Faith at both the image and question levels. Byte-identical images are identified using cryptographic hashes, while potential near-duplicates are detected through perceptual-similarity matching. For questions, we combine normalized exact matching with semantic-similarity matching. Potential overlaps identified through approximate image or question matching are manually reviewed. After construction and quality verification, the final **MCoT-ViG** corpus contains 10,240 records. Source-stratified sampling preserves the initial source proportions up to integer rounding, yielding 1,229 CUB, 2,458 Flickr30k, 3,072 GQA, 1,843 OpenImages, and 1,638 Visual7W records.

B.2 Construction Protocols and Prompt Designs

This subsection documents the construction protocols, prompt templates, and structured-output schemas used for four-stage MCoT generation with grounding annotations, attribute transformation, and controlled resampled-image generation. Each textual output was parsed and checked against its prescribed schema using deterministic rules. If parsing or schema validation failed, the corresponding generation was attempted up to four times.

B.2.1 Four-stage MCoT Generation and Grounding Annotation

For each source record, GPT-4o (OpenAI, 2024) serves as the teacher model and generates a structured MCoT rationale together with object-level grounding records. The model receives the question and verified reference answer, the original image, a grounding-guide image overlaid with the target bounding boxes, the corresponding object identifiers and coordinates, and optional source reasoning evidence when available. The teacher model returns a single JSON object containing grounding records for all supplied targets and a four-stage rationale comprising *Summary*, *Caption*, *Reasoning*, and *Conclusion*. When a record contains multiple target bounding boxes, the JSON object includes a separate grounding record for each annotated target. The *Caption* reports detailed, question-relevant visual evidence directly observable in the original image. The *Reasoning* then connects this evidence to the verified answer through explicit intermediate steps. At the first mention of each target object in *Reasoning*, its object label is followed by the corresponding reasoning-object identifier and complete bounding-box coordinates. Subsequent mentions retain the same identifier but omit the coordinates, thereby preserving cross-step co-reference. Table 11 provides the complete prompt template and required JSON output schema.

B.2.2 Attribute Transformation

The preceding four-stage generation procedure produces one grounding record per annotated target, including its identity, bounding box, aliases, and a concise description of its visually observable properties. Each target grounding record is submitted to GPT-4o (OpenAI, 2024) together with the question and verified reference answer. For source records containing multiple annotated targets, the procedure is applied independently to each target description. GPT-4o identifies explicitly stated visual attributes that admit answer-preserving transformations. Attributes directly queried by the question or required to derive the verified answer are ineligible. For each transformable target, the model replaces exactly one eligible attribute with a perceptually distinct yet plausible alternative. The resulting edits must preserve the verified reference answer, target categories and identities, spatial relations, and all unselected descriptive content. A description is marked as non-transformable if no explicitly stated attribute admits an answer-preserving edit. Records that cannot satisfy the transformation requirements are excluded from subsequent resampled-image generation. The minimally revised descriptions specify the intended localized edits and are used to construct the editing instructions for resampled-image generation. The original four-stage rationale remains unchanged throughout this procedure. Of the 16,794 source records entering this stage, 272 were excluded,

leaving 16,522 records for resampled-image generation. Table 12 provides the complete prompt template and required JSON output schema.

B.2.3 Controlled Resampled-Image Generation

For each transformable record, target-specific editing instructions are constructed from the attribute-transformation outputs. For records containing multiple annotated targets, these instructions cover all transformed targets. Each instruction identifies its target using the object label and normalized bounding-box coordinates. The source description, transformed description, and structured transformation record specify exactly one attribute substitution per target. Qwen-Image-Edit generates a controlled visual variant of the original image I using these localized editing instructions. The model is instructed to modify only the specified attribute within each designated target region. It must preserve the scene layout, object identities, spatial relations, all unselected target attributes, and all non-target content. The resulting image I^{rs} is paired with I as its controlled visual variant. Table 13 provides the complete editing prompt.

B.3 Quality Verification Details

This subsection details the non-compensatory verification procedure formalized in Equation 2. Each record should first satisfy three mandatory constraints: structural validity, single-stage completeness, and complete coverage of annotated targets in *Reasoning*. Cross-stage consistency and evidence coverage are then evaluated as continuous quality criteria. A record is retained only when all mandatory constraints hold and both continuous scores meet their calibrated thresholds. Section B.3.1 defines the mandatory constraints, Section B.3.2 details the continuous criteria, and Section B.3.3 presents threshold calibration and final acceptance.

B.3.1 Mandatory Acceptance Constraints

Structure Validation. A deterministic validator checks each generated record before model-based evaluation. The output parses as a single JSON object conforming to the construction schema. The object needs to contain exactly one non-empty field for each of the four MCoT stages in the required order. The *Reasoning* field comprises consecutively numbered steps, and the *Conclusion* must follow the constant format. For grounding annotations, every reasoning object id is unique and maps to exactly one supplied ground-truth object id. At the first occurrence in *Reasoning*, each target reference includes the object label, identifier, and four integer $xyxy_{1000}$ coordinates. These coordinates exactly match the corresponding normalized reference bounding boxes. Subsequent references may omit the coordinates but resolve to the same object identifier. Any parsing, schema, stage-format, identifier-resolution, or coordinate-matching failure sets the structural-validity indicator $S_{\mathrm{str},n}$ to zero; otherwise, $S_{\mathrm{str},n} = 1$. This constraint validates the structure of recorded references; complete target coverage is checked separately below.

Single-stage Completeness. A frozen Llama-3 (Grattafiori et al., 2024) verifier evaluates each stage independently using a role-specific rubric. For stage i of record x_n, it returns a binary role-fulfillment label $s^{\mathrm{sc}}_{i,n} \in \{0, 1\}$ and a concise failure code identifying any unmet requirement. The *Summary* needs to provide a question-conditioned reasoning plan without reporting visual observations, intermediate results, or the final answer. The *Caption* needs to describe question-relevant visual evidence without introducing inference or answer statements. The *Reasoning* needs to connect the reported evidence to the candidate answer through explicit intermediate steps. The *Conclusion* needs to state the answer directly in the prescribed format without adding further reasoning. A stage receives a label of 1 only when every requirement of the prescribed role is satisfied. The verifier assesses role fulfillment, not writing style, verbosity, cross-stage coherence, or image-level faithfulness. The record-level indicator is

$$S_{\mathrm{sc},n} = \prod_{i=1}^{4} s^{\mathrm{sc}}_{i,n}. \tag{12}$$

Therefore, $S_{\mathrm{sc},n} = 1$ only when all four stages satisfy their respective rubrics.

Object Coverage. Let $\mathcal{B}_n = \{(o_m, b_m)\}_{m=1}^{M_n}$ denote the annotated target objects and their source boxes for record x_n. A deterministic grounding parser scans *Reasoning* for coordinate-bearing and identifier-only object references. The parsed coordinate-bearing references form

$$\widehat{\mathcal{B}}_n^{\mathrm{rea}} = \{(\hat{o}_k, \hat{b}_k)\}_{k=1}^{K_n},$$

where $\hat{o}_k$ and $\hat{b}_k$ denote the predicted object reference and bounding box, respectively. A coordinate-bearing reference is parsed only when its first occurrence contains an object identifier and four integer xyxy_{1000} coordinates. Subsequent identifier-only mentions are linked to the corresponding parsed reference and do not create additional object box pairs. For $(\hat{o}_k, \hat{b}_k) \in \widehat{\mathcal{B}}_n^{\mathrm{rea}}$ and $(o_m, b_m) \in \mathcal{B}_n$, write $\hat{o}_k \equiv o_m$ when the predicted identifier resolves to the annotated target o_m. The set of successfully covered annotated targets is

$$\mathcal{C}_n^{\mathrm{rea}} = \left\{ (o_m, b_m) \in \mathcal{B}_n : \exists\, (\hat{o}_k, \hat{b}_k) \in \widehat{\mathcal{B}}_n^{\mathrm{rea}},\ \hat{o}_k \equiv o_m \wedge \hat{b}_k = b_m \right\}. \tag{13}$$

Since records without valid target annotations are removed during preprocessing, $|\mathcal{B}_n| > 0$. The object coverage score is

$$S_{\mathrm{ocr},n} = \frac{|\mathcal{C}_n^{\mathrm{rea}}|}{|\mathcal{B}_n|}. \tag{14}$$

The mandatory constraint requires $S_{\mathrm{ocr},n} = 1$, meaning that every annotated target has a valid, exactly matched reasoning-stage reference. Predictions referring to additional objects or unmatched boxes cannot increase this score, whose visual faithfulness is assessed separately by the evidence coverage criterion.

B.3.2 Continuous Verification Criteria

Cross-stage Consistency. Let $(t_{1,n}, t_{2,n}, t_{3,n}, t_{4,n})$ denote the *Summary*, *Caption*, *Reasoning*, and *Conclusion* texts of record x_n, respectively. Cross-stage consistency evaluates the local coherence of the three adjacent transitions: *Summary→Caption*, *Caption→Reasoning*, and *Reasoning→Conclusion*. For transition $i \in \{1, 2, 3\}$, the frozen Llama-3 verifier receives the question q_n, the adjacent stage texts $t_{i,n}$ and $t_{i+1,n}$, and a transition-specific consistency rubric. The first transition assesses whether the caption supplies the visual information required by the proposed reasoning plan. The second determines whether the inference consistently uses the captioned evidence without contradiction or an unsupported shift in focus. The third evaluates whether the conclusion follows from the preceding inference and preserves the derived answer. The image and verified reference answer are not provided to this verifier, so the criterion measures internal trajectory coherence rather than visual faithfulness or answer correctness. Let $\mathcal{Z}_{\mathrm{cc}} = \{1, 2, 3, 4, 5\}$ denote the score label set. For transition i of record x_n, let $\ell_{i,n}^{(z)}$ be the first-step logit assigned by the frozen verifier to label z. We normalize the logits over the five score labels and compute the expected transition score:

$$s_{i,n}^{\mathrm{cc}} = \sum_{z \in \mathcal{Z}_{\mathrm{cc}}} z \, \frac{\exp\left(\ell_{i,n}^{(z)}\right)}{\sum_{z' \in \mathcal{Z}_{\mathrm{cc}}} \exp\left(\ell_{i,n}^{(z')}\right)}. \tag{15}$$

Therefore, $s_{i,n}^{\mathrm{cc}} \in [1, 5]$, with larger values indicating stronger coherence between the corresponding adjacent stages. Because the three transitions form a sequential reasoning trajectory, an arithmetic mean may allow highly coherent transitions to compensate excessively for a weak one. We therefore normalize each transition score to $[0, 1]$, aggregate the normalized scores using their geometric mean, and map the result back to the five-point scale:

$$\begin{aligned} u_{i,n} &= \frac{s_{i,n}^{\mathrm{cc}} - 1}{4}, \qquad u_{i,n} \in [0, 1], \\ S_{\mathrm{cc},n} &= 1 + 4 \left(\prod_{i=1}^{3} u_{i,n} \right)^{1/3}. \end{aligned} \tag{16}$$

This aggregation preserves equal transition scores while assigning a greater penalty to locally inconsistent transitions. Accordingly, $S_{\mathrm{cc},n} \in [1, 5]$, with larger values indicating a more coherent four-stage rationale trajectory.

Evidence Coverage. We use GPT-5.1 (OpenAI, 2025) as a separate multimodal verifier to assess the visual evidence provided by each complete four-stage rationale. For each record, the verifier receives the original image, question, target-object annotations, and generated rationale. The verified reference answer is withheld to reduce outcome-conditioned assessment and require each score to be justified by the supplied visual evidence. The verifier independently assigns three integer scores on a five-point anchored scale: (1) s_n^{ac} measures the coverage of question-relevant attributes and relations; (2) s_n^{vf} measures the faithfulness of visual claims to the original image; (3) s_n^{rs} measures whether the cited visual evidence supports the intermediate reasoning. For each dimension, 1 indicates that the criterion is not satisfied, whereas 5 indicates complete satisfaction. Because these dimensions represent complementary requirements, we treat them as jointly necessary rather than freely substitutable evidence signals. Let $\mathcal{D}_{\mathrm{ecr}} = \{\mathrm{ac}, \mathrm{vf}, \mathrm{rs}\}$ denote the set of evidence-coverage dimensions. We first normalize each score to $[0, 1]$, aggregate the normalized values using their geometric mean, and map the result back to the original five-point scale:

$$v_n^d = \frac{s_n^d - 1}{4}, \qquad d \in \mathcal{D}_{\mathrm{ecr}} = \{\mathrm{ac}, \mathrm{vf}, \mathrm{rs}\},$$
$$S_{\mathrm{ecr},n} = 1 + 4 \left(\prod_{d \in \mathcal{D}_{\mathrm{ecr}}} v_n^d \right)^{1/3}. \tag{17}$$

This geometric aggregation acts as a continuous conjunction across the three dimensions, assigning a greater penalty to a weak component. It preserves equal component scores and yields $S_{\mathrm{ecr},n} = 1$ whenever any component score equals 1. Consequently, $S_{\mathrm{ecr},n} \in [1, 5]$, with larger values indicating more complete, faithful, and reasoning-relevant visual evidence. Table 14 provides the complete evaluation prompt and output schema.

B.3.3 Threshold Calibration and Final Acceptance

Let $\mathcal{D}_{\mathcal{H}}$ denote the set of records satisfying all three mandatory constraints, and let $\mathcal{K} = \{\mathrm{cc}, \mathrm{ecr}\}$ denote the two continuous criteria. We select the quantile level $q^\star$ using a human-annotated development subset. Let $Q_{q^\star}(\cdot)$ denote the empirical quantile function. For each $k \in \mathcal{K}$, the acceptance threshold is then computed as the empirical $q^\star$-quantile of its scores over $\mathcal{D}_{\mathcal{H}}$:

$$\tau_k = Q_{q^\star}\left(\{S_{k,n} \,|\, x_n \in \mathcal{D}_{\mathcal{H}}\}\right), \qquad k \in \mathcal{K}, \tag{18}$$

Using $q^\star = 0.10$ yields $\tau_{\mathrm{cc}} = 3.97$ for cross-stage consistency and $\tau_{\mathrm{ecr}} = 3.73$ for evidence coverage. Under the normalization $\widetilde{S}_{k,n} = (S_{k,n} - 1)/4$, these thresholds correspond to 0.74 and 0.68 on the $[0, 1]$ scale. The verifier models, prompt templates, scoring rubrics, and decoding settings are frozen during quality verification. Following Equation 2, a record is retained if and only if it satisfies all three mandatory constraints and simultaneously meets $S_{\mathrm{cc},n} \geq \tau_{\mathrm{cc}}$ and $S_{\mathrm{ecr},n} \geq \tau_{\mathrm{ecr}}$.

B.4 Dataset Characterization and Quality Validation

This subsection characterizes the resulting MCoT-ViG dataset and evaluates the validity of automatic verifier scores and the quality of controlled visual variants. Section B.4.1 reports record retention throughout the construction and verification pipeline. Section B.4.2 summarizes the source composition and sequence-length distributions of the final dataset. Section B.4.3 evaluates the alignment between the automatic verifier scores and human judgements. Section B.4.4 examines whether the controlled visual variants realize the intended attribute transformations while preserving target identity and non-target scene content. Section B.4.5 presents representative records illustrating paired original and resampled images, four-stage rationales, and object-level grounding annotations.

B.4.1 Construction and Verification Statistics

Table 3 reports record retention throughout the MCoT-ViG construction and quality verification pipeline. Each complete record contains a valid four-stage rationale, traceable object-level grounding annotations, an eligible answer-preserving attribute transformation, and a valid resampled image. Each excluded record is counted once at the earliest processing step that makes it ineligible for further processing. Starting from $50,000$ candidate records, source preprocessing, benchmark-overlap filtering, rationale generation, and grounding alignment retain $16,794$ records. Attribute transformation excludes 272 records without an eligible answer-preserving attribute, leaving $16,522$ records

Table 3: **Record flow through the MCoT-ViG construction and verification pipeline.** Conditional retention is calculated relative to the input of each processing step, whereas overall retention is calculated relative to the initial pool.

Processing step	Input	Rejected	Retained	Conditional retention (%)	Overall retention (%)
Initial candidate pool	50,000	0	50,000	100.00	100.00
Source preprocessing and stratified selection	50,000	28,249	21,751	43.50	43.50
Benchmark-overlap filtering	21,751	1,275	20,476	94.14	40.95
Four-stage rationale generation	20,476	493	19,983	97.59	39.97
Grounding alignment and annotation	19,983	3,189	16,794	84.04	33.59
Attribute transformation	16,794	272	16,522	98.38	33.04
Resampled-image generation	16,522	3,999	12,523	75.80	25.05
Mandatory acceptance constraints	12,523	149	12,374	98.81	24.75
Continuous verification criteria	12,374	2,134	10,240	82.75	20.48

Table 4: **Source composition and text-token length statistics for MCoT-ViG.** Candidate-pool and final-corpus counts refer to records used in the MCoT-ViG construction pipeline. Token counts are computed using the Qwen3-VL tokenizer and reported as median $[Q_1, Q_3]$, rounded to the nearest token. Total denotes the per-record sum of the four stage lengths before dataset-level aggregation.

Source	Number of records		Input length		Four-stage MCoT length				
	Initial	Final	Question	Answer	Summary	Caption	Reasoning	Conclusion	Total
CUB	6,000	1,229	15 [12, 17]	1 [1, 1]	22 [19, 24]	56 [47, 66]	76 [66, 88]	6 [6, 6]	161 [142, 183]
Flickr30k	12,000	2,458	17 [16, 18]	14 [9, 26]	27 [23, 31]	45 [44, 48]	86 [80, 92]	18 [13, 30]	175 [162, 199]
GQA	15,000	3,072	7 [6, 10]	1 [1, 2]	21 [20, 23]	57 [51, 68]	79 [70, 91]	6 [6, 6]	167 [150, 184]
OpenImages	9,000	1,843	6 [6, 7]	2 [1, 2]	20 [18, 22]	50 [42, 60]	72 [63, 84]	7 [6, 7]	149 [132, 172]
Visual7W	8,000	1,638	7 [7, 10]	5 [5, 6]	21 [19, 24]	53 [45, 63]	76 [66, 88]	10 [9, 11]	160 [141, 184]
Overall	50,000	10,240	9 [6, 15]	2 [1, 6]	22 [20, 24]	50 [45, 61]	80 [70, 90]	7 [6, 11]	166 [148, 185]

for resampled-image generation. Among these records, $12,523$ produce valid resampled images and enter quality verification. The mandatory acceptance constraints retain $12,374$ records. Applying the calibrated thresholds for cross-stage consistency and evidence coverage further retains $10,240$ records in the final MCoT-ViG dataset, corresponding to 20.48% of the initial candidate pool.

B.4.2 Source Composition and Sequence-Length Statistics

Table 4 summarizes the source composition and textual length distributions of MCoT-ViG. Through stratified selection, the final corpus preserves the initial source proportions of 12%, 24%, 30%, 18%, and 16% for CUB, Flickr30k, GQA, Open Images, and Visual7W, respectively. The answer-length differences reflect the construction characteristics of the source subsets. Flickr30k records contain the longest answers because their question–answer pairs are generated from descriptive captions. In contrast, CUB, GQA, and OpenImages records predominantly contain concise category, attribute, or relation answers, while Visual7W instances more often contain short answer phrases. Despite these differences, the shared four-stage schema yields comparable rationale lengths, with median total lengths ranging from 149 to 175 tokens. Among the four MCoT stages, the cross-source variation in median length is greatest for *Reasoning*, whose median ranges from 72 to 86 tokens.

B.4.3 Alignment between Human judgements and Automatic Verification

To examine the alignment between automatic verification scores and expert assessments, we conducted a stratified audit of 300 records that passed all mandatory acceptance constraints. We constructed three non-overlapping audit sets of 100 records using random seeds 17, 29, and 43. Within each set, records were jointly stratified by source subset and automatic-score quintile. This procedure preserved the source composition while covering the observed score range. Each record was independently evaluated by three domain experts using criterion-specific rubrics for cross-stage consistency and evidence coverage. The experts were blinded to the automatic verifier scores. For each record and criterion, the human consensus score was computed as the arithmetic mean of the three expert ratings. Both the human consensus scores and automatic verifier scores were then linearly mapped to $[0, 1]$. Figure 5 compares the automatic and human scores. The left panel reports the paired difference $\Delta_{n,k} = S^{\text{auto}}_{n,k} - S^{\text{human}}_{n,k}$ for record n and criterion k. Across the three audit draws, the mean paired differences remained close to zero for both criteria. Cross-stage consistency showed small positive offsets in the first two draws and a small negative offset in the third,

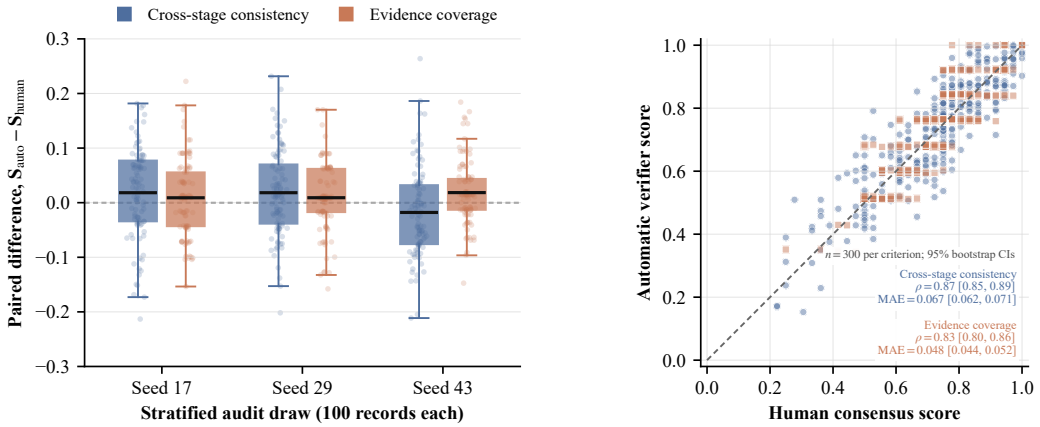


Figure 5: **Alignment between continuous verification scores and human judgements. Left:** Distributions of paired differences between automatic scores and human consensus scores across three stratified audit sets of 100 records. **Right:** Automatic verifier scores plotted against human consensus scores for the pooled audit set ($n = 300$ records per criterion).

whereas evidence coverage exhibited consistently small positive offsets. These values indicate limited overall directional bias, although the sign change for cross-stage consistency under Seed 43 reveals modest variation across audit draws. Cross-stage consistency also exhibited greater dispersion than evidence coverage, indicating larger record-level deviations from the human assessments. The right panel evaluates pooled record-level alignment. Cross-stage consistency achieved a Spearman correlation of $\rho = 0.87$ with a 95% confidence interval of $[0.85, 0.89]$, together with an MAE of 0.067 $[0.062, 0.071]$. Evidence coverage achieved $\rho = 0.83$ $[0.80, 0.86]$ and an MAE of 0.048 $[0.044, 0.052]$. Therefore, cross-stage consistency exhibited stronger rank association, while evidence coverage showed smaller absolute deviations from the human consensus scores.

B.4.4 Quality of Controlled Visual Variants

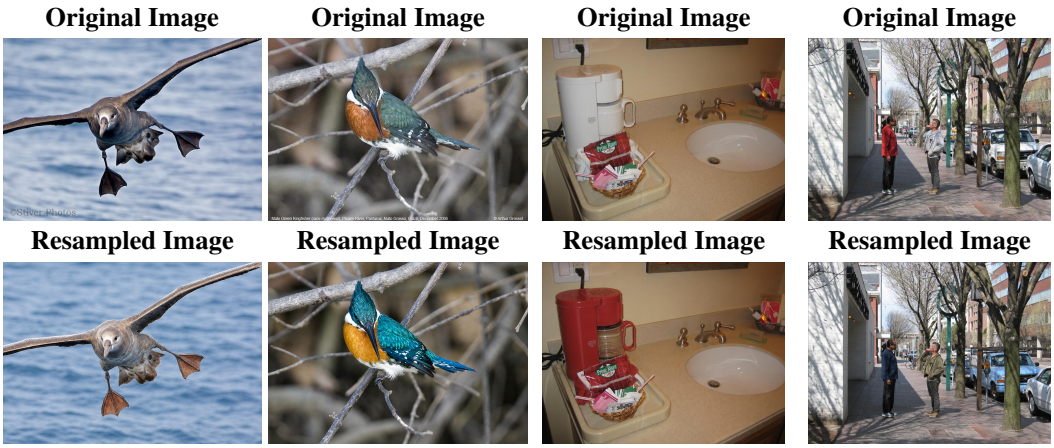
Each original image I and its controlled variant I^{rs} were evaluated against four conjunctive criteria: edit realization, target preservation, non-target preservation, and answer preservation. A pair was retained only when all specified attribute transformations were visibly realized within their corresponding target regions while preserving the validity of the verified answer. The controlled variant was also required to preserve target identities, spatial relations, visual style, and content outside the edited regions. We rejected pairs containing missing or incomplete edits, target deletion or duplication, unintended changes to question-relevant evidence, or conspicuous generation artifacts. Figure 6 qualitatively illustrates representative pairs involving bird appearance, feather color, appliance color, and attributes of multiple objects in a complex street scene.

B.4.5 Representative MCoT-ViG Samples

Figures 9 and 10 pair an original-image rationale with a controlled visual variant. The edited image changes an answer-irrelevant color attribute, while the four-stage rationale remains conditioned on the original image and is therefore kept unchanged. Figure 11 instead focuses on grounding a small target within a cluttered scene. At the first occurrence of each target object in *Reasoning*, its object reference includes the reasoning-object identifier and complete bounding-box coordinates. Subsequent references retain only the identifier, thereby maintaining cross-step co-reference without repeatedly inserting the coordinates. Together, these examples illustrate how MCoT-ViG combines original-image rationales, controlled visual variants, and persistent object-level grounding within a unified record structure.

C Perception-aware Intermediate Process Optimization Details

This appendix provides the complete formulation and implementation details of **Perception-aware Intermediate Process Optimization** (**PaIPO**). Motivated by the observed discrepancy between



(a) Plumage and Foot (b) Feather Color (c) Appliance Color (d) Clothing and Vehicle

Figure 6: **Representative controlled visual variants from MCoT-ViG.** Each column pairs an original image I (top) with its controlled variant I^{rs} (bottom). **(a)** Changes to the appearance of a flying bird. **(b)** A localized change in feather color. **(c)** A color transformation applied to a coffee maker. **(d)** Attribute transformations applied to annotated objects in a complex street scene.

final-answer correctness and intermediate visual faithfulness, PaIPO supplements conventional RLVR with verifiable supervision over the intermediate MCoT trajectory. The **Caption Perception Reward** (**CPR**) evaluates whether the generated *Caption* is semantically aligned with the verified reference and preferentially supported by the original image over the controlled variant. The **Reasoning Grounding Reward** (**RGR**) assesses whether visual references in *Reasoning* are spatially aligned with annotated objects and remain consistent across reasoning steps. The **Step Reward** (**SR**) complements these perception-oriented signals by evaluating semantic trajectory alignment, sequential order consistency, and terminal answer entailment. Appendix C.1 details the semantic-alignment and counterfactual-discrimination components of CPR. Appendix C.2 formalizes object matching, spatial agreement, and cross-step grounding consistency for RGR. Appendix C.3 defines the complete formulation of Step Reward, while Appendix C.4 integrates these signals into the unified process and trajectory rewards and presents the GRPO and DAPO instantiations.

C.1 Caption Perception Reward

The **Caption Perception Reward** (**CPR**) evaluates two complementary properties of a generated *Caption*: semantic fidelity to the verified reference and sensitivity to the supplied visual evidence. Reference alignment alone may reward descriptions supported primarily by linguistic priors, while visual discrimination alone may favor distinctive but incomplete content. CPR therefore combines linguistic reference alignment with counterfactual visual discrimination. Both components evaluate the caption sampled from the original image. The controlled variant is used to rescore the same generated caption without sampling an additional trajectory.

Let $c_i = o_i^{\mathrm{cap}}$ denote the *Caption* of the i-th completion and $c^\star$ the corresponding reference caption. A frozen atomic extractor $\mathcal{E}_{\mathrm{atom}}$ decomposes them into visual proposition sets:

$$\mathcal{A}_i = \mathcal{E}_{\mathrm{atom}}(c_i), \qquad \mathcal{A}^\star = \mathcal{E}_{\mathrm{atom}}(c^\star).$$

Linguistic Reference Alignment. For propositions u and v, let $p_\psi(u \Rightarrow v) \in [0,1]$ denote the entailment probability assigned by a frozen textual verifier. We define their symmetric semantic agreement as

$$M_\psi(u,v) = \sqrt{p_\psi(u \Rightarrow v)\, p_\psi(v \Rightarrow u)}. \tag{19}$$

The bidirectional construction assigns a high score only when the propositions mutually support one another, while remaining tolerant of valid paraphrases. A one-way entailment score may treat a broader or more specific proposition as semantically equivalent, while requiring entailment in both directions reduces this asymmetry and favors propositions with comparable semantic content. Using

M_ψ as a soft matching score, we compute candidate precision and reference coverage:

$$P_i^{\mathrm{ref}} = \frac{1}{|\mathcal{A}_i|} \sum_{u \in \mathcal{A}_i} \max_{v \in \mathcal{A}^\star} M_\psi(u, v),$$
$$C_i^{\mathrm{ref}} = \frac{1}{|\mathcal{A}^\star|} \sum_{v \in \mathcal{A}^\star} \max_{u \in \mathcal{A}_i} M_\psi(u, v). \tag{20}$$

Candidate precision penalizes generated propositions unsupported by the reference, while reference coverage penalizes omitted reference evidence. Their harmonic mean defines the linguistic reference-alignment reward:

$$r_i^{\mathrm{ref}} = \frac{2 P_i^{\mathrm{ref}} C_i^{\mathrm{ref}}}{P_i^{\mathrm{ref}} + C_i^{\mathrm{ref}} + \epsilon_{\mathrm{num}}}. \tag{21}$$

Here, $\epsilon_{\mathrm{num}} > 0$ ensures numerical stability. When either proposition set is empty, we define $r_i^{\mathrm{ref}} = 0$ without evaluating P_i^{ref} or C_i^{ref}.

Counterfactual Visual Discrimination. Semantic agreement does not establish whether the generated caption depends on the original visual evidence. We therefore compare the likelihood of the same caption tokens under the original image I and its controlled variant I^{rs}.

Let $\mathcal{T}_i^{\mathrm{cap}}$ denote the token indices belonging to the generated *Caption*, and let $\pi_{\bar{\theta}}$ denote a frozen snapshot of the rollout policy. For every $t \in \mathcal{T}_i^{\mathrm{cap}}$, we compute

$$\delta_{i,t} = \log \pi_{\bar{\theta}}\left(o_{i,t} \mid q, I, o_{i,<t}\right) - \log \pi_{\bar{\theta}}\left(o_{i,t} \mid q, I^{\mathrm{rs}}, o_{i,<t}\right), \qquad e_{i,t} = \tanh\left(\frac{\delta_{i,t}}{2\tau_{\mathrm{cf}}}\right). \tag{22}$$

The question, generated token, and preceding textual context are identical in both evaluations. The temperature $\tau_{\mathrm{cf}} > 0$ controls sensitivity to the counterfactual likelihood difference. Accordingly, $e_{i,t} > 0$ indicates stronger support from the original image, $e_{i,t} < 0$ favors the controlled variant, and $e_{i,t} \approx 0$ indicates visual invariance.

Because the controlled transformation is localized, many caption tokens should be similarly supported by both images. We therefore use confidence-weighted aggregation so that visually invariant tokens contribute little:

$$D_i^{\mathrm{cf}} = \frac{\sum_{t \in \mathcal{T}_i^{\mathrm{cap}}} |e_{i,t}|^\gamma e_{i,t}}{\sum_{t \in \mathcal{T}_i^{\mathrm{cap}}} |e_{i,t}|^\gamma + \epsilon}, \qquad r_i^{\mathrm{vis}} = \left[\frac{D_i^{\mathrm{cf}} - m_{\mathrm{cf}}}{1 - m_{\mathrm{cf}}}\right]_{[0,1]}. \tag{23}$$

Here, $\gamma \geq 1$ increases the contribution of tokens with stronger counterfactual evidence, and $\epsilon > 0$ ensures numerical stability. The clipping operator is defined as $[x]_{[0,1]} = \min\{1, \max\{0, x\}\}$, and $m_{\mathrm{cf}} \in [0, 1)$ is an indifference margin, which discounts the intrinsic preference for I induced by sampling the completion under the original visual condition. Therefore, visually invariant tokens contribute little to D_i^{cf}, while tokens preferentially supported by I^{rs} contribute negatively. A high r_i^{vis} requires the aggregate caption evidence to favor the original image beyond the calibrated margin.

Caption Perception Reward. Semantic fidelity and counterfactual visual sensitivity are treated as jointly necessary properties. We combine them using the harmonic mean:

$$r_i^{\mathrm{CPR}} = \frac{2 r_i^{\mathrm{ref}} r_i^{\mathrm{vis}}}{r_i^{\mathrm{ref}} + r_i^{\mathrm{vis}} + \epsilon}. \tag{24}$$

Consequently, a high CPR requires the generated caption to cover the verified reference evidence without unsupported content and to remain preferentially supported by the original image.

C.2 Reasoning Grounding Reward

Caption perception reward encourages the generated *Caption* to remain sensitive to visual evidence, but cannot ensure that the subsequent *Reasoning* remains spatially grounded in the relevant image

regions. We therefore introduce the **Reasoning Grounding Reward** (**RGR**) to evaluate two complementary properties of the reasoning trajectory: spatial agreement between predicted and annotated grounding regions, and identifier consistency across reasoning steps.

For a source record, let

$$\mathcal{G} = \{(u_k, o_k, b_k)\}_{k=1}^{K} \tag{25}$$

denote the reference grounding map, where u_k is a textual reference to target object o_k, b_k is its annotated bounding box, and $K \geq 1$ is the number of annotated targets. From the i-th completion $\mathbf{o}_i$, a deterministic parser extracts the coordinate-bearing grounding records

$$\widehat{\mathcal{G}}_i = \left\{ (\widehat{u}_{i,j}, \widehat{o}_{i,j}, \widehat{b}_{i,j}) \right\}_{j=1}^{\widehat{K}_i}, \tag{26}$$

where $\widehat{K}_i$ is the number of parsed predictions. Under the required output format, these records correspond to the first coordinate-bearing occurrences of referenced objects. Subsequent identifier-only mentions are collected separately for cross-step consistency evaluation. All predicted and annotated boxes use the same normalized coordinate system.

Pairwise Spatial Grounding. For predicted box $\widehat{b}_{i,j}$ and annotated box b_k, we define the pairwise grounding score through their spatial overlap:

$$\omega_{i,j,k} = \mathrm{IoU}\left(\widehat{b}_{i,j}, b_k\right) = \frac{\left|\widehat{b}_{i,j} \cap b_k\right|}{\left|\widehat{b}_{i,j} \cup b_k\right|}. \tag{27}$$

Here, $|\cdot|$ denotes the area of a region. Accordingly, $\omega_{i,j,k} \in [0,1]$ is a continuous measure of spatial agreement. A value of zero indicates no overlap, whereas a value of one indicates exact alignment.

Multi-target Grounding Matching. When a sample contains multiple predicted and annotated boxes, independently assigning each prediction to the highest overlap target may produce duplicate assignments. We therefore construct the complete pairwise IoU matrix

$$\mathbf{\Omega}_i = [\omega_{i,j,k}]_{j=1,\ldots,\widehat{K}_i;\, k=1,\ldots,K} \in [0,1]^{\widehat{K}_i \times K}. \tag{28}$$

Let $[m] = \{1, \ldots, m\}$, and let $\mathbb{I}[\cdot]$ denote the indicator function. The set of valid matchings is

$$\mathfrak{M}_i = \left\{ \mathcal{M} \subseteq [\widehat{K}_i] \times [K] \;\middle|\; \begin{array}{l} \displaystyle\sum_{k=1}^{K} \mathbb{I}[(j,k) \in \mathcal{M}] \leq 1, \quad \forall j \in [\widehat{K}_i], \\ \displaystyle\sum_{j=1}^{\widehat{K}_i} \mathbb{I}[(j,k) \in \mathcal{M}] \leq 1, \quad \forall k \in [K] \end{array} \right\}. \tag{29}$$

The first constraint allows each predicted box to match at most one annotated box, while the second ensures each annotated box is matched with at most one prediction. We select the correspondence that maximizes the total spatial agreement:

$$\mathcal{M}_i^{\star} = \arg\max_{\mathcal{M} \in \mathfrak{M}_i} \sum_{(j,k) \in \mathcal{M}} \omega_{i,j,k}. \tag{30}$$

Consequently, $|\mathcal{M}_i^{\star}| \leq \min(\widehat{K}_i, K)$. When several local assignments compete for the same annotated region, the matching selects the combination with the largest IoU over the complete sample.

Localization-Weighted Precision and Recall. We define the total matched grounding mass as

$$W_i^{\mathrm{grd}} = \sum_{(j,k) \in \mathcal{M}_i^{\star}} \omega_{i,j,k}. \tag{31}$$

The localization-weighted precision, recall, and F-score are then

$$P_i^{\mathrm{grd}} = \frac{W_i^{\mathrm{grd}}}{\max(1, \widehat{K}_i)}, \qquad R_i^{\mathrm{grd}} = \frac{W_i^{\mathrm{grd}}}{K}, \qquad F_i^{\mathrm{grd}} = \frac{2 P_i^{\mathrm{grd}} R_i^{\mathrm{grd}}}{P_i^{\mathrm{grd}} + R_i^{\mathrm{grd}} + \epsilon}. \tag{32}$$

Here, $\epsilon > 0$ ensures numerical stability. Since $K \geq 1$, the recall denominator is always well-defined. This formulation handles unequal set sizes without introducing an additional discrete penalty. When $\widehat{K}_i > K$, unmatched predictions remain in the precision denominator and reduce P_i^{grd}. When $\widehat{K}_i < K$, unmatched annotated targets remain in the recall denominator and reduce R_i^{grd}. Multiple predictions competing for the same annotated region cannot all receive matched grounding mass because of the one-to-one constraint. Similarly, one predicted box cannot simultaneously account for several annotated targets. If no coordinate-bearing prediction is parsed, then $\widehat{K}_i = 0$, $\mathcal{M}_i^\star = \varnothing$, and $W_i^{\mathrm{grd}} = P_i^{\mathrm{grd}} = R_i^{\mathrm{grd}} = F_i^{\mathrm{grd}} = 0$.

Cross-Step Identifier Consistency. Spatial overlap alone does not ensure that later reasoning steps continue to refer to the same grounded objects. MCoT-ViG therefore requires each subsequent object mention to preserve the identifier established at the first coordinate-bearing occurrence. Let $\mathcal{R}_i^{\mathrm{sub}}$ denote the set of subsequent object references expected to reuse an established identifier in the i-th reasoning trajectory. For each $r \in \mathcal{R}_i^{\mathrm{sub}}$, define $\zeta_{i,r} \in \{0, 1\}$ as

$$\zeta_{i,r} = \begin{cases} 1, & \text{if reference } r \text{ resolves to the identifier established} \\ & \text{for the same object at its first occurrence,} \\ 0, & \text{otherwise.} \end{cases} \tag{33}$$

Unresolved identifiers, identifier substitutions, and references that conflict with the established object mapping receive a value of zero. We compute cross-step identifier consistency as

$$C_i^{\mathrm{id}} = \begin{cases} \dfrac{1}{|\mathcal{R}_i^{\mathrm{sub}}|} \displaystyle\sum_{r \in \mathcal{R}_i^{\mathrm{sub}}} \zeta_{i,r}, & |\mathcal{R}_i^{\mathrm{sub}}| > 0, \\ 1, & |\mathcal{R}_i^{\mathrm{sub}}| = 0. \end{cases} \tag{34}$$

The second case avoids penalizing an object that is grounded correctly but mentioned only once, because such a trajectory contains no subsequent co-reference to verify. Importantly, C_i^{id} measures identifier persistence relative to the first coordinate-bearing occurrence and does not alter the pairwise IoU scores or box matching.

Reasoning Grounding Reward. The final RGR combines localization quality with cross-step identifier consistency:

$$r_i^{\mathrm{RGR}} = F_i^{\mathrm{grd}} \left[(1 - \lambda_{\mathrm{id}}) + \lambda_{\mathrm{id}} C_i^{\mathrm{id}} \right], \tag{35}$$

where $\lambda_{\mathrm{id}} \in [0, 1]$ controls the contribution of identifier consistency. Since $F_i^{\mathrm{grd}}, C_i^{\mathrm{id}} \in [0, 1]$, the resulting reward satisfies $r_i^{\mathrm{RGR}} \in [0, 1]$. When all subsequent references are consistent, $C_i^{\mathrm{id}} = 1$ and RGR reduces to F_i^{grd}. When identifier consistency decreases, it proportionally attenuates the spatial grounding reward. The multiplicative construction prevents consistent identifiers from compensating for missing or inaccurate localization, since $F_i^{\mathrm{grd}} = 0$ always implies $r_i^{\mathrm{RGR}} = 0$.

C.3 Step Reward for Reasoning Steps

The reasoning grounding reward validates the spatial grounding of object references within the reasoning trajectory, but spatial alignment alone does not ensure the semantic completeness, sequential coherence, or answer support of multi-step inference. A spatially grounded trajectory may still omit necessary reasoning links, order otherwise valid steps incorrectly, or provide insufficient evidential support for the verified answer. We therefore define the **Step Reward** (**SR**), a structured metric that evaluates three complementary properties of the generated *Reasoning* trajectory: ***Semantic Trajectory Alignment***, ***Sequential Order Consistency***, and ***Terminal Answer Entailment***.

A deterministic parser decomposes both the generated and reference *Reasoning* into ordered step sequences:

$$\mathbf{P}_i = (p_{i,1}, \ldots, p_{i,M_i}), \qquad \mathbf{P}^\star = (p_1^\star, \ldots, p_{M^\star}^\star), \tag{36}$$

where M_i and $M^\star$ denote the number of parsed steps for the generated and reference reasoning, respectively, with $M^\star \geq 1$. Only successfully parsed reasoning steps are carried forward to subsequent computations; structural format compliance is evaluated independently by the format reward.

Let $p_\psi(p \Rightarrow p') \in [0,1]$ denote the entailment probability assigned by the frozen textual NLI verifier shared with CPR. We define the symmetric semantic similarity between generated step $p_{i,j}$ and reference step $p_k^\star$ as

$$s_{i,j,k} = \sqrt{p_\psi\left(p_{i,j} \Rightarrow p_k^\star\right) p_\psi\left(p_k^\star \Rightarrow p_{i,j}\right)}. \tag{37}$$

This bidirectional formulation mitigates the inherent asymmetry between broader and more specific statements, ensuring that high similarity scores are assigned only when the two steps convey comparable semantic content.

Semantic Trajectory Alignment. Semantic trajectory alignment quantifies the coverage of the reference reasoning steps while penalizing unmatched generated steps. Let $0 < \tau_{\text{step}} \leq 1$ denote the minimum similarity threshold for an eligible step pair. We define the candidate edge set as

$$\mathcal{E}_i^{\text{step}}(\tau_{\text{step}}) = \left\{(j,k) \in [M_i] \times [M^\star] : s_{i,j,k} \geq \tau_{\text{step}}\right\}, \tag{38}$$

where $[m] = \{1, \ldots, m\}$ for $m \geq 1$, $[0] = \emptyset$, and $\mathbb{I}[\cdot]$ denotes the indicator function. We next formalize the set of admissible one-to-one step matchings, which follows standard bipartite matching constraints to ensure fair and unambiguous scoring:

$$\begin{aligned}\mathfrak{M}_i^{\text{step}}(\tau_{\text{step}}) = \Bigg\{ \mathcal{M} \subseteq \mathcal{E}_i^{\text{step}}(\tau_{\text{step}}) \;\Bigg|\; & \sum_{k=1}^{M^\star} \mathbb{I}\left[(j,k) \in \mathcal{M}\right] \leq 1, \quad \forall j \in [M_i], \\ & \sum_{j=1}^{M_i} \mathbb{I}\left[(j,k) \in \mathcal{M}\right] \leq 1, \quad \forall k \in [M^\star] \Bigg\}.\end{aligned} \tag{39}$$

The first constraint ensures that each generated step is matched with at most one reference step. The second ensures that each reference step is counted at most once, preventing duplicate credit from being assigned to repeated generated steps.

We compute the maximum-weight semantic matching and the corresponding total matched semantic mass as

$$\mathcal{M}_i^{\text{step},\star} = \underset{\mathcal{M} \in \mathfrak{M}_i^{\text{step}}(\tau_{\text{step}})}{\arg\max} \sum_{(j,k)\in\mathcal{M}} s_{i,j,k}, \qquad S_i^{\text{sem}} = \sum_{(j,k)\in\mathcal{M}_i^{\text{step},\star}} s_{i,j,k}. \tag{40}$$

When multiple matchings yield identical total semantic weights within a predefined numerical tolerance, we apply a deterministic hierarchical tie-breaking rule. We first retain the candidate matchings with the largest cardinality, thereby using the greatest number of eligible semantic correspondences without materially changing the primary matching objective. If multiple candidates remain, the pairs in each matching are sorted first by generated-step index and then by reference-step index, and the lexicographically smallest sequence is selected.

The resulting alignment-weighted precision, recall, and F_1 score are defined as:

$$P_i^{\text{step}} = \frac{S_i^{\text{sem}}}{\max(1, M_i)}, \qquad R_i^{\text{step}} = \frac{S_i^{\text{sem}}}{M^\star}, \qquad F_i^{\text{step}} = \frac{2 P_i^{\text{step}} R_i^{\text{step}}}{P_i^{\text{step}} + R_i^{\text{step}} + \epsilon}. \tag{41}$$

Since $S_i^{\text{sem}} \leq \min(M_i, M^\star)$, all three metrics are bounded within $[0,1]$. Unmatched generated steps degrade precision, whereas unmatched reference steps lower recall. The one-to-one constraint prevents redundant generated steps from matching the same reference step multiple times. If no generated reasoning steps are successfully parsed (*i.e.*, $M_i = 0$), the candidate edge set is empty, and all metrics evaluate to zero: $S_i^{\text{sem}} = P_i^{\text{step}} = R_i^{\text{step}} = F_i^{\text{step}} = 0$.

Sequential Order Consistency. While semantic trajectory alignment establishes step-level semantic matches, it imposes no constraints on the sequential order of matched steps relative to the reference. We accordingly design sequential order consistency to quantify the degree to which matched steps preserve the reference ordering. We first extract the sequence of matched step pairs from the optimal semantic matching $\mathcal{M}_i^{\text{step},\star}$, sorted by the generated reasoning step in ascending order. Formally, let $(j_1, k_1), \ldots, (j_{m_i}, k_{m_i})$ denote this sorted sequence, satisfying $j_1 < \ldots < j_{m_i}$, where $m_i = |\mathcal{M}_i^{\text{step},\star}|$ denotes the total number of matched step pairs. The corresponding sequence of reference step indices is denoted as

$$\boldsymbol{\kappa}_i = (k_1, \ldots, k_{m_i}).$$

To quantify ordering deviations from the reference, we compute the inversion count of $\boldsymbol{\kappa}_i$, which counts the number of pairwise order violations relative to the reference sequence:

$$\mathrm{Inv}\left(\boldsymbol{\kappa}_i\right) = \sum_{1 \le a < b \le m_i} \mathbb{I}\left[k_a > k_b\right]. \tag{42}$$

Since the semantic matching is one-to-one, the reference indices are distinct.

We then derive the order-consistency score by normalizing the inversion count against its theoretical maximum:

$$O_i^{\mathrm{step}} = \begin{cases} 1 - \dfrac{2\,\mathrm{Inv}\left(\boldsymbol{\kappa}_i\right)}{m_i(m_i-1)}, & m_i \ge 2, \\ 1, & M_i > 0 \text{ and } m_i < 2, \\ 0, & M_i = 0. \end{cases} \tag{43}$$

For $m_i \ge 2$, the denominator corresponds to twice the maximum possible inversion count. Therefore, $O_i^{\mathrm{step}} = 1$ when all matched steps preserve the reference order, while $O_i^{\mathrm{step}} = 0$ when their order is completely reversed. For edge cases with fewer than two matched pairs, relative ordering cannot be evaluated. We adopt a neutral scoring policy: non-empty generated trajectories receive a score of 1, as any lack of semantic coverage is already penalized by the semantic trajectory alignment F_i^{step}. Empty trajectories with no parsed steps are assigned a score of 0.

Terminal Answer Entailment. Terminal answer entailment assesses whether the final step of the generated reasoning trajectory provides explicit semantic support for the verified answer. Let $\mathrm{Statement}(q, y)$ denote the function that converts question q and verified answer y into a declarative answer statement. We define the terminal entailment score as

$$T_i^{\mathrm{step}} = \begin{cases} p_\psi\left(p_{i,M_i} \Rightarrow \mathrm{Statement}(q,y)\right), & M_i > 0, \\ 0, & M_i = 0. \end{cases} \tag{44}$$

This directional entailment check verifies that the final reasoning step supports the answer statement, rather than the reverse direction where the answer would imply the preceding reasoning. This component complements the answer-correctness reward by evaluating whether the reasoning trajectory terminates with explicit step-level support for the verified answer.

Step Reward. The three sub-metrics quantify complementary properties of the reasoning process: F_i^{step} measures semantic trajectory alignment, O_i^{step} measures sequential order consistency, and T_i^{step} measures terminal answer entailment. Let $\mathcal{D}_{\mathrm{SR}} = \{\mathrm{sem}, \mathrm{ord}, \mathrm{ter}\}$ denote the set of component indices, with:

$$x_{i,\mathrm{sem}} = F_i^{\mathrm{step}}, \qquad x_{i,\mathrm{ord}} = O_i^{\mathrm{step}}, \qquad x_{i,\mathrm{ter}} = T_i^{\mathrm{step}}.$$

Each component takes values in $[0, 1]$. To mitigate cross-component compensation while preventing a zero score in any single dimension from nullifying all signal from the remaining dimensions, we apply an affine smoothing transformation:

$$\widetilde{x}_{i,d} = \delta_{\mathrm{SR}} + (1 - \delta_{\mathrm{SR}})\, x_{i,d}, \qquad d \in \mathcal{D}_{\mathrm{SR}}, \qquad 0 < \delta_{\mathrm{SR}} < 1. \tag{45}$$

The final Step Reward is formulated as a normalized weighted geometric mean:

$$r_i^{\mathrm{SR}} = \frac{\prod_{d \in \mathcal{D}_{\mathrm{SR}}} \widetilde{x}_{i,d}^{\lambda_d} - \delta_{\mathrm{SR}}}{1 - \delta_{\mathrm{SR}}}, \qquad \lambda_d > 0, \qquad \sum_{d \in \mathcal{D}_{\mathrm{SR}}} \lambda_d = 1. \tag{46}$$

The component weights λ_d control the relative contribution of each dimension, while preserving the conjunctive behavior of geometric aggregation. Given that $\widetilde{x}_{i,d} \in [\delta_{\mathrm{SR}}, 1]$ and the weights sum to unity, it holds that:

$$\delta_{\mathrm{SR}} \le \prod_{d \in \mathcal{D}_{\mathrm{SR}}} \widetilde{x}_{i,d}^{\lambda_d} \le 1. \tag{47}$$

It follows that $r_i^{\mathrm{SR}} \in [0, 1]$. This normalized geometric aggregation rewards reasoning trajectories that jointly exhibit strong semantic alignment, sequential order consistency, and terminal answer entailment. It also limits the degree to which strength in one dimension can compensate for weakness in another. For the edge case $M_i = 0$, all three component scores equal zero, their smoothed geometric mean equals δ_{SR}, and the resulting Step Reward evaluates to zero.

C.4 Unified Reward and Policy Optimization

PaIPO introduces an intermediate-process supervision paradigm that is architecturally decoupled from the underlying group-relative policy optimizer. Formally, let $\mathcal{B} \in \{\mathrm{G}, \mathrm{D}\}$ index the two instantiations of the base optimizer: GRPO and DAPO, respectively.

Answer Correctness. We evaluate answer correctness based on answers deterministically extracted from the *Conclusion* stage of the generated reasoning trajectory. Formally, let $\widehat{y}_i = \mathcal{E}_{\mathrm{ans}}(o_i^{\mathrm{con}})$ denote the answer extracted from the i-th sample, and let $\mathcal{Y}^\star = \left\{y_1^\star, \ldots, y_{L_y}^\star\right\}$ denote the set of verified reference answers, with $L_y = 1$ for single-answer samples. To mitigate surface-form variability, we first apply a deterministic normalization function $\nu(\cdot)$ to both generated and reference answers. This function implements Unicode normalization, case folding, whitespace standardization, terminal punctuation removal, and canonicalization of equivalent numerical expressions, while preserving all semantically salient content such as quantities, units, and negations. We then compute an exact-match indicator for each reference answer ℓ:

$$m_{i,\ell}^{\mathrm{EM}} = \mathbb{I}\left[\nu(\widehat{y}_i) = \nu(y_\ell^\star)\right]. \tag{48}$$

Exact matching provides near-perfect precision but suffers from low recall, as it cannot recognize semantically equivalent paraphrases or alternative phrasings. To address this limitation, we supplement exact matching with a semantic equivalence score computed using the same frozen textual NLI verifier reused from the Step Reward module. To avoid evaluating short answers out of context, we first convert each question-answer pair into a full declarative statement via the $\mathrm{Statement}(q, y)$ operator. For the generated answer and each reference answer, we define:

$$a_i = \mathrm{Statement}(q, \widehat{y}_i), \qquad a_\ell^\star = \mathrm{Statement}(q, y_\ell^\star).$$

We then compute a bidirectional NLI equivalence score as the geometric mean of forward and reverse entailment probabilities:

$$e_{i,\ell}^{\mathrm{NLI}} = \sqrt{p_\psi\left(a_i \Rightarrow a_\ell^\star\right) p_\psi\left(a_\ell^\star \Rightarrow a_i\right)}. \tag{49}$$

This bidirectional design serves two key purposes: forward entailment confirms the generated answer justifies the reference answer, while reverse entailment excludes answers that are too general, too specific, or semantically partial. The resulting score lies in $[0, 1]$ and is high only when the two answers are semantically comparable in context.

We combine the two signals into an exact-match-anchored correctness score:

$$s_i^{\mathrm{ans}} = \begin{cases} \max\limits_{\ell \in [L_y]} \left[m_{i,\ell}^{\mathrm{EM}} + \left(1 - m_{i,\ell}^{\mathrm{EM}}\right) e_{i,\ell}^{\mathrm{NLI}}\right], & \widehat{y}_i \text{ is valid}, \\ 0, & \text{otherwise}, \end{cases} \tag{50}$$

where $[L_y] = \{1, \ldots, L_y\}$. This formulation has three desirable properties: (1) an exact match automatically yields full correctness, prioritizing lexical precision; (2) when exact matching fails, semantic equivalence provides a robust fallback; (3) taking the maximum over references implements a disjunctive evaluation rule appropriate for multi-answer benchmarks. This design avoids penalizing semantically correct paraphrases for lacking surface-form identity with the reference.

The final binary answer-correctness reward is:

$$r_i^{\mathrm{ans}} = \mathbb{I}\left[s_i^{\mathrm{ans}} \geq \tau_{\mathrm{ans}}\right], \qquad \tau_{\mathrm{ans}} \in (0, 1), \tag{51}$$

By construction, exact matches always pass the threshold, while non-exact answers require strong bidirectional semantic equivalence. Invalid, empty, or unparseable answers receive a reward of 0.

Unified Process Reward. All three constituent process rewards (*i.e.*, Caption Perception Reward (CPR), Reasoning Grounding Reward (RGR), and Step Reward (SR)) are normalized to the unit interval $[0, 1]$, and each captures a complementary facet of reasoning process quality. Let $\mathcal{D}_{\mathrm{proc}} = \{\mathrm{CPR}, \mathrm{RGR}, \mathrm{SR}\}$ index the three components.

To aggregate them into a single unified process reward, we adopt a smoothed weighted geometric mean approach. This design addresses two competing concerns: first, avoiding complete reward

annihilation when any single component returns zero, and second, retaining the conjunctive nature of geometric aggregation to limit cross-dimensional compensation. We first apply an affine smoothing transformation to each raw component:

$$\widetilde{r}_i^d = \delta_{\text{proc}} + (1 - \delta_{\text{proc}})\, r_i^d, \qquad 0 < \delta_{\text{proc}} < 1. \tag{52}$$

where δ_{proc} is a smoothing parameter that sets the minimum value of each smoothed component. The unified process reward is then obtained by rescaling the weighted geometric mean of the smoothed components back to the unit interval:

$$r_i^{\text{proc}} = \frac{\prod_{d \in \mathcal{D}_{\text{proc}}} \left(\widetilde{r}_i^d\right)^{\alpha_d} - \delta_{\text{proc}}}{1 - \delta_{\text{proc}}}, \qquad \alpha_d > 0, \qquad \sum_{d \in \mathcal{D}_{\text{proc}}} \alpha_d = 1. \tag{53}$$

where α_d are component weights that control the relative importance of each process dimension. This aggregation scheme has three key properties: (1) the smoothing parameter ensures that no single zero component can fully erase the signal from the other dimensions; (2) the geometric structure inherently limits compensation, *e.g.*, strong performance in caption perception cannot fully offset deficiencies in spatial grounding or step-level reasoning; (3) the formulation is mathematically well-behaved, which preserves the relative ordering of component scores, maps an all-zero component vector to zero reward, and is guaranteed to lie within $[0, 1]$.

Trajectory Reward. Let $r_i^{\text{fmt}} \in \{0, 1\}$ denote a binary format compliance indicator, which takes a value of 1 if and only if the generated completion adheres to the required four-stage output structure. We frame answer correctness and intermediate process quality as two independent supervision signals. This design ensures that completions with incorrect final answers can obtain partial credit for valid visual perception, accurate object grounding, or well-aligned intermediate reasoning steps. Conversely, completions with correct answers do not automatically receive high process quality scores. The total trajectory reward under optimizer $\mathcal{B}$ is

$$R_i^{(\mathcal{B})} = \lambda_{\text{ans}} r_i^{\text{ans}} + \lambda_{\text{fmt}} r_i^{\text{fmt}} + \lambda_{\text{proc}} r_i^{\text{proc}} + r_{i,\text{aux}}^{(\mathcal{B})}, \tag{54}$$

where $\lambda_{\text{ans}}, \lambda_{\text{fmt}}, \lambda_{\text{proc}} \geq 0$ are fixed non-negative weight coefficients for each reward component. The optimizer-specific auxiliary reward term is defined as:

$$r_{i,\text{aux}}^{(\mathcal{B})} = \begin{cases} 0, & \mathcal{B} = \text{G}, \\ \lambda_{\text{len}} r_i^{\text{len}}, & \mathcal{B} = \text{D}, \end{cases} \qquad \lambda_{\text{len}} \geq 0, \qquad r_i^{\text{len}} \leq 0, \tag{55}$$

where r_i^{len} is the overlong-response penalty inherited from DAPO.

Under this formulation, incorrect answers forfeit the full answer-correctness reward but can still accrue process reward proportional to their verified intermediate quality. This partial credit is not unconditional: the terminal answer entailment component of SR explicitly penalizes trajectories whose final reasoning step fails to provide semantic support for the reference answer. CPR and RGR, by contrast, can still reward sound visual perception and spatial grounding even when subsequent inferences or the final conclusion are erroneous.

Group-relative Advantage. For a rollout group consisting of G completions, we first compute the sample mean and standard deviation of the unified trajectory rewards:

$$\mu_R^{(\mathcal{B})} = \frac{1}{G} \sum_{j=1}^{G} R_j^{(\mathcal{B})}, \qquad \sigma_R^{(\mathcal{B})} = \sqrt{\frac{1}{G} \sum_{j=1}^{G} \left(R_j^{(\mathcal{B})} - \mu_R^{(\mathcal{B})}\right)^2}. \tag{56}$$

The group-relative advantage of completion i is

$$A_i^{(\mathcal{B})} = \frac{R_i^{(\mathcal{B})} - \mu_R^{(\mathcal{B})}}{\sigma_R^{(\mathcal{B})} + \epsilon_{\text{adv}}}, \qquad \epsilon_{\text{adv}} > 0. \tag{57}$$

Since $R_i^{(\mathcal{B})}$ incorporates independently evaluated process quality, completions with identical answer correctness labels can still receive distinct advantage values. Groups where all sampled answers

are incorrect can still provide informative training signals, as long as completions vary in caption perception quality, object grounding accuracy, or reasoning step quality. This property allows PaIPO to reinforce desirable intermediate behaviors even before a fully correct trajectory is sampled. When all trajectory rewards within a group are identical, all advantage values equal zero and the group contributes no policy gradient signal. For the DAPO instantiation, dynamic sampling is therefore applied to the variance of the unified trajectory reward, rather than to answer correctness alone. Formally, a rollout group $\mathcal{G}$ is retained for training if:

$$\mathrm{Keep}(\mathcal{G}) = \mathbb{I}\left[\sigma_R^{(\mathrm{D})} > \tau_{\mathrm{dyn}}\right], \qquad \tau_{\mathrm{dyn}} \geq 0, \tag{58}$$

where τ_{dyn} is a fixed numerical tolerance. Consequently, an all-correct or all-incorrect group is retained whenever its process rewards induce a meaningful relative preference.

Policy Optimization. For each generated token $o_{i,t}$, the importance ratio is defined as:

$$\rho_{i,t}(\theta) = \frac{\pi_\theta\left(o_{i,t} \mid q, I, o_{i,<t}\right)}{\pi_{\theta_{\mathrm{old}}}\left(o_{i,t} \mid q, I, o_{i,<t}\right)}, \qquad \widetilde{\rho}_{i,t}^{(\mathcal{B})}(\theta) = \mathrm{clip}\left(\rho_{i,t}(\theta), 1-\varepsilon_{\mathrm{low}}^{(\mathcal{B})}, 1+\varepsilon_{\mathrm{high}}^{(\mathcal{B})}\right). \tag{59}$$

GRPO employs symmetric clipping, such that $\varepsilon_{\mathrm{low}}^{(\mathrm{G})} = \varepsilon_{\mathrm{high}}^{(\mathrm{G})}$, whereas DAPO allows for asymmetric lower and upper clipping bounds. The token-level clipped surrogate objective is then given by:

$$\mathcal{U}_{i,t}^{(\mathcal{B})}(\theta) = \min\left(\rho_{i,t}(\theta)A_i^{(\mathcal{B})}, \widetilde{\rho}_{i,t}^{(\mathcal{B})}(\theta)A_i^{(\mathcal{B})}\right). \tag{60}$$

PaIPO preserves the loss reduction scheme of the chosen base optimizer. For GRPO, the clipped surrogate terms are first averaged within each completion and then across the rollout group:

$$\mathcal{L}_{\mathrm{G}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\mathcal{U}_{i,t}^{(\mathrm{G})}(\theta). \tag{61}$$

For GRPO, we additionally apply reference-policy KL regularization. Let $\widehat{d}_{i,t}^{\mathrm{KL}}(\theta)$ denote the token-level KL estimator. The completion-level aggregation is

$$\mathcal{K}_{\mathrm{G}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\widehat{d}_{i,t}^{\mathrm{KL}}(\theta). \tag{62}$$

DAPO instead aggregates over all valid generated tokens in each retained group:

$$\mathcal{L}_{\mathrm{D}}(\theta) = \frac{\displaystyle\sum_{i=1}^{G}\sum_{t=1}^{T_i}\mathcal{U}_{i,t}^{(\mathrm{D})}(\theta)}{\displaystyle\sum_{i=1}^{G}T_i}. \tag{63}$$

where T_i denotes the number of valid generated tokens in the i-th completion. The overall training objective is:

$$\begin{aligned}\mathcal{J}_{\mathrm{PaIPO_G}}(\theta) &= \mathbb{E}\left[\mathcal{L}_{\mathrm{G}}(\theta) - \beta_{\mathrm{KL}}\mathcal{K}_{\mathrm{G}}(\theta)\right], \\ \mathcal{J}_{\mathrm{PaIPO_D}}(\theta) &= \mathbb{E}\left[\mathcal{L}_{\mathrm{D}}(\theta)\right],\end{aligned} \qquad \beta_{\mathrm{KL}} = 0.04. \tag{64}$$

with optimizer-specific clipping, regularization, and loss reduction schemes all inherited from the corresponding base method. For $\mathrm{PaIPO_D}$, the original dynamic sampling principle is maintained, but applied to the variance of the unified trajectory reward instead of answer correctness. Overall, PaIPO defines which answer-level and intermediate behaviors are credited, while the underlying base optimizer governs how these preferences are translated into policy updates.

D Additional Experimental Analyses

This appendix supplements the main experiments with detailed implementation settings and analyses of the behaviors optimized by PaIPO. Section D.1 documents the policy-training configuration, reward computation, comparison settings, and benchmark-specific evaluation protocols. Section D.2 examines the evolution of caption and reasoning stage rewards and compares the counterfactual NLL separation of PaIPO and PAPO under matched GRPO and DAPO optimization. Section D.3 analyzes post-training VAM profiles across MCoT stages and model scales. Finally, Sections D.4 and D.5 compare caption perception and reasoning grounding before and after PaIPO training.

D.1 Experimental Setup and Hyperparameters

This subsection specifies the training and evaluation configurations used throughout the experiments. We first describe the common policy-training configuration and the optimizer-specific settings for GRPO and DAPO. We then specify the reward models and scalar hyperparameters used to compute the answer and process rewards. Finally, we detail the comparison method configurations and benchmark-specific evaluation and reporting protocols.

D.1.1 Policy Training Configuration

Training Data and Policy Models. All post-trained policy variants use the same fixed subset of $10{,}000$ records from the MCoT-ViG corpus. Each record contains an image–question pair, a verified reference answer, a four-stage rationale comprising *Summary*, *Caption*, *Reasoning*, and *Conclusion*, object-level grounding annotations, and a corresponding controlled visual variant. We instantiate all methods with the 2B, 4B, and 8B variants of Qwen3-VL-Instruct (Bai et al., 2025). The four-stage output schema and rollout budget are held fixed across model scales and comparison methods. For each prompt, the current policy samples $G = 4$ candidate completions, which are scored independently before group-relative advantage estimation. We use LoRA (Hu et al., 2022) with rank 32, scaling factor 64, and dropout 0. Adapters are inserted into the query, key, value, output, gate, up-projection, and down-projection modules. All non-LoRA policy parameters remain frozen during post-training. Images are processed using the native Qwen3-VL preprocessing pipeline. The supported visual resolution ranges from 4×28^2 to 768×28^2 pixels. We use scaled dot-product attention and bfloat16 numerical precision without gradient checkpointing.

Optimization and Rollout Configuration. Each model is trained for one epoch, corresponding to $2{,}500$ optimizer steps over the training subset. Training is distributed across four NVIDIA A100-SXM4 GPUs with $80\,\mathrm{GB}$ of memory per device. The per-device batch size is 4, gradient accumulation is 1, and each optimizer step contains four distinct prompts and 16 sampled completions. We use AdamW with a learning rate of 1×10^{-5}, zero weight decay, a cosine learning-rate schedule, and a warm-up ratio of 0.03. The gradient norm is clipped at 1.0. During rollout generation, we use a temperature of 0.7, top-p of 0.8, and top-k of 20. The maximum completion length is 384 tokens, while prompts are not truncated to a fixed token length. The training seed is fixed at 42 for all methods. Table 10 summarizes the common and optimizer-specific configurations.

Optimizer-specific Settings. The GRPO instantiation uses symmetric importance-ratio clipping with lower and upper clipping coefficients of 0.20. The surrogate terms are first averaged over the tokens of each completion and then averaged across the rollout group. A KL coefficient of 0.04 regularizes the updated policy against the reference policy. The DAPO instantiation uses asymmetric clipping with lower and upper coefficients of 0.20 and 0.28, respectively. The surrogate objective is reduced over all valid response tokens in the rollout batch, and no additional reference-policy KL penalty is applied. DAPO additionally performs dynamic sampling according to the within-group standard deviation of the complete trajectory reward. A rollout group is treated as informative when its reward standard deviation exceeds 0.03. Groups that fail this condition are resampled using fresh prompts for at most two attempts. Because the statistic is computed from the complete reward rather than answer correctness alone, process-quality differences can retain a group even when its completions share the same answer-correctness outcome. The DAPO overlong-response penalty is inactive for completions of at most 320 tokens and increases linearly over the remaining 64-token buffer. A response reaching the 384-token limit receives the maximum penalty.

D.1.2 Reward Models and Hyperparameters

Frozen Reward Models. All auxiliary models used to compute rewards remain frozen throughout policy optimization. The atomic proposition extractor is a T5-base (Raffel et al., 2020) claim-extraction model with deterministic decoding. Its maximum input and output lengths are 512 and 128 tokens, respectively. We use one decoding beam, a repetition penalty of 1.10, a four-gram repetition constraint, and an inference batch size of 32. Textual entailment for CPR, SR, and answer evaluation is computed using a DeBERTa-v3-base (He et al., 2023) verifier trained on MNLI, FEVER, and ANLI. The verifier uses a maximum sequence length of 256, an inference batch size of 128, and bidirectional entailment scoring whenever semantic equivalence is required.

Table 5: **Reward and aggregation hyperparameters used by PaIPO.** All process reward components and the unified process reward lie in $[0, 1]$.

Component	Hyperparameter	Value
Numerical stability	ϵ_{num}	10^{-8}
CPR	Counterfactual temperature τ_{cf}	0.25
	Confidence exponent γ	2
	Counterfactual indifference margin m_{cf}	0.10
RGR	Identifier-consistency weight λ_{id}	0.25
SR	Step-matching threshold τ_{step}	0.50
	Smoothing constant δ_{SR}	0.01
	Semantic-trajectory weight λ_{sem}	0.60
	Sequential-order weight λ_{ord}	0.25
	Terminal-entailment weight λ_{ter}	0.15
Unified process reward	Smoothing constant δ_{proc}	0.01
	CPR exponent α_{CPR}	0.30
	RGR exponent α_{RGR}	0.40
	SR exponent α_{SR}	0.30
Trajectory reward	Answer weight λ_{ans}	1.00
	Format weight λ_{fmt}	0.10
	Process weight λ_{proc}	0.80
	DAPO overlong weight λ_{len}	1.00

Table 6: **Settings of the comparison methods.** Answer-correctness and format rewards are shared across all post-trained methods. Methods within each optimizer family use matched training and optimization configurations.

Method	Optimizer	Process supervision	Distinguishing component
Qwen3-VL-Instruct	–	–	Original pretrained instruction model
Step GRPO	GRPO	Step Reward	Reference-only process supervision
Step DAPO	DAPO	Step Reward	Dynamic sampling and overlong penalty
PAPO_{G}	GRPO	Textual reference alignment; SR	Perception-aware masked-image policy loss
PAPO_{D}	DAPO	Textual reference alignment; SR	PAPO loss with DAPO optimization
PaIPO_{G}	GRPO	CPR, RGR, and SR	Unified perception-aware process reward
PaIPO_{D}	DAPO	CPR, RGR, and SR	Unified reward with DAPO optimization

Answer Correctness. Answer correctness follows the exact-match anchored NLI procedure defined in Appendix C.4. The final answer is deterministically extracted from *Conclusion* and normalized through Unicode normalization, whitespace normalization, and numerical canonicalization. An exact match with any verified reference answer receives full credit. When exact matching fails, the question and answer are converted into declarative statements and compared using bidirectional NLI entailment. The geometric mean of the two entailment probabilities defines their semantic-equivalence score. A non-exact answer is accepted when this score reaches $\tau_{\text{ans}} = 0.80$. Invalid or unparseable answers receive zero reward. The resulting answer reward is binary: $r_i^{\text{ans}} \in \{0, 1\}$. Table 5 reports the scalar hyperparameters used by PaIPO. All process reward components are normalized to $[0, 1]$.

D.1.3 Comparison Settings

We compare PaIPO with the original Qwen3-VL-Instruct backbones, Step GRPO, Step DAPO, and PAPO (Wang et al., 2026). All post-trained methods use the same answer-correctness and four-stage format rewards. Within each optimizer family, they also share the policy initialization, LoRA configuration, training data, rollout budget, decoding parameters, and optimizer hyperparameters. Table 6 summarizes their intermediate-process supervision and distinguishing components.

Table 7: **Benchmark-specific evaluation protocols.** All reported metrics are expressed on a 0–100 scale. An upward arrow indicates that larger values are preferred.

Benchmark	Evaluation setting	Reported metric	Evaluation target
MMMU-Pro	Standard 10-option	10-option accuracy ↑	Correct option selection
MMStar	Official evaluation split	Accuracy ↑	Vision-dependent multimodal understanding
VRC-Bench	Answer and reasoning evaluation	Answer accuracy ↑	Final prediction correctness
VRC-Bench	Four-completion reasoning protocol	Step Score ↑	Intermediate rationale quality
VFaith-Bench	Original-image condition	Raw accuracy ↑	Original visual reasoning
VFaith-Bench	Edited-image condition	Edit accuracy ↑	Adaptation to modified visual evidence
VFaith-Bench	Paired comparison	Δ_{pp}	Change from raw to edited accuracy
SPD-Faith	Symmetric-query evaluation	CR ↑	Cross-prompt reasoning consistency
SPD-Faith	Difference-description evaluation	DRF ↑	Faithfulness to actual visual differences
SPD-Faith	Combined evaluation	Average ↑	Mean of CR and DRF

Qwen3-VL Backbones. The backbone rows use the original Qwen3-VL-Instruct checkpoints without post-training. They are evaluated using the same four-stage prompt, image preprocessing, and decoding configuration as the post-trained policies.

Step GRPO and Step DAPO. The Step baselines combine answer-correctness and format rewards with reference-based supervision for the intermediate MCoT stages. For *Summary*, a frozen textual verifier measures bidirectional entailment between the generated and reference summaries. For *Caption*, a frozen atomic extractor decomposes the generated and reference captions into visual propositions. Their semantic correspondence is evaluated through bidirectional entailment, followed by candidate-precision and reference-coverage aggregation. For *Reasoning*, the baselines use the SR defined in Appendix C.3. These baselines therefore provide reference-based textual supervision but do not compare the original and controlled images or evaluate predicted grounding regions. Step GRPO and Step DAPO differ only in the optimizer-specific settings reported in Table 10.

PAPO Variants. PAPO augments the same answer, format, and reference-based process rewards used by the corresponding Step baseline with a perception-aware policy loss. This auxiliary objective compares token likelihoods under the original image and a randomly masked visual input. The masked-image branch re-evaluates the same completion through teacher forcing and requires no additional rollout. We use a visual-patch masking probability of 0.60. The perception-KL coefficient is 0.02 for $\mathrm{PAPO_G}$ and 0.01 for $\mathrm{PAPO_D}$. Both variants use the low-variance KL estimator, with log-ratio clipping at 20 and KL-value clipping at 10. The entropy coefficients for the original and masked branches are both 0.03. Unlike PaIPO, the perception-aware component of PAPO modifies the policy objective rather than contributing to the group-relative trajectory reward.

PaIPO Variants. The PaIPO variants replace the reference-only intermediate reward configuration with the unified process reward comprising CPR, RGR, and SR. CPR retains reference-based Caption alignment while adding counterfactual visual discrimination between the original and controlled images. RGR evaluates spatial agreement with annotated target regions and cross-step identifier consistency. SR preserves semantic, sequential, and terminal supervision over the generated reasoning trajectory. The answer-correctness and format rewards remain unchanged. $\mathrm{PaIPO_G}$ otherwise follows the Step-GRPO optimization configuration, while $\mathrm{PaIPO_D}$ follows the Step-DAPO configuration. These matched comparisons isolate the effect of perception-aware intermediate-process supervision from differences attributable to the underlying group-relative optimizer.

D.1.4 Benchmark-specific Evaluation

Table 7 summarizes the evaluated capability and principal metric for each benchmark.

MMMU-Pro and MMStar. We evaluate MMMU-Pro (Yue et al., 2025) under the Standard setting containing ten candidate options per question. MMStar (Chen et al., 2024) is evaluated on the official vision-dependent multiple-choice split. For both benchmarks, the answer extracted from *Conclusion* is mapped to a canonical option identifier. Accuracy is computed as

$$\mathrm{Acc} = \frac{100}{N} \sum_{n=1}^{N} \mathbb{I}\left[\widehat{y}_n = y_n\right], \tag{65}$$

where $\widehat{y}_n$ and y_n denote the predicted and reference option identifiers.

VRC-Bench. VRC-Bench (Thawakar et al., 2025) evaluates both final-answer correctness and the quality of the generated reasoning process. Answer accuracy is computed using the benchmark evaluator after extracting the final prediction from *Conclusion*. For Step Score, four completions are sampled for each question and assessed using the ten benchmark-specified rationale-quality criteria. Let $s_{n,g,c} \in [0,1]$ denote the normalized score assigned to criterion c for completion g of question n. The reported Step Score is

$$\mathrm{StepScore} = \frac{100}{N G_{\mathrm{eval}} |\mathcal{C}_{\mathrm{VRC}}|} \sum_{n=1}^{N} \sum_{g=1}^{G_{\mathrm{eval}}} \sum_{c \in \mathcal{C}_{\mathrm{VRC}}} s_{n,g,c}, \tag{66}$$

where $G_{\mathrm{eval}} = 4$ and $|\mathcal{C}_{\mathrm{VRC}}| = 10$. The averaging is performed uniformly over questions, sampled completions, and evaluation criteria.

VFaith-Bench. VFaith-Bench (Yu et al., 2025a) pairs each original image with an edited image in which question-relevant visual evidence has been modified. Raw accuracy evaluates the prediction obtained from the original image against the original reference answer:

$$\mathrm{RawAcc} = \frac{100}{N} \sum_{n=1}^{N} \mathbb{I}\left[\widehat{y}_n^{\mathrm{raw}} = y_n^{\mathrm{raw}}\right]. \tag{67}$$

Edit accuracy analogously evaluates predictions under the edited image and the updated reference answer:

$$\mathrm{EditAcc} = \frac{100}{N} \sum_{n=1}^{N} \mathbb{I}\left[\widehat{y}_n^{\mathrm{edit}} = y_n^{\mathrm{edit}}\right]. \tag{68}$$

The paired accuracy difference is reported in percentage points:

$$\Delta_{\mathrm{pp}} = \mathrm{EditAcc} - \mathrm{RawAcc}, \qquad \Delta_{\mathrm{pp}} \in [-100, 100]. \tag{69}$$

A less negative Δ_{pp} indicates a smaller performance reduction following the visual intervention. Raw and Edit accuracy remain the primary correctness measures since the difference alone does not identify the absolute performance level.

SPD-Faith. SPD-Faith (Lv et al., 2026) evaluates whether reasoning about differences between paired images remains internally consistent and grounded in the actual modifications. We report the two benchmark dimensions directly associated with reasoning faithfulness. Consistency Ratio (CR) measures whether answers to complementary queries remain mutually consistent. Difference Reasoning Faithfulness (DRF) evaluates whether the generated reasoning is supported by the annotated visual differences rather than by an unsupported shortcut. Both metrics are computed using the released benchmark evaluators and reported on a 0–100 scale. Since our analysis focuses on reasoning faithfulness, we additionally report their arithmetic mean.

Repeated Evaluation and Reporting. For each benchmark record, we obtain $K_{\mathrm{eval}} = 3$ independent stochastic outputs under the same checkpoint, prompt, and decoding hyperparameters. For every metric, we first compute the dataset-level score within each repetition and then report the mean and standard deviation over the three repetitions. For VRC-Bench, each repetition independently executes the four-completion Step Score protocol. All LLM-based benchmark evaluators are frozen, use deterministic decoding, and remain identical across comparison methods. No benchmark test example is used for reward calibration or post-training. The benchmark-overlap controls described in Appendix B.1 are applied before constructing the training subset.

D.2 Reward Dynamics

We examine the training dynamics of the stage-specific process signals on Qwen3-VL-4B-Instruct under the GRPO and DAPO configurations in Figure 7. **The upper-left panel** reports the *Caption* reward trajectories. Both PaIPO variants exhibit sustained reward growth, while Step GRPO and Step DAPO remain close to their initial levels. In particular, $\mathrm{PaIPO_D}$ increases rapidly during

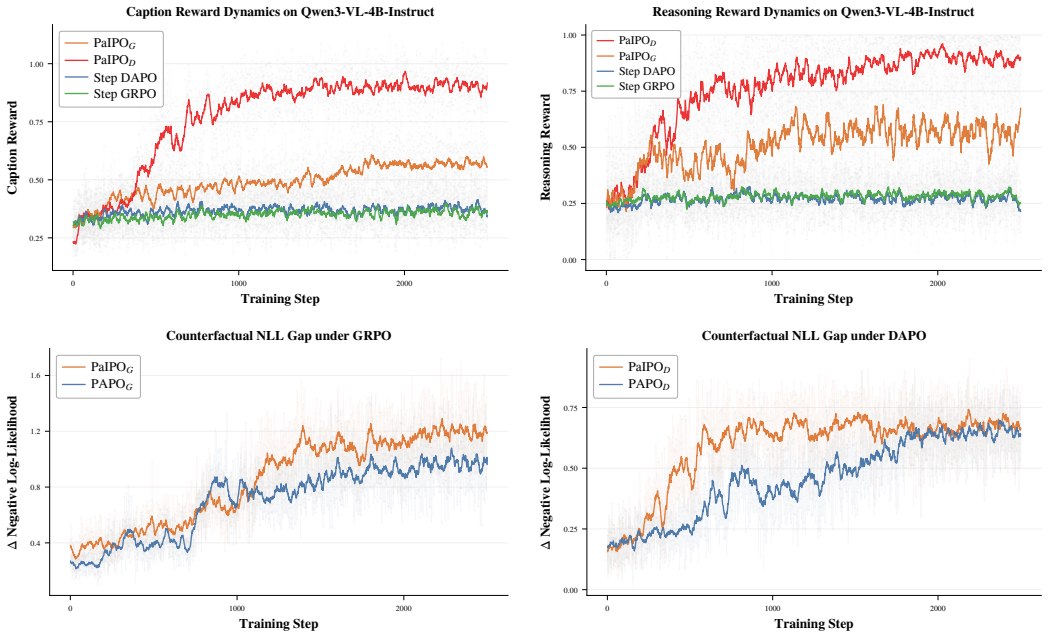


Figure 7: **Reward and counterfactual likelihood dynamics on Qwen3-VL-4B-Instruct.** The upper panels compare the *Caption* and *Reasoning* reward trajectories of PaIPO with the corresponding Step baselines. The lower panels compare the counterfactual NLL gaps of PaIPO and PAPO under GRPO and DAPO, respectively. A positive counterfactual NLL gap indicates that the generated caption is more strongly supported by the original visual evidence.

the early stage of training and subsequently stabilizes at the highest reward level. **The upper-right panel** presents a similar pattern for the *Reasoning* reward. The Step baselines remain nearly stationary throughout training, while both PaIPO variants progressively increase their reasoning rewards. $\mathrm{PaIPO_D}$ shows the fastest increase and the highest late-stage reward, while $\mathrm{PaIPO_G}$ achieves a smaller but persistent improvement over its matched GRPO baseline. These trajectories show that PaIPO strengthens both caption-level perception and reasoning-level grounding during policy optimization.

The lower panels compare PaIPO and PAPO using the counterfactual negative-log-likelihood gap. Positive values indicate that the generated caption is assigned a higher likelihood under the original image than under the counterfactual visual input. Under GRPO, as shown in **the lower-left panel**, the counterfactual NLL gap increases for both methods, but $\mathrm{PaIPO_G}$ develops a larger separation during the later stages of training. Under DAPO, the **lower-right panel** shows that $\mathrm{PaIPO_D}$ develops visual discrimination earlier than $\mathrm{PAPO_D}$. The two trajectories become comparable near the end of training, indicating that the principal PaIPO advantage under DAPO lies in faster early optimization. Overall, the reward and likelihood dynamics indicate that PaIPO provides earlier and stronger process-level learning signals under both optimizers.

D.3 Post-training Visual Attention

We examine whether the perception-aware improvements observed in the reward trajectories are accompanied by changes in direct visual-attention allocation after policy training. Using the same four-stage generation protocol and VAM computation described in Section 2.1, we evaluate the PaIPO-trained Qwen3-VL-Instruct models at the 2B, 4B, and 8B scales. Figure 8 reports the token-level observations, smoothed stage-wise trajectories, and overall mean VAM for each model scale. Across all three models, *Caption* continues to receive the largest concentration of direct visual attention. More importantly, after the transient reduction at the transition from *Caption* to *Reasoning*, reasoning-stage VAM recovers and remains at a comparatively elevated level throughout most of the reasoning trajectory. Compared with the corresponding backbone trajectories in Figure 1, the post-trained models exhibit higher overall VAM and a more persistent allocation of direct visual attention during *Reasoning*. Direct visual attention nevertheless remains lower during *Reasoning*

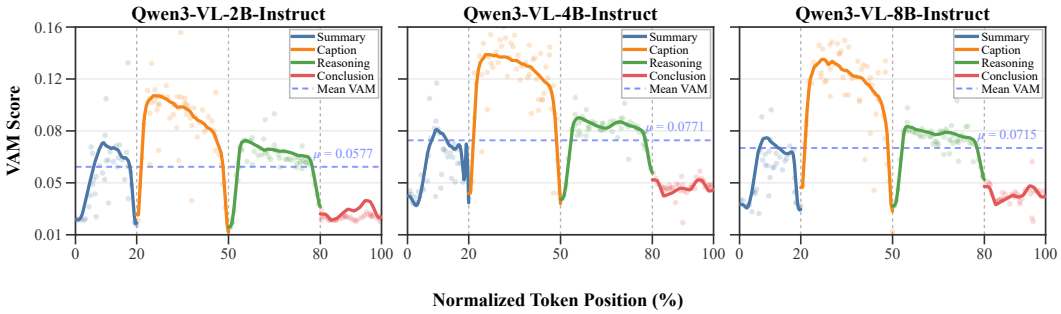


Figure 8: **Post-training VAM profiles across Qwen3-VL model scales.** Across model scales, direct visual attention remains concentrated during *Caption* and is sustained through a larger portion of *Reasoning*.

Table 8: **Caption perception before and after PaIPO training.** Component and overall scores are reported on a 1–10 scale and averaged over the evaluation records. Paired change denotes the mean record-level difference in overall caption perception relative to the corresponding Qwen3-VL-Instruct backbone.

Backbone	Checkpoint	Visual faithfulness	Evidence coverage	Perceptual specificity	Overall caption perception	Paired change (95% CI)
Qwen3-VL-2B	-	6.14	6.04	5.73	5.95	–
	PaIPO_G	6.91	6.67	6.58	6.61	+0.67 [0.61, 0.72]
	PaIPO_D	7.07	6.76	6.85	6.79	+0.84 [0.79, 0.89]
Qwen3-VL-4B	-	6.67	6.51	6.25	6.45	–
	PaIPO_G	7.34	7.13	7.03	7.07	+0.62 [0.57, 0.67]
	PaIPO_D	7.51	7.25	7.19	7.23	+0.78 [0.73, 0.83]
Qwen3-VL-8B	-	7.09	6.91	6.69	6.88	–
	PaIPO_G	7.66	7.44	7.40	7.41	+0.53 [0.48, 0.58]
	PaIPO_D	7.79	7.51	7.52	7.52	+0.64 [0.59, 0.69]

than during *Caption*, and a further decrease occurs upon entering *Conclusion*. Since VAM measures attention allocation rather than causal visual influence, these results are interpreted as evidence of a post-training redistribution of direct visual attention, not as proof that the attention changes causally determine reasoning improvements.

D.4 Caption Perception Before and After Training

We further evaluate whether PaIPO improves the perceptual quality of the generated *Caption* on VRC-Bench. For each image–question pair, the backbone and post-trained checkpoints generate a complete four-stage trajectory under the same prompt and decoding configuration. A deterministic parser then extracts the generated *Caption*. We employ a multimodal judge model with a caption-specific evaluation prompt. The scoring format is adapted from the model-based evaluation protocol of MM-Vet (Yu et al., 2024), but the evaluated dimensions and scoring anchors are redesigned specifically for question-relevant visual captions. For each record, the judge receives the original image, the question, and the extracted caption. The judge model assigns four integer scores on a 1–10 anchored scale: (1) *Visual faithfulness* measures whether the caption's visual claims are supported by the image and free from hallucinated content; (2) *Evidence coverage* measures whether the caption includes the objects, attributes, and relations needed to address the question; (3) *Perceptual specificity* measures whether the caption provides concrete and visually discriminative details rather than generic descriptions; (4) *Overall caption perception* provides a holistic assessment of whether the caption is visually faithful, sufficiently detailed, and perceptually complete. A score of 1 indicates that the criterion is almost entirely unsatisfied, while a score of 10 indicates complete satisfaction.

Table 8 shows consistent improvements after PaIPO training across all model scales and evaluation dimensions. The paired overall improvements range from $+0.53$ to $+0.67$ for PaIPO_G and from $+0.64$ to $+0.84$ for PaIPO_D. The largest component improvements occur in perceptual specificity. Under PaIPO_D, specificity increases by 1.12, 0.94, and 0.83 for the 2B, 4B, and 8B models, respec-

Table 9: **Reasoning grounding before and after PaIPO training.** All metrics are computed from bounding boxes extracted from complete four-stage trajectories. Scores are reported on a 0–100 scale. The paired change in $F_1@0.5$ is measured relative to the corresponding backbone before policy training, together with the 95% paired confidence interval.

Backbone	Checkpoint	mIoU ↑	$P@0.5$ ↑	$R@0.5$ ↑	$F_1@0.5$ ↑	$F_1@0.75$ ↑	Paired change in $F_1@0.5$ (95% CI)
Qwen3-VL-2B	–	27.26	38.18	31.63	34.60	19.74	–
	$\mathrm{PaIPO_G}$	29.94	40.34	34.72	37.32	21.45	+2.72 [2.24, 3.22]
	$\mathrm{PaIPO_D}$	31.48	42.82	35.63	38.90	23.67	+4.30 [3.74, 4.88]
Qwen3-VL-4B	–	31.12	43.58	37.21	40.14	22.72	–
	$\mathrm{PaIPO_G}$	33.23	45.44	40.87	43.03	23.69	+2.89 [2.39, 3.41]
	$\mathrm{PaIPO_D}$	34.71	46.86	41.51	44.02	25.86	+3.88 [3.36, 4.43]
Qwen3-VL-8B	–	34.36	45.82	40.73	43.13	24.23	–
	$\mathrm{PaIPO_G}$	35.68	47.91	42.34	44.95	25.61	+1.83 [1.40, 2.24]
	$\mathrm{PaIPO_D}$	36.91	48.02	43.03	45.39	27.48	+2.26 [1.81, 2.72]

tively. These results indicate that PaIPO improves question-relevant caption perception, consistent with the objective of CPR: to explicitly reward captions that distinguish the original visual evidence from its controlled counterpart.

D.5 Reasoning Grounding Before and After Training

To evaluate whether PaIPO improves the spatial grounding of visual references during *Reasoning*, we construct an evaluation set of 1,000 samples. This set comprises the 240 MCoT-ViG records excluded from the fixed policy-training subset and 760 additional records sampled from the source Visual-CoT dataset. None of these evaluation records overlap with the policy-training data. For each image–question pair, the models generate a complete four-stage trajectory under the same prompt and decoding configuration. A deterministic parser extracts the coordinate-bearing bounding boxes from the generated *Reasoning*. We evaluate these predictions against the annotated target boxes using mean intersection over union (mIoU), precision, recall, and F_1 at an IoU threshold of 0.5, together with F_1 at the stricter threshold of 0.75. Table 9 shows that both PaIPO variants improve every grounding metric across all three model scales. For instance, $\mathrm{PaIPO_G}$ improves $F_1@0.5$ by $+2.72$, $+2.89$, and $+1.83$ for the 2B, 4B, and 8B backbones under the GRPO optimizer. The improvements arise from concurrent improvements in precision and recall, indicating that PaIPO both reduces unsupported predictions and recovers a larger proportion of the annotated targets. Improvements also persist under the stricter localization criterion. In particular, $\mathrm{PaIPO_D}$ increases $F_1@0.75$ by $+3.93$, $+3.14$, and $+3.25$ at the 2B, 4B, and 8B scales, demonstrating more accurate spatial alignment rather than only coarse target localization. Overall, these results are consistent with the intended effect of PaIPO: explicit process-level grounding supervision strengthens the connection between object references in the generated reasoning trajectory and their supporting visual regions.

E Related Work

Our work connects structured multimodal reasoning, reinforcement learning with process supervision, and visual faithfulness. We organize the discussion around how prior methods represent intermediate reasoning, supervise the rationale quality, and assess its dependence on visual evidence.

E.1 Structured multimodal reasoning and grounded rationales

Multimodal chain-of-thought (MCoT) reasoning exposes intermediate reasoning steps by incorporating visual signals into rationale generation. Multimodal-CoT (Zhang et al., 2024) separates rationale generation from answer prediction, while LLaVA-CoT (Xu et al., 2025) organizes generation into *Summary*, *Caption*, *Reasoning*, and *Conclusion* and introduces stage-wise retracing at inference time. Such methods offer principled frameworks to disentangle visual interpretation and logical deduction. The well-defined stage boundaries of structured MCoT enable uniform comparisons of visual attention across reasoning trajectories, alongside targeted supervision over perceptual de-

scriptions and their downstream utilization in reasoning. For this reason, we adopt LLaVA-CoT's four-stage paradigm as a unified interface for analysis and supervision.

Grounded reasoning datasets provide a complementary foundation. Visual CoT (Shao et al., 2024a) supplies intermediate bounding-box annotations and a multi-turn pipeline for focusing on relevant image regions. Built upon the samples of Visual CoT and target-object annotations, MCoT-ViG augments the dataset with four-stage rationales, controllable visual variants, and explicit cross-step object references. This dataset design is uniquely tailored to facilitate auxiliary reward computation: original–variant image pairs evaluate the visual sensitivity of the *Caption* stage, while object annotations validate spatial grounding and reference consistency within the *Reasoning* stage.

E.2 Reinforcement Learning and Intermediate Process Supervision

Reinforcement learning with verifiable rewards offers a pragmatic framework for enhancing reasoning capabilities via automatically computed feedback signals. GRPO, first introduced in DeepSeekMath (Shao et al., 2024b), estimates relative advantages from batches of sampled completions without requiring a separate critic network. DAPO (Yu et al., 2025b) further advances optimization and sampling techniques tailored for large-scale reasoning reinforcement learning. While these algorithms define the mechanics of policy updates, the inductive biases forged during training are fundamentally shaped by reward design. PaIPO occupies this complementary design space: it defines perception-aware intermediate rewards while preserving the underlying optimization backbone of GRPO or DAPO.

Process supervision broadens the scope of feedback beyond binary final-answer correctness. R1-VL (Zhang et al., 2025) proposes StepGRPO, which assigns rewards based on key-step coverage and reasoning validity, incentivizing intermediate trajectories to include all necessary steps and follow a coherent logical progression. PaIPO aligns with this core objective of intermediate reasoning supervision. The Step Reward metric of PaIPO evaluates three complementary dimensions: semantic trajectory alignment, sequential order consistency, and entailment of the terminal answer. A key distinguishing feature of PaIPO is the emphasis on explicit visual verification. Specifically, the Caption Perception Reward (CPR) verifies that a semantically appropriate caption is preferentially grounded in the original image, while the Reasoning Grounding Reward (RGR) confirms that reasoning references correspond accurately to annotated objects and image regions. The unified process reward employs a smoothed geometric aggregation mechanism to mitigate undesired trade-offs across three components: semantic reasoning quality, caption perception, and reasoning grounding. This design addresses a well-known limitation of reference-based textual supervision: alignment with a gold-standard rationale does not inherently guarantee sensitivity to the visual evidence that underpins that rationale.

E.3 Perception-Aware Optimization and Spatial Grounding

Visual supervision has already been incorporated into reinforcement learning through both task rewards and grounded reasoning representations. Visual-RFT (Liu et al., 2025) devises verifiable reward signals for visual tasks, notably including IoU-based rewards for object detection. Point-RFT (Ni et al., 2025) combines format fine-tuning on visually grounded rationales with reinforcement fine-tuning for visual document understanding. These works collectively validate the effectiveness of spatially grounded outputs and visual feedback. PaIPO advances this line of research by assigning distinct verification roles to the *Caption* and *Reasoning* stages. In particular, the RGR applies localization-weighted one-to-one object matching to references extracted from reasoning trajectories, while also accounting for the consistency of object identifiers across subsequent steps. The RGR jointly evaluates four dimensions within the reasoning process: target coverage, unsupported references, localization accuracy, and cross-step co-reference.

PAPO (Wang et al., 2026) is especially closely related in the objective of strengthening visual dependence. It introduces an implicit perception loss that amplifies the divergence between output distributions conditioned on original and corrupted visual inputs, paired with entropy regularization to stabilize training. This mechanism operates without requiring additional perception annotations or external reward models. PaIPO instead uses explicit stage-specific supervision: the CPR compares the likelihood of an identical generated caption given the original image versus a controlled attribute-edited variant, integrates this comparison with semantic reference alignment, and feeds the

resulting signal into the trajectory reward used to compute group-relative advantages. The RGR further provides annotated spatial targets that cannot be captured by a distributional divergence objective alone. The distinction is therefore both computational and supervisory: PAPO modifies the policy objective using visual corruption, while PaIPO constructs intermediate rewards from controlled visual comparisons, reference semantics, and object annotations.

E.4 Visual Faithfulness, Attention Diagnostics, and Evaluation

Visual faithfulness describes the degree to which intermediate reasoning claims are grounded in visual evidence, a property that cannot be inferred from final-answer accuracy alone. Uppaal et al. (2026) evaluate the visual faithfulness of perception steps and introduce training-free self-reflection to revise unsupported content. Jia et al. (2026) investigate degradation in visual context utilization and decouple reasoning from perception via an agentic framework that actively queries visual evidence during inference. SPD-Faith (Lv et al., 2026) diagnoses perceptual blindness and perception–reasoning dissociation, analyzes degraded visual attention, and introduces a training-free evidence-calibration method. Our VAM analysis is related to these diagnoses but focuses on direct text-to-visual attention across explicitly separated generation stages and its association with intermediate reasoning quality across model scales. VAM serves as a diagnostic rather than a training reward or a causal measure of visual reliance. PaIPO addresses the observed problem through policy training with explicit perception and grounding rewards.

Counterfactual evaluation further tests whether models adapt their predictions to changed visual evidence. VFaith-Bench (Yu et al., 2025a) manipulates answer-relevant visual cues, while SPD-Faith (Lv et al., 2026) leverages fine-grained image-difference reasoning to examine faithfulness beyond response correctness. We adopt these benchmarks for external evaluation, and additionally construct dedicated controlled visual variants within MCoT-ViG for training supervision. The two serve distinct purposes: benchmark interventions evaluate behavioral adaptation to visual changes, while CPR fixes the generated caption and quantifies its relative evidential support under two contrasting visual conditions. Combined with reasoning quality and grounding evaluations, these tests capture complementary dimensions of the targeted improvements. Neither stronger attention nor accurate region references alone guarantees faithful reasoning. PaIPO therefore combines visual sensitivity, spatial grounding, and semantic process supervision.

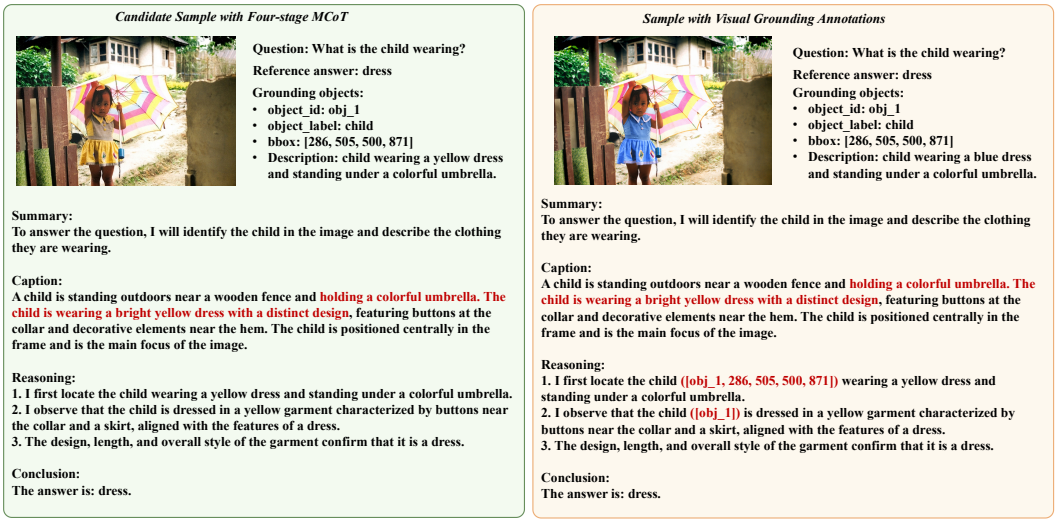


Figure 9: **MCoT-ViG record for object-category recognition.** The left panel presents the original image and its four-stage rationale. The right panel pairs an answer-preserving dress-color variant with the grounding-annotated version of the same rationale. The controlled edit changes the dress from yellow to blue, while the reference answer remains *dress*.

Candidate Sample with Four-stage MCoT



Question: Is the location where the man is filling up his car an indoor or outdoor setting?

Grounding objects:
- **object_id: obj_1**
- **object_label: man**
- **bbox: [2,10,338,1000]**
- **Description: a man wearing a light-colored shirt and dark pants, standing next to a car while holding a fuel pump.**

Summary:
To determine whether the location where the man is filling up his car is an indoor or outdoor setting, I will examine the surroundings for features indicative of either environment.

Caption:
A man wearing a light-colored shirt and dark pants is standing next to a black car while holding a fuel pump. The car is parked at a gas station, which is an open area with visible fuel pumps and a concrete ground. The background shows a wall with advertisements and an open sky above, indicating an outdoor setting.

Reasoning:
1. The man is standing next to a black car while holding a fuel pump, indicating that he is refueling the vehicle.
2. The surrounding environment includes visible fuel pumps, a concrete ground, and a wall with advertisements.
3. The open sky above the scene and the lack of enclosing walls or a roof suggest that the location is outdoors.
4. Based on these observations, it is evident that the man is filling up his car in an outdoor setting at a gas station.

Conclusion:
The answer is: It is an outdoor setting; the man is at a gas station.

Sample with Visual Grounding Annotations



Question: Is the location where the man is filling up his car an indoor or outdoor setting?

Grounding objects:
- **object_id: obj_1**
- **object_label: man**
- **bbox: [2,10,338,1000]**
- **Description: a man wearing a light-colored shirt and dark pants, standing next to a car while holding a fuel pump.**

Summary:
To determine whether the location where the man is filling up his car is an indoor or outdoor setting, I will examine the surroundings for features indicative of either environment.

Caption:
A man wearing a light-colored shirt and dark pants is standing next to a black car while holding a fuel pump. The car is parked at a gas station, which is an open area with visible fuel pumps and a concrete ground. The background shows a wall with advertisements and an open sky above, indicating an outdoor setting.

Reasoning:
1. The man ([obj_1, 2, 10, 338, 1000]) is standing next to a black car while holding a fuel pump, indicating that he is refueling the vehicle.
2. The surrounding environment includes visible fuel pumps, a concrete ground, and a wall with advertisements.
3. The open sky above the scene and the lack of enclosing walls or a roof suggest that the location is outdoors.
4. Based on these observations, it is evident that the man ([obj_1]) is filling up his car in an outdoor setting at a gas station.

Conclusion:
The answer is: It is an outdoor setting; the man is at a gas station.

Figure 10: **MCoT-ViG record for scene-level reasoning.** The left panel presents the original image and its four-stage rationale. The right panel pairs a controlled shirt-color variant with the grounding-annotated version of the same rationale. The edit changes the man's shirt color while preserving the visual evidence supporting the outdoor gas-station answer.

Candidate Sample with Four-stage MCoT



Question: What device is to the right of the book?
Reference answer: charger
Grounding objects:
object_id: obj_1, object_label: charger,
bbox: [752,403,790,547]
Description: a small charger with a pink casing and a visible cable port.

Summary:
Identifying the book and examining the object positioned to its right will determine the type of device located there.

Caption:
A variety of items are arranged on a patterned surface, including a book titled 'Information Architecture' placed near the center. To the right of the book, there is a small charger with a pink casing and a visible cable port. Other objects such as pens, electronic devices, and water bottles are scattered around.

Reasoning:
1. The book is clearly visible near the center of the arrangement, titled 'Information Architecture'.
2. To the right of the book, there is a small charger with a pink casing and a visible cable port.
3. The position and proximity of the charger confirm that it is the device located to the right of the book.
4. Therefore, the device to the right of the book is the charger.

Conclusion:
The answer is: charger.

Sample with Visual Grounding Annotations



Grounding objects:
object_id: obj_1, object_label: charger,
bbox: [752,403,790,547]
Description: a small charger with a pink casing and a visible cable port.

Summary:
Identifying the book and examining the object positioned to its right will determine the type of device located there.

Caption:
A variety of items are arranged on a patterned surface, including a book titled 'Information Architecture' placed near the center. To the right of the book, there is a small charger with a pink casing and a visible cable port. Other objects such as pens, electronic devices, and water bottles are scattered around.

Reasoning:
1. The book is clearly visible near the center of the arrangement, titled 'Information Architecture'.
2. To the right of the book, there is a small charger ([obj_1, 752, 403, 790, 547]) with a pink casing and a visible cable port.
3. The position and proximity of the charger ([obj_1]) confirm that it is the device located to the right of the book.
4. Therefore, the device to the right of the book is the charger ([obj_1]).

Conclusion:
The answer is: charger.

Figure 11: **MCoT-ViG record for spatial-relation reasoning.** The left panel presents the original image and its four-stage rationale, while the right panel shows the corresponding grounding-annotated record. Red circles identify the small target charger. Red text marks its coordinate-bearing first occurrence and the identifier-only references used in subsequent reasoning steps.

Table 10: **Policy-training hyperparameters.** Settings above the separator are shared across model scales and optimizer families. The rollout batch denotes the total number of sampled completions contributing to one optimizer step.

Hyperparameter	$\mathbf{PaIPO_G}$	$\mathbf{PaIPO_D}$
Policy backbones	Qwen3-VL-Instruct 2B, 4B, and 8B	
Training records	10,000	
Training epochs	1	
Optimizer steps	2,500	
Hardware	4 × NVIDIA A100-SXM4 80 GB	
Per-device batch size	1 prompt and 4 completions	
Gradient accumulation	1	
Completions per prompt G	4	
Prompts per optimizer step	4	
Global rollout batch	16 completions	
Maximum completion length	384 tokens	
Maximum prompt length	No fixed truncation	
Learning rate	1×10^{-5}	
Optimizer	AdamW	
Learning-rate schedule	Cosine	
Warm-up ratio	0.03	
Weight decay	0	
Maximum gradient norm	1.0	
Numerical precision	bfloat16	
Sampling temperature	0.7	
Top-p / top-k	0.8 / 20	
LoRA rank / scaling / dropout	32 / 64 / 0	
Random seed	42	
Checkpoint interval / retention	250 steps / 3 checkpoints	
Policy-loss reduction	Completion-level	Token-level
KL coefficient β	0.04	0
Lower clipping coefficient	0.20	0.20
Upper clipping coefficient	0.20	0.28
Importance-ratio interval	$[0.80, 1.20]$	$[0.80, 1.28]$
Reward normalization	Group-relative	Group-relative
Dynamic sampling	Disabled	Enabled
Dynamic-sampling statistic	–	Group reward standard deviation
Dynamic-sampling threshold	–	0.03
Maximum resampling attempts	–	2
Overlong-penalty onset	–	320 tokens
Overlong buffer	–	64 tokens
Overlong-penalty weight	–	1.0

Table 11: **Prompt template and output schema for four-stage MCoT generation with object-level grounding annotations.** The template specifies the model inputs, stage-specific generation requirements, grounding-annotation format, and required JSON response.

Prompt Component	Template Content
System Prompt	You are a vision annotation expert responsible for constructing high-quality MCoT-ViG training data for multimodal reasoning and object grounding.
Input Specification	You receive: 1. an original image; 2. a grounding-guide image showing one or more question-relevant bounding boxes; 3. a question and its verified answer; 4. the ground-truth object identifier, local reasoning-object identifier, and exact bounding box in integer `xyxy_1000` format for every target object; and 5. optional source reasoning evidence.
Generation Objective	Identify the primary visible object represented by each supplied bounding box. Construct one grounding-object record for every target and generate a single four-stage MCoT rationale comprising *Summary*, *Caption*, *Reasoning*, and *Conclusion*. Preserve the one-to-one correspondence between each object record, reasoning-object identifier, and bounding box. The rationale must be grounded in the visible image content and consistent with the verified answer. Return exactly one valid JSON object and no additional text.
Stage Requirements	***Summary*.** Write a concise, question-conditioned reasoning plan rather than a general image summary. The *Summary* must contain only the reasoning plan and must not report observations, intermediate results, or the answer. ***Caption*.** Write a detailed visual description focusing on the objects, attributes, and relationships relevant to the question. The *Caption* must describe every grounded target using its exact object label and only directly visible evidence. It must not contain reasoning, interpretation, verification, answer prediction, object identifiers, or bounding-box coordinates. ***Reasoning*.** Write fluent natural-language reasoning that connects the visual evidence to the verified answer. Return the reasoning as a numbered sequence in one JSON string, with lines separated by `"\n"`; never return it as a JSON array. Each step must begin with an Arabic numeral followed by a period. The reasoning must satisfy the following requirements: 1. **Ground all relevant objects.** Every supplied target object must appear at least once. At the first occurrence of each target, use its exact object label and immediately append the corresponding identifier and bounding-box coordinates: `object_label ([obj_k, x1, y1, x2, y2])`. Copy the identifier and all four integer coordinates exactly. 2. **Examine the question-relevant evidence.** Use only attributes, states, actions, counts, positions, or relationships already described in the *Caption*. Subsequent references to a target should use a natural phrase containing its exact object label followed by the corresponding identifier, such as `the bird ([obj_k])`. 3. **Perform the required inference.** Compare, verify, count, identify, or relate the visual evidence according to the question and derive the answer. Additional intermediate steps may be included when multiple objects, comparisons, or relations must be considered. ***Conclusion*.** Write a concise natural-language conclusion that directly answers the question using the fixed format `"The answer is: ..."`.

Output Schema	Return a JSON object with exactly the following structure. The `grounding_objects` array must contain one record for every supplied target, ordered consistently with the input annotations.
	`{`
	`  "grounding_objects": [`
	`  {`
	`  "gt_object_id": "copied GT ID",`
	`  "reasoning_object_id": "copied local object ID",`
	`  "object_label": "generated object label",`
	`  "description": "generated object description",`
	`  "aliases": ["alias 1", "alias 2"],`
	`  "bbox": [x1, y1, x2, y2]`
	`  }`
	`  ],`
	`  "mcot": {`
	`  "summary": "question-conditioned reasoning plan",`
	`  "caption": "observation-only visual description",`
	`  "reasoning": "1. ...\n2. ...\n3. ...",`
	`  "conclusion": "The answer is: ..."`
	`  }`
	`}`

Table 12: **Prompt template and output schema for answer-preserving controlled attribute transformation.** Given a question, its verified reference answer, and a selected grounding-object record, GPT-4o modifies exactly one eligible visual attribute while preserving the verified answer and all unselected content.

Prompt Component	Template Content
System Prompt	You are a controlled visual-attribute transformation assistant. Given a question, its verified reference answer, and a selected grounding-object record, replace exactly one explicitly stated visual attribute with a perceptually distinct yet plausible alternative. The transformation must preserve the validity of the verified answer. Do not modify any attribute that is directly queried or required to derive the answer. Preserve the object category, identity, spatial relations, and all unselected content. Return exactly one valid JSON object with no additional text.
Generation Objective	Identify all explicit visual attributes in the selected target description. Determine which attributes can be changed without altering the verified answer to the supplied question. Exclude any attribute that is directly queried or provides evidence required to support the verified answer. From the remaining candidates, select exactly one eligible attribute and replace its value with a clearly distinguishable alternative. Revise the description minimally so that it reflects only the selected change. Eligible attributes include color, material, texture, pattern, and non-defining size or shape. Do not modify the object category, count, location, spatial relations, pose, action, or functional state. If no explicit attribute admits an answer-preserving transformation, mark the description as non-transformable rather than introducing an unsupported or answer-altering change.

Input Formatting	The input contains the question, its verified reference answer, and one selected grounding-object record:
	`{`
	`  "question": "<copied question>",`
	`  "verified_answer": "<copied verified reference answer>",`
	`  "grounding_objects": [`
	`  {`
	`  "gt_object_id": "copied GT ID",`
	`  "reasoning_object_id": "copied local object ID",`
	`  "object_label": "generated object label",`
	`  "description": "generated object description",`
	`  "aliases": ["alias 1", "alias 2"],`
	`  "bbox": [x1, y1, x2, y2]`
	`  }`
	`  ]`
	`}`

Response Formatting	Return exactly one JSON object in one of the following formats.
	Transformable description:
	`{`
	`  "is_transformable": true,`
	`  "source_description": "<copied input description>",`
	`  "attribute_transformation": {`
	`  "attribute_type": "<attribute type>",`
	`  "original_value": "<original value>",`
	`  "transformed_value": "<new value>"`
	`  },`
	`  "transformed_description": "<minimally revised description>"`
	`}`
	Non-transformable description:
	`{`
	`  "is_transformable": false,`
	`  "source_description": "<copied input description>",`
	`  "reason": "<explanation of why no explicit attribute>"`
	`}`

Table 13: **Prompt template for controlled resampled-image generation with Qwen-Image-Edit.** The prompt localizes the selected attribute transformation to the designated target region while constraining all unselected visual content to remain unchanged.

Prompt Component	Template Content
System Prompt	You are an image-editing assistant for controlled visual-attribute modification. Given an original image and structured editing specifications, modify every supplied target according to its corresponding attribute transformation. Apply exactly one specified attribute substitution per target within its designated bounding box. Preserve the scene layout, object identities, spatial relations, and all unselected visual content. Return only the edited image without textual explanations or rendered annotations.
Generation Objective	Generate one edited image containing all attribute substitutions specified in the supplied target records. For each target, use its object label and bounding box to identify the corresponding visual region. Use its source description, transformed description, and structured transformation record to determine the intended attribute substitution. The transformation changes are already specified; do not select additional targets, substitute different attribute values, or introduce additional edits. Preserve the original background, object categories, object counts, object locations, poses, spatial relations, and overall visual style. Do not introduce new objects, remove existing objects, or modify unspecified attributes. Leave all non-target regions and all unselected target properties unchanged.

Input Formatting	The input contains the original image and a structured editing specification. The `grounding_objects` contains one entry for every supplied target. Its length determines the number of targets to edit. Each entry combines the target's grounding annotation with the corresponding attribute-transformation output. Bounding boxes use integer `xyxy_1000` coordinates.
	`Original Image: {image}`
	`Editing Specification:`
	The following template illustrates one target entry. Repeat this entry structure for all supplied targets.
	`{ "grounding_objects": [{ "gt_object_id": "<copied GT ID>", "reasoning_object_id": "<copied local object ID>", "object_label": "<copied object label>", "bbox": [x1, y1, x2, y2], "source_description": "<copied source description>", "transformed_description": "<transformed description>", "attribute_transformation": { "attribute_type": "<copied attribute type>", "original_value": "<copied original value>", "transformed_value": "<copied transformed value>" } }] }`

Editing Instruction	Apply all transformations specified in the `grounding_objects` array to the original image. For each entry, locate the target using its object label and bounding box. Replace the specified `original_value` with the corresponding `transformed_value` for the stated `attribute_type`. The edited target should reflect its `transformed_description` while retaining all properties not selected for transformation. Keep each transformation associated with its own target; do not transfer attribute changes between objects. Use bounding boxes and object identifiers only to localize and distinguish the targets. Do not render bounding boxes, coordinates, object identifiers, labels, or other annotations in the output. Preserve all non-target content and return one image containing the requested target-specific edits.
Response Formatting	Return only the edited image without any textual explanation. The construction pipeline stores the resulting controlled visual variant as {`resampled_image`}.

Table 14: **Prompt template and output schema for multimodal visual-evidence verification.** GPT-5.1 evaluates attribute–relation coverage, visual faithfulness, and reasoning supportiveness using independent five-point anchored scores.

Prompt Component	Template Content
System Prompt	You are a multimodal visual-evidence verification assistant. Given a question, an original image, annotated target objects, and a four-stage MCoT rationale, evaluate whether the rationale contains complete, faithful, and reasoning-relevant visual evidence. Base every assessment exclusively on the supplied image, target annotations, and rationale. Return exactly one valid JSON object with no additional text.
Evaluation Objective	Evaluate the generated rationale separately along three dimensions: attribute–relation coverage, visual faithfulness, and reasoning supportiveness. Assign each dimension an integer score from 1 to 5 using the anchored rubric below. Do not allow stronger performance on one dimension to influence the score assigned to another. Do not infer evidence quality from the final answer alone.

Input Formatting	The input is formatted as:
	`Question: {question}`
	`Original Image: {image}`
	`Target Annotations: {obj_id, obj_labels, bounding_boxes}`
	`Summary: {summary}`
	`Caption: {caption}`
	`Reasoning: {reasoning}`
	`Conclusion: {conclusion}`

Evaluation Criteria	**Attribute–relation coverage.** Assess whether the *Caption* reports the question-relevant objects, attributes, quantities, object relations, and spatial relations required by the subsequent reasoning. **Visual faithfulness.** Assess whether the visual claims made in the *Caption* and *Reasoning* are supported by the original image and target annotations. Penalize hallucinated objects, incorrect attributes, and unsupported relations. **Reasoning supportiveness.** Assess whether the visual evidence reported in the *Caption* and referenced in the *Reasoning* is sufficient to support the intermediate inference. Use the following shared scoring anchors: **1:** The criterion is not satisfied; the required evidence is absent or materially incorrect. **2:** The criterion is weakly satisfied, with major omissions or errors. **3:** The criterion is partially satisfied; relevant evidence is present, but noticeable omissions or unsupported claims remain. **4:** The criterion is strongly satisfied, with only minor omissions and no material errors. **5:** The criterion is fully satisfied, with complete, supported, and directly relevant evidence.

Response Formatting	Return exactly one valid JSON object with the following structure. Each `score` must be an integer from 1 to 5, and each `reason` must provide a concise evidence-based justification.
	`{`
	`"attribute_relation_coverage": {`
	`"score": SCORE,`
	`"reason": "<brief evidence-based justification>"`
	`},`
	`"visual_faithfulness": {`
	`"score": SCORE,`
	`"reason": "<brief evidence-based justification>"`
	`},`
	`"reasoning_supportiveness": {`
	`"score": SCORE,`
	`"reason": "<brief evidence-based justification>"`
	`}`
	`}`

Scoring Constraints	**1.** Use the original image as the primary source for verifying visual claims and the target annotations for object identification and localization. **2.** Score the three dimensions independently; a high score on one dimension must not influence the score assigned to another. **3.** Penalize missing key evidence, hallucinated objects, incorrect attributes, and unsupported visual relations. **4.** Do not assign a high score solely because the final answer is correct. **5.** Do not reward linguistic fluency or verbosity unless the stated content is supported by visual evidence.