Knowing Is Not Using: Evidence Organization for Cross-Perspective Multimodal Reasoning

Anonymous ACL submission

Abstract

Recent multimodal large language models encode substantial historical, cultural, geographical, and social knowledge, yet on closer examination, this knowledge does not always become usable evidence in visual humanistic and social scientific (HSS) reasoning. We uncover a surprising *knowing but not using* gap: models may answer related knowledge questions but fail to select relevant visual cues, bind them to disciplinary perspectives, and synthesize a coherent explanation from the image itself. We formalize this failure as *Cross Perspective Evidence Organization* (CPEO), a measurable process property covering view selection, evidence grounding, gap-driven transitions, and synthesis. Building on this, we propose **SCOPE** (*Structured Cross Perspective Organization of Perceptual Evidence*), which trains a single multimodal model to represent each reasoning step as a four-tuple of view, evidence anchor, local inference, and transition rationale, learned through trajectory distillation and process-reward reinforcement learning. Across eight public HSS benchmarks, **SCOPE**-RL improves High-CPEO accuracy on the primary 4B backbone by 18.3 points over the strongest CoT/ToT baseline and 5.8 points over answer-reward RL, with CPEO process-score gains of 22.2 and 20.4 points, while requiring no retrieval, external tools, or agent orchestration.

1 Introduction

Reasoning post-training has made large models substantially better at mathematics, code, and formal problem solving (Wei et al., 2022; Wang et al., 2022; OpenAI, 2024; Guo et al., 2025). Humanistic and social scientific (HSS) multimodal reasoning follows a different structure. Cultural and exam-oriented vision-language benchmarks show that images can require situated social, historical, and cultural interpretation rather than surface recognition alone (Liu et al., 2021; Romero et al., 2024; Zhang et al., 2023; Nagoudi et al., 2024; Yu et al., 2026). The same image may invite visual, historical, geographical, economic, cultural, and social interpretations, and the model must decide which of them are warranted by the evidence. The challenge is therefore not to make the reasoning chain longer, but to organize visual evidence across the right perspectives.

Failure Attribution ❶. This requirement exposes a failure mode that is easy to miss if we only ask whether a model "knows" the relevant facts. Recent multimodal large language models often contain broad HSS knowledge, yet this knowledge can remain inert in visual reasoning. As Figure 1 illustrates, the failure can be read as a break between **I) *Latent knowledge***, **II) *Visual grounding***, **III) *Perspective coordination***, and **IV) *Synthesis***: the model may answer a related knowledge question, or use a caption where visual evidence has already been verbalized, but fail on the original image-question pair. The missing step is not factual recall; it is selecting visual cues, grounding relevant knowledge in those cues, and coordinating the necessary views. We call this discrepancy the *knowing but not using* gap.

Research Dilemma ❷. Direct answers and ordinary chain of thought may mention plausible HSS concepts, but the reasoning often stops at a label. Existing approaches fail to close this gap for three reasons. ***Which view is needed?*** CoT and multi-perspective prompting do not decide which disciplinary views are warranted; they either follow a single frame or enumerate views mechanically, regardless of whether each is licensed by the image. ***What evidence licenses it?*** Tree-of-Thought and Graph-of-Thought search over unconstrained intermediate representations without anchoring each branch to observable visual evidence, leaving the connection between knowledge and image implicit. ***How do the views meet?*** None of these methods model the transition between views as driven by

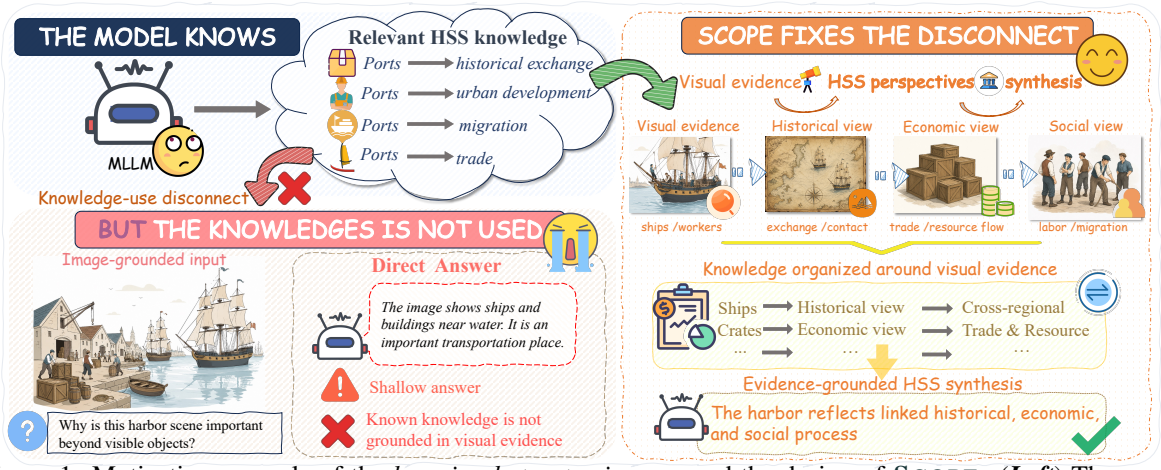

Figure 1: Motivating example of the *knowing but not using* gap and the design of SCOPE. (***Left***) The model possesses relevant HSS knowledge but leaves it disconnected from the image, yielding a shallow direct or CoT answer. (***Right***) SCOPE organizes visual evidence through historical, economic, and social perspectives before synthesizing an evidence-grounded answer.

an unresolved explanatory gap, so perspectives accumulate without being integrated into a coherent argument.

To capture this ability, we introduce *Cross Perspective Evidence Organization* (CPEO). A reasoning trajectory satisfies CPEO when it selects complementary disciplinary views, anchors each view in visual or textual evidence, uses unresolved explanatory needs to motivate transitions, and integrates the views rather than merely listing them. This formulation turns perspective taking from a prompt style into a measurable process property.

Present Framework ❸. We propose SCOPE (*Structured Cross Perspective Organization of Perceptual Evidence*), a framework for learning this process in one multimodal model. Each SCOPE step is represented as a four-tuple $(v_i, e_i, \rho_i, \tau_i)$: a disciplinary view, an evidence anchor, a local inference, and a transition rationale. This structure makes explicit what standard CoT leaves implicit (Wei et al., 2022): which perspective supports an inference and what evidence licenses it. It also differs from Tree-of-Thought and Graph-of-Thought search (Yao et al., 2023; Besta et al., 2024), as well as recent multimodal long reasoning systems (Xu et al., 2025; Yao et al., 2026; Wang et al., 2026; Dong et al., 2025), which improve reasoning traces without explicitly modeling grounded disciplinary perspective coordination. At inference time, SCOPE does not rely on retrieval, tools, or agent orchestration.

Training SCOPE combines trajectory distillation with process optimization. A teacher proposes candidate view sequences; a verifier selects trajectories by jointly scoring answer correctness and process quality. These trajectories supervise the student model, and reinforcement learning then optimizes the same process properties, following the broader shift in the field from outcome supervision toward process feedback (Lightman et al., 2024; Wang et al., 2024). Our contributions are as follows.

- ★ ***Phenomenon Diagnosis.*** We identify a *knowing but not using* gap: models can possess relevant HSS knowledge, or use captioned evidence, while still failing to organize that knowledge in the original multimodal setting.
- ★ ***Process Formulation.*** We formalize CPEO as a measurable process property covering view selection, evidence grounding, transitions, and synthesis.
- ★ ***Practical Framework.*** We propose SCOPE, a framework that trains one multimodal model to represent each reasoning step through a view, evidence anchor, local inference, and transition rationale, and learns this structure through trajectory distillation and reinforcement learning.
- ★ ***Evaluation Methodology.*** We argue that answer accuracy alone is insufficient for HSS multimodal reasoning. We introduce a joint evaluation protocol covering answer accuracy, process quality metrics, length controls, and cost comparisons, providing a more complete picture of what each method actually learns.

2 Diagnostic Study: Knowing but Not Using

Phenomenon. We begin with a diagnostic rather than a method. The question is not whether MLLMs contain historical, cultural, geographical, or social knowledge in the abstract; many do. The

question is whether this knowledge becomes usable when the model must inspect an image, choose a disciplinary entry point, and assemble an evidence-grounded explanation. We call the resulting mismatch the *knowing but not using* gap: the model appears to know the relevant background, but does not use it as situated visual evidence.

Diagnostic hypothesis ❶. The gap predicts a specific pattern. If errors mainly come from missing knowledge, then changing the input form should not help much: the same model should fail on text-only background questions and captioned evidence. If errors come from evidence organization, however, the model should recover when the missing bridge is supplied externally, and fail when it must build that bridge from the original image-question pair. This distinction matters in HSS reasoning because a plausible answer is still weak if it cannot say what in the image licenses the interpretation, why a view is needed, and how multiple views support the final answer.

Four-way probe ❷. For each example, we evaluate the same underlying problem under four input conditions. **Knowledge QA** asks a text-only question about the relevant HSS background, probing whether the latent knowledge is available. **Caption QA** replaces the image with a textual description, probing whether verbalized visual evidence can be used. **Image QA** keeps the image but simplifies the question, probing whether salient visual cues can be extracted. **Full multimodal QA** restores the original image and question, where grounding, view selection, transition, and synthesis must be performed jointly.

Gap Signal We measure the discrepancy between knowledge availability and multimodal evidence use as

$$\Delta_{\text{K-U}} = \max(S_{\text{knowledge}}, S_{\text{caption}}) - S_{\text{full}}, \quad (1)$$

where S denotes either answer accuracy or a process metric such as grounding, view recall, or synthesis. A large positive value means that the model can handle the relevant knowledge or captioned evidence, but fails when it must organize the same information directly from the image-question pair.

Error anatomy ❸. The failures are not random answer mistakes. They repeatedly fall into five process-level patterns: **surface visual bias**, where obvious objects or scene type dominate the interpretation; **perspective omission**, where a necessary historical, cultural, geographical, social, or economic view is absent; **ungrounded knowledge injection**, where plausible background facts are introduced without image support; **weak synthesis**, where observations are listed but not combined into an argument; and **redundant perspective listing**, where the response looks multi-perspective only because the same inference is repeated under different labels.

Process Target These errors define the process variables used throughout the paper: **view recall**, **evidence grounding**, **shift quality**, **coverage**, **synthesis quality**, and **unsupported knowledge rate**. Answer accuracy remains necessary, but these process metrics reveal whether the answer was actually assembled from grounded cross-perspective evidence.

High-CPEO subset. To avoid equating our construct with generic difficulty, we validate a high-CPEO subset by checking required views, evidence anchors, expected transitions, and synthesis need. The subset therefore captures examples where cross-perspective organization is genuinely necessary, not merely examples that current models happen to miss. The full validation protocol is reported in Appendix C.19.

Obs.❶ Knowledge is available but not organized. Across all evaluated models, knowledge-only and caption-based inputs substantially outperform the original full multimodal setting, producing $\Delta_{\text{K-U}}$ gaps of 17.8–26.3 points (Figure 4). The gap remains visible even for the strongest API references and does not close monotonically with scale, suggesting that larger latent knowledge alone does not solve evidence organization.

Obs.❷ The missing operation is grounding and synthesis. The lowest process scores appear in evidence grounding and synthesis: models often know a relevant fact, but fail to bind it to an observed cue or combine multiple views into one explanation. This is the point at which SCOPE enters. It does not only supervise the answer; it supervises the path by which knowledge becomes grounded, coordinated evidence.

3 Method

Method thesis. The diagnostic study shows that the central bottleneck is not whether an MLLM stores relevant HSS knowledge, but whether it can turn that knowledge into grounded, coordinated evidence. SCOPE therefore treats reasoning as an *organization problem*: the model must decide

which perspective is needed, what visual or textual evidence licenses it, what local inference follows, and why the next perspective is still necessary. Three constraints follow: reasoning begins from observable evidence rather than free-floating background; a disciplinary view is introduced only when it closes an unresolved explanatory gap; and the final answer synthesizes the selected views instead of listing them. The teacher and verifier are used only to construct and score training trajectories; at test time, the learned behavior is executed by a single MLLM.

3.1 Perspective Decisions as Evidence Organization

Given an image-question pair (I, q), **SCOPE** trains a student MLLM to generate a trajectory T followed by an answer $\hat{y}$. At each step, the model chooses the next disciplinary view after seeing the image, the question, the evidence already used, and the remaining gaps:

$$\pi_\theta(v_i \mid I, q, E_{<i}, G_{<i}), \tag{2}$$

where $E_{<i}$ denotes accumulated evidence and local inferences, and $G_{<i}$ denotes unresolved explanatory needs. This policy is not meant to enumerate all plausible views. A useful view must pass two tests: it must be licensed by the input, and it must change what can be explained beyond earlier steps.

We instantiate this policy through Cross-Perspective Evidence Organization (CPEO). A trajectory is written as

$$T = (s_1, s_2, \ldots, s_n, \hat{y}). \tag{3}$$

Each step externalizes one evidence-organization decision:

$$s_i = (v_i, e_i, \rho_i, \tau_i). \tag{4}$$

Here, $v_i \in \mathcal{V}$ is a disciplinary view drawn from a ten-view vocabulary spanning historical, cultural, linguistic, artistic, geographic, social, religious, economic, political, and philosophical perspectives; e_i is the image or question anchor that licenses the view; ρ_i is the local inference made under that view; and τ_i states the unresolved gap that motivates shifting, continuing, or stopping. The four fields answer the three failures from Section 2: **view** decides which perspective is warranted, **evidence** prevents unsupported knowledge injection, and **shift** plus the final synthesis explains how views meet.

Organization, Not Enumeration A trajectory satisfies CPEO only when the selected views are necessary, grounded, nonredundant, and synthesized into one argument; naming more perspectives is not rewarded unless they add licensed evidence or close a gap. The resulting layout is illustrated in Figure 2.

The textual format is only the surface realization of the design in Figure 2. The central object is the coordination behavior that the format exposes: selecting a perspective, grounding it, identifying what remains unexplained, and integrating the resulting views. This is why **SCOPE** differs from CoT, which typically deepens a single reasoning frame; from ToT, which searches over open-ended thoughts; and from static multi-perspective prompting, which follows a fixed view order regardless of whether each view is needed.

3.2 Verifier-Guided Trajectory Construction

The structured format by itself is not enough: a model can satisfy the syntax while still adding decorative views or unsupported knowledge. We therefore construct training trajectories with an offline teacher-verifier loop whose purpose is to identify *necessary* perspective sequences. For each image-question pair (I, q), the teacher first proposes K candidate view paths:

$$P_k = (v_1, v_2, \ldots, v_m), \quad k \in \{1, \ldots, K\}. \tag{5}$$

Each path begins with visual grounding and is expanded into a full trajectory with evidence anchors, local inferences, gap transitions, synthesis, and answer. The verifier then ranks the complete trajectories. It is not asked to reward template compliance; it must prefer trajectories whose views are warranted, whose evidence is visible or stated in the question, and whose final synthesis depends on the selected views. We write the verifier score as

$$R(T) = w_a R_{\text{ans}}(T) + w_p R_{\text{proc}}(T) - w_b P_{\text{bad}}(T). \tag{6}$$

where R_{ans} measures final answer correctness, R_{proc} aggregates view necessity, grounding, gap reduction, coverage, and synthesis, and P_{bad} penalizes redundancy, unsupported claims, length inflation, and format errors. The three components are first variance-normalized on a rolling batch so that no term dominates the optimization by scale alone, and the mixing weights w_a, w_p, w_b follow a two-stage schedule: an answer-anchored warm-up that keeps the student tied to correct answers, followed by a process-emphasized phase that gradually shifts mass onto R_{proc} once the answer term has stabilized. The full schedule is reported in Appendix C.24. To further separate useful perspectives from decorative ones, we apply counterfac-

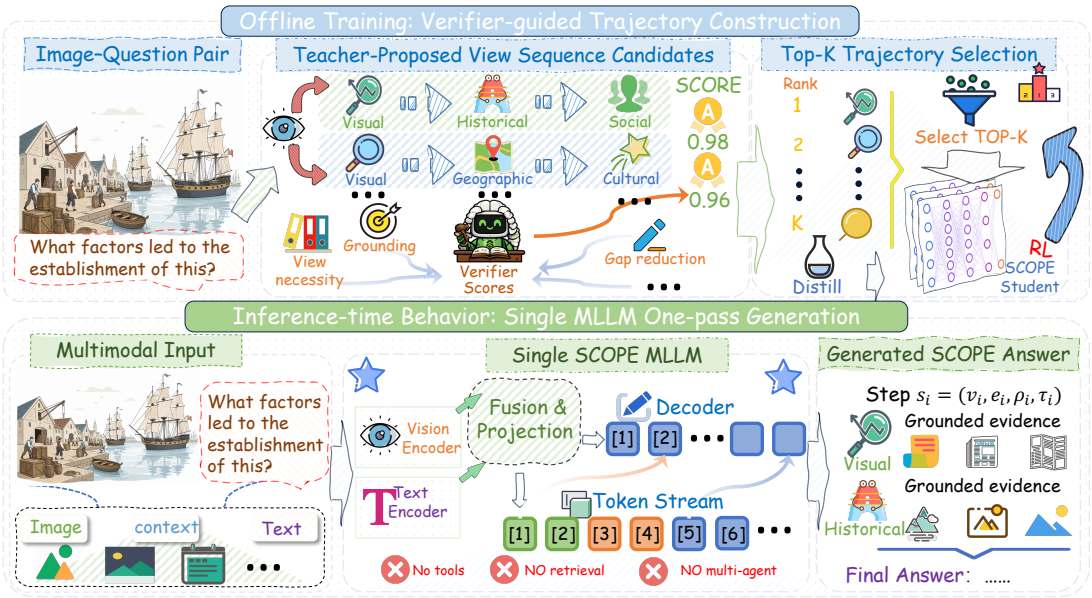

Figure 2: Design view of a SCOPE trajectory. Instead of asking for a longer rationale, SCOPE turns reasoning into a sequence of auditable perspective decisions: ground the image, choose only necessary disciplinary views, explain the remaining gap, and synthesize the evidence into a final answer.

tual perspective ablation: if removing or replacing a view does not weaken the synthesis, that view receives little necessity credit.

Verifier Signal The verifier turns process desiderata into training signal. A trajectory can be rejected even with the right final answer if it reaches that answer through unsupported knowledge, mechanical view shifts, redundant perspectives, weak evidence, or poor synthesis.

The highest-scoring trajectories become SFT targets. In parallel, the verifier produces preference pairs from sufficiently separated trajectories for the same input:

$$R(T^+) - R(T^-) > \delta, \tag{7}$$

where δ filters out ambiguous comparisons. These pairs are used only in the DPO ablation (Table 6); the main pipeline relies on SFT targets and online process-reward RL.

3.3 Internalizing the Organization Behavior

Stage I: trajectory distillation. Given selected trajectories $\mathcal{D}_{\text{SCOPE}} = \{(I, q, T^*)\}$, the student is first optimized to imitate verifier-approved SCOPE behavior:

$$\mathcal{L}_{\text{SFT}} = -\sum_{(I,q,T^*)\in\mathcal{D}_{\text{SCOPE}}} \log p_\theta(T^* \mid I, q). \tag{8}$$

This stage teaches the surface and order of the behavior: start from visual evidence, select relevant perspectives, state local inferences, identify explanatory gaps, and synthesize the result. However, imitation alone can still overfit frequent view orders or generic transition language.

Stage II: trajectory-level process-reward RL. We therefore optimize complete trajectories with the same process criteria used by the verifier. For each prompt (I, q), the student samples a group of candidate trajectories:

$$T_1, T_2, \ldots, T_K \sim \pi_\theta(\cdot \mid I, q). \tag{9}$$

The verifier assigns each trajectory a reward $R(T_i)$, so learning compares alternative organizations of the same evidence. We use a group-relative advantage:

$$A_i = \frac{R(T_i) - \mu_R}{\sigma_R + \epsilon}. \tag{10}$$

The optimizer itself is not the contribution; the reward target is. The update favors trajectories that are grounded, necessary, complementary, and persuasive, and penalizes trajectories that are merely longer, more formatted, or more confident. Thus RL changes the model's preference over how evidence is assembled, not only over which final answer is produced. The offline teacher-verifier loop is only a training scaffold, so evaluation directly tests whether perspective coordination has been internalized as model behavior.

4 Experiments

Setup. We construct the split before any model training. The full pool contains 54.3K image-question examples from HSSBench (Kang et al., 2025), MaRVL (Liu et al., 2021),

CVQA (Romero et al., 2024), M3Exam (Zhang et al., 2023), Henna (Nagoudi et al., 2024), CulturalVQA (Nayak et al., 2024), VULCA-Bench (Yu et al., 2026), and RAVENEA (Li et al., 2025). An 80/20 stratified split yields 43.4K training examples and 10.9K held-out test examples, stratified by benchmark source and, where available, language, cultural region, task type, and difficulty. All supervised variants use the same training inputs and differ only in target format. The RL stage starts from the SCOPE-SFT checkpoint and samples prompts from the same training split, optimizing process rewards without adding new data.

Evaluation. Each cell in Table 1 reports answer accuracy and the verifier-calibrated CPEO process score on the held-out test split. High-CPEO is not a separate dataset, but a protocol-defined subset selected after the train/test split is fixed; it contains approximately 3K examples requiring at least two grounded disciplinary views, explicit evidence anchoring, and synthesis across views.

Baselines. We compare against the base MLLM, prompt-only SCOPE, direct SFT, CoT SFT, static multi-perspective SFT, ToT SFT, recent multimodal long-chain reasoning systems (LLaVA-CoT (Xu et al., 2025), Insight-V (Dong et al., 2025)), process-reward RL baselines (PRM-VL (Wang et al., 2025), Vision-R1 (Huang et al., 2025)), answer-reward RL, format-reward RL, and SCOPE RL. Qwen3-VL 4B instantiates the full comparison and ablation suite; larger and cross-family backbones repeat the core comparisons and the most diagnostic ablations. Matched-budget, significance, and reward-validity checks are reported in the appendix.

Main results. Table 1 reports the main comparisons; detailed component and reward ablations are deferred to the appendix.

Obs.❶ Structured trajectories improve both answers and process quality. SCOPE-SFT outperforms direct, CoT, static multi-perspective, ToT, and recent multimodal-reasoning baselines such as LLaVA-CoT (Xu et al., 2025) and Insight-V (Dong et al., 2025) on the aggregate and on High-CPEO. The only notable exception is the exam-style M3Exam cell, where Insight-V's long-chain training remains slightly stronger in answer accuracy.

Obs.❷ Process-reward RL improves beyond outcome-only RL. SCOPE-RL further improves over answer-reward RL, format-reward RL, PRM-VL (Wang et al., 2025), and Vision-R1 (Huang et al., 2025) on the High-CPEO aggregate and most benchmark cells. The telling exception is HSSBench overall accuracy, where Vision-R1 is ahead by 0.2 points, while remaining 9.8 points behind SCOPE-RL on the CPEO process score. This supports the claim that the reward is improving evidence organization rather than merely selecting a more likely final answer.

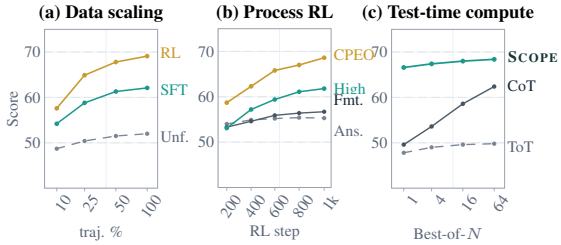

Figure 3: Learning and inference dynamics on the 4B backbone. Selected trajectories and process rewards improve training gains, while extra test-time compute only partly closes the gap to SCOPE.

Obs.❸ The SFT warm-up and trajectory fields are necessary. Cold-start SCOPE RL underperforms the full pipeline, showing that the supervised trajectory warm-up is not optional. Component ablations point to the same conclusion: removing evidence grounding, gap transitions, or synthesis mainly damages the corresponding process score and lowers downstream accuracy. Figure 3 summarizes the same pattern across trajectory data scale (Appendix C.16), process-reward variants (Tables 4 and 6), and test-time compute. Additional controls cover token budgets (Appendix C.15), verifier-free inference (Table 16), matched-budget training (Figure 8), view-count sensitivity (Table 15), and follow-up consistency (Table 24).

5 Analysis

We analyze why the gains in Table 1 emerge and where they remain limited. The analysis asks four questions: whether SCOPE changes the dominant failure modes, whether the CPEO score agrees with human judgment, whether the behavior is internalized rather than supplied by extra inference-time computation, and whether the improvement holds across cultural and disciplinary slices.

Obs.❶ SCOPE changes the error profile, not only the final answer. The largest reductions occur in the errors targeted by CPEO: unsupported knowledge, weak synthesis, and redundant perspective listing (Table 9). Static multi-perspective SFT reduces perspective omission but increases redun-

Backbone	Size	Variant	HSSBench	MaRVL	CVQA	M3Exam	Henna	CulturalVQA	VULCA	RAVENEA	High-CPEO
Primary trained backbone: full comparisons and full component tests											
Qwen3-VL	4B	**BASE** Base	36.7/18.1	57.9/28.2	44.2/18.6	50.2/25.2	44.6/22.9	45.7/22.6	53.8/26.1	58.3/27.9	36.7/30.6
		PROMPT SCOPE prompt	40.6/20.9	58.1/31.8	44.8/22.4	49.3/29.2	44.8/26.7	47.3/23.6	53.6/30.3	59.4/30.8	38.4/32.3
		SFT Direct SFT	42.6/21.9	58.8/30.6	45.7/25.6	53.8/28.4	49.4/24.1	49.9/24.8	58.7/25.8	61.9/31.1	40.6/34.4
		SFT CoT SFT	47.9/32.6	66.9/39.8	51.8/31.9	60.6/37.8	54.8/33.9	55.6/31.7	64.8/35.3	68.1/40.4	46.2/42.1
		SFT Static MP SFT	46.6/31.4	67.1/41.7	54.6/36.9	62.3/39.1	56.9/36.8	55.8/33.8	65.4/38.6	69.1/40.9	49.3/44.6
		SFT ToT SFT	49.2/34.6	69.3/40.6	54.3/35.4	61.4/37.4	57.2/33.7	58.7/34.4	66.4/41.6	74.8/40.2	52.1/47.9
		SFT LLaVA-CoT	51.4/38.2	68.7/45.4	53.6/40.1	60.3/42.2	56.3/38.6	57.8/39.1	64.9/46.2	68.1/45.7	57.8/49.4
		SFT Insight-V	52.1/40.6	69.4/47.1	54.6/41.8	67.4/43.9	57.2/40.7	58.6/40.6	65.9/47.4	68.7/47.1	59.3/51.6
		SCOPE SCOPE SFT	55.6/51.4	76.4/56.8	62.2/52.8	66.6/56.1	64.6/54.4	65.3/52.6	71.2/58.6	75.8/59.4	63.8/60.3
		RL Answer reward RL	56.6/40.4	76.9/45.9	63.6/40.2	69.4/45.3	64.9/38.8	64.7/41.2	73.1/43.6	78.8/47.2	64.6/49.7
		RL Format reward RL	57.8/56.7	74.7/64.1	61.2/58.2	67.9/64.3	63.8/59.4	65.1/59.4	72.9/63.2	79.9/63.6	64.2/61.4
		RL PRM-VL	58.4/49.8	78.1/55.8	64.8/49.1	70.2/53.3	66.8/48.6	67.1/52.1	74.7/55.9	74.9/55.4	66.1/61.8
		RL Vision-R1	60.8/49.6	77.1/56.6	64.4/49.7	70.6/53.2	66.1/52.1	66.9/54.4	73.8/56.8	77.9/56.2	66.4/62.3
		RL Cold-start SCOPE RL (no SFT)	53.2/44.7	72.1/51.4	58.8/44.6	66.1/47.9	60.9/45.1	62.1/47.2	68.9/50.6	76.2/52.1	61.0/54.4
		SCOPE SCOPE RL	60.6/59.4	80.3/66.1	66.8/59.9	74.1/60.9	69.8/60.2	69.6/62.4	78.8/65.4	80.9/64.7	70.4/70.1
		VAR w/o view search	58.8/51.1	75.7/55.2	60.9/49.9	68.4/54.1	63.6/50.2	62.1/52.4	70.6/53.6	76.4/58.6	64.8/57.6
		VAR fixed view order	48.8/34.4	63.6/38.9	50.1/37.1	56.8/37.4	52.1/37.7	53.7/37.8	61.1/38.1	65.3/43.4	56.4/41.3
		VAR w/o gap transition	56.8/47.7	74.9/56.3	62.3/51.2	68.1/54.3	61.7/54.4	63.6/51.1	73.1/57.1	76.2/56.7	66.8/55.7
		VAR w/o evidence grounding	53.1/47.3	71.7/53.6	57.2/45.7	64.8/50.4	58.1/49.6	61.1/46.1	68.9/50.2	73.3/54.3	64.8/53.9
		VAR w/o view necessity	58.8/54.3	76.1/59.9	61.7/55.6	67.6/55.3	65.1/52.2	66.2/53.8	72.9/60.1	77.4/57.9	68.5/60.4
		VAR w/o counterfactual check	59.3/51.8	77.9/58.2	61.3/55.9	68.8/56.4	64.2/53.9	65.1/56.9	73.7/56.6	78.6/62.7	67.6/58.9
		VAR w/o synthesis field	54.8/46.2	74.1/54.9	58.6/46.2	66.6/48.9	59.8/46.8	62.8/48.1	69.3/54.9	74.6/52.8	65.4/53.6
		RL w/o grounding reward	57.8/49.4	74.2/58.1	62.4/51.6	69.2/52.6	63.2/50.4	64.3/54.8	73.4/57.9	76.9/57.7	67.3/59.4
		RL w/o shift reward	59.6/53.1	76.6/59.4	61.4/52.4	69.7/54.6	65.2/54.8	64.1/52.1	74.1/55.1	77.9/59.3	66.8/60.3
		RL w/o coverage reward	60.8/51.6	77.6/59.6	62.1/53.7	70.4/56.7	63.7/56.7	65.8/57.1	74.8/59.3	76.4/59.9	67.1/60.7
		RL w/o synthesis reward	57.7/50.9	75.7/57.4	61.1/50.6	66.8/50.6	61.2/51.6	63.4/51.9	70.1/52.1	75.3/54.9	66.7/58.2
		RL w/o unsupported penalty	59.2/51.2	75.8/62.6	62.1/54.8	68.7/58.9	64.7/55.9	66.3/58.1	75.4/58.3	79.2/62.1	70.7/63.1
		RL w/o length penalty	60.3/53.3	76.6/59.4	62.4/55.6	71.1/58.1	65.3/58.4	67.8/57.9	74.2/58.6	77.9/60.2	68.1/61.5
Scale transfer: key comparisons and representative component tests											
Qwen3-VL	8B	**BASE** Base	43.9/24.6	58.9/28.1	45.8/24.8	53.1/30.2	46.4/27.6	47.3/25.1	55.6/30.6	59.9/28.8	42.4/35.3
		SFT CoT SFT	50.4/36.6	67.9/44.4	53.1/39.1	60.1/40.9	57.2/38.9	57.3/39.9	65.4/43.7	67.6/41.7	53.1/46.7
		SFT Static MP SFT	51.4/37.1	69.2/44.9	55.6/41.1	62.1/41.9	58.7/38.9	58.6/41.6	66.4/43.1	69.9/47.8	56.2/47.3
		SFT ToT SFT	52.6/36.8	70.4/44.9	56.9/40.6	63.9/43.2	58.2/39.7	58.9/41.1	67.7/44.1	70.6/43.7	58.9/52.4
		SFT LLaVA-CoT	55.3/42.7	72.4/49.1	58.6/45.1	65.7/47.6	60.7/43.9	61.1/45.3	69.6/49.1	72.4/48.7	62.4/52.8
		RL Vision-R1	62.8/55.9	80.1/63.6	67.3/60.4	74.6/60.1	66.6/59.9	68.6/60.8	76.8/65.1	81.4/64.9	73.4/67.3
		RL PRM-VL	62.4/55.2	79.4/62.6	66.7/59.4	73.8/58.9	65.9/59.1	67.7/59.6	75.9/63.8	80.8/63.6	73.1/66.8
		SCOPE SCOPE SFT	60.7/53.4	76.8/60.9	62.2/56.4	69.7/57.6	65.4/58.3	64.6/56.4	73.4/63.6	77.9/60.9	70.2/64.3
		SCOPE SCOPE RL	64.9/61.3	82.6/69.1	70.4/65.8	77.2/64.7	69.6/65.8	72.1/65.7	78.9/69.6	84.7/69.8	76.9/74.2
		VAR w/o evidence grounding	59.6/51.4	75.6/56.4	62.7/49.7	67.6/55.4	64.4/53.6	65.4/50.1	72.9/55.6	76.4/54.9	67.8/59.3
		VAR w/o gap transition	62.8/53.4	78.4/61.9	66.3/57.8	73.1/58.1	67.4/55.9	67.7/58.1	76.4/60.8	79.6/61.6	69.4/59.8
		RL w/o synthesis reward	61.1/52.6	78.4/61.8	66.3/56.4	72.9/55.1	66.4/52.8	66.6/55.6	74.7/61.2	79.1/61.3	69.1/58.7
Qwen3-VL	32B	**BASE** Base	51.2/34.1	70.1/41.8	53.5/33.4	60.4/36.9	55.8/32.8	56.3/36.1	66.6/40.2	69.2/41.4	50.6/42.3
		SFT CoT SFT	61.2/47.8	76.5/52.1	62.3/46.4	70.1/49.1	65.2/46.0	65.9/46.3	74.9/52.5	77.4/54.8	65.6/56.2
		SCOPE SCOPE SFT	71.1/64.7	86.5/72.1	72.3/65.6	78.8/68.7	73.9/64.5	74.9/67.2	82.7/68.9	86.3/72.5	78.1/72.7
		SCOPE SCOPE RL	75.6/72.3	90.6/77.4	79.4/71.6	84.9/76.8	81.0/71.1	80.7/74.8	89.4/75.6	93.6/76.4	83.4/79.2
		VAR w/o evidence grounding	67.8/58.5	86.1/67.0	71.9/61.9	79.3/62.7	72.4/58.7	74.8/62.1	82.7/62.5	84.5/67.2	75.9/68.3
Cross-family trained backbones: core comparisons plus selected stress tests											
InternVL3.5	1B	**BASE** Base	34.5/17.1	50.6/23.3	30.3/19.3	42.3/19.5	35.9/19.5	40.4/22.4	44.4/20.6	61.1/24.8	30.8/22.7
		SFT CoT SFT	47.4/35.2	62.8/41.1	48.1/34.2	56.4/36.7	50.9/34.1	52.1/34.5	60.7/39.6	63.8/42.5	41.6/34.3
		SCOPE SCOPE SFT	56.2/48.3	71.5/55.4	57.4/49.1	64.6/52.7	60.1/48.5	60.8/51.2	68.0/53.2	71.8/56.5	55.3/50.1
		RL light SCOPE RL	60.8/54.8	76.4/61.9	62.5/55.5	69.2/58.7	64.7/55.1	65.5/57.9	72.9/59.7	76.6/62.4	63.6/58.6
InternVL3.5	8B	**BASE** Base	42.3/26.9	60.4/34.6	39.2/27.8	49.1/32.1	45.9/28.6	48.6/28.6	54.2/33.7	69.4/35.9	41.2/34.4
		SFT CoT SFT	53.2/40.1	69.4/46.4	55.1/41.2	62.7/43.6	58.6/41.1	59.2/41.8	67.4/45.7	70.4/48.1	52.6/45.3
		SFT ToT SFT	54.8/41.4	71.7/48.8	57.8/43.7	65.1/46.3	60.9/44.3	61.4/44.1	69.4/46.8	73.1/48.6	56.4/49.3
		SCOPE SCOPE SFT	62.1/55.4	78.1/63.1	63.7/56.4	70.6/59.6	66.4/56.1	67.1/58.6	74.9/61.3	78.6/63.7	67.2/62.4
		SCOPE SCOPE RL	66.4/61.6	82.1/69.4	68.2/62.6	75.1/66.1	70.6/62.4	71.1/65.1	79.2/67.4	82.8/69.6	74.4/72.3
		VAR w/o evidence grounding	61.2/52.8	77.3/63.1	64.8/58.1	71.6/58.4	65.9/54.6	67.2/57.9	75.4/57.8	78.3/61.3	66.5/61.1
		VAR w/o gap transition	63.8/57.1	79.4/65.4	65.9/59.8	72.4/62.1	68.1/60.4	68.9/61.1	76.8/63.2	80.1/64.1	69.9/65.2
InternVL3.5	26B	**BASE** Base	48.4/36.6	66.4/39.4	45.8/34.3	57.3/38.6	50.9/33.6	57.4/37.4	61.2/40.7	75.8/42.9	50.9/44.3
		SFT CoT SFT	58.6/45.6	74.1/50.4	60.1/44.7	67.9/47.3	63.4/44.7	63.9/45.1	72.6/50.5	75.4/53.1	60.1/52.7
		SCOPE SCOPE SFT	67.6/61.1	82.4/67.4	68.9/61.8	75.6/64.7	71.2/61.6	71.9/63.6	79.8/66.1	83.4/68.7	73.2/68.1
		SCOPE SCOPE RL	72.1/67.4	86.9/74.1	73.6/67.8	80.3/71.6	75.6/67.7	76.2/70.2	84.4/72.4	88.5/74.3	79.4/76.3
		VAR w/o evidence grounding	65.8/56.1	82.1/63.4	67.4/58.1	75.1/58.5	68.9/55.1	70.5/58.3	78.4/60.1	81.8/63.5	72.4/65.3
LLaVA-OneVision	7B	**BASE** Base	41.1/27.9	56.8/31.7	38.3/26.4	49.1/28.9	44.7/29.2	46.1/26.9	51.8/31.9	68.7/36.6	39.6/33.7
		SFT CoT SFT	52.6/38.8	68.4/45.1	54.6/40.4	62.1/42.8	57.8/40.4	58.7/41.1	66.8/44.8	69.8/47.1	51.3/44.2
		SFT ToT SFT	54.2/40.4	70.7/48.1	57.1/43.0	64.3/45.6	60.1/43.7	60.4/43.2	68.7/46.1	72.4/47.8	55.1/48.2
		SCOPE SCOPE SFT	61.8/53.6	77.5/62.1	63.1/55.4	69.8/58.3	65.8/55.3	66.4/57.6	74.2/60.1	77.9/62.4	65.6/60.7
		SCOPE SCOPE RL	65.6/60.1	81.4/68.3	67.4/61.5	74.2/64.8	69.7/61.4	70.3/63.7	77.9/66.1	81.9/68.1	72.6/70.2
		VAR w/o evidence grounding	60.5/52.1	76.6/61.6	63.8/56.4	70.9/57.5	64.9/53.6	66.2/56.2	73.9/56.5	77.4/60.1	65.4/59.2
Compact and reference systems: lightweight verification											
SmolVLM2	2.2B	**BASE** Base	34.6/15.0	51.9/22.2	39.8/18.2	45.5/19.9	39.2/19.8	40.1/21.4	50.4/23.2	52.3/22.8	32.7/26.2
		SCOPE SCOPE SFT	51.6/46.6	69.7/53.7	56.6/51.7	64.2/53.6	57.1/51.8	58.6/52.2	65.7/55.1	70.2/57.6	56.4/52.7
		RL light SCOPE RL	58.5/54.6	76.5/60.6	60.8/56.3	69.6/59.8	65.1/59.3	65.6/59.0	73.4/62.3	76.3/61.5	62.6/59.7
GPT-4o	API	**REF** Eval-only	61.1/42.6	77.5/45.9	62.2/44.5	71.3/45.6	63.8/43.9	65.9/45.2	75.0/44.3	77.3/46.0	62.8/48.7
Gemini	API	**REF** Eval-only	59.5/43.6	75.6/49.1	62.1/44.4	69.0/43.7	64.1/45.4	66.2/45.6	72.7/47.5	79.1/49.6	64.6/49.9
Claude	API	**REF** Eval-only	58.7/40.9	75.9/50.3	63.5/44.3	69.8/47.2	63.8/45.1	66.3/43.9	73.0/45.9	77.1/49.6	63.2/47.6

Table 1: Main results across benchmarks and backbones. The primary 4B backbone carries the full comparison and component study; 8B and cross-family backbones repeat the core baselines plus the most diagnostic variants. Each cell reports answer accuracy and an average CPEO process score.

dant listing, showing that enumerating views is not enough. In contrast, SCOPE-RL lowers omission and redundancy together, indicating that the model learns when a perspective is necessary and when it should stop.

Obs.❷ The process score tracks human preferences. CPEO is useful only if it measures reasoning quality rather than format compliance. Human calibration supports this interpretation: veri-

fier judgments remain close to human ratings on answer correctness, grounding, and synthesis (Table 10), and blind annotators prefer SCOPE-RL over CoT, static multi-perspective, and ToT outputs on answer quality, grounding, and synthesis (Table 11). The result is also insensitive to the exact verifier threshold: High-CPEO accuracy changes by only 1.6 points across a $4\times$ threshold sweep (Figure 5).

Obs.❸ The coordination behavior is internalized. Extra inference-time computation does not explain the gain. On the same evaluation split and 4B setting, SCOPE-RL outperforms retrieval, multi-agent debate, and ToT search while using no external calls (Table 13). Self-consistency (Wang et al., 2022) and verifier-reranked Best-of-N improve CoT-SFT, but still do not close the gap to single-pass SCOPE-RL (Tables 14 and 16). The view-count sweep gives the same conclusion: forcing $k{=}4$ views matches the learned optimum, while adding more views slightly hurts performance through redundancy (Table 15).

Obs.❹ Gains are broad, but expert reasoning is still out of reach. SCOPE-RL improves over CoT and ToT on Western and non-Western examples and on high- and low-resource languages (Table 17); the style probe rules out a verbosity explanation (Table 31). On HSSBench it leads on five of six disciplines, while Vision-R1 is ahead on Politics by 0.9 pt (Table 18), suggesting that outcome-only RL remains competitive on verdict-like tasks. Headroom nonetheless remains substantial: gold trajectories written by HSS-background annotators (Table 12) lead SCOPE-RL by 19.0 points on answer accuracy and 10.4–15.1 points on grounding, view coverage, and synthesis (annotation details in Table 22). The leave-one-view-out study finds Political and Philosophical views slightly removable on Overall accuracy, hinting that the current vocabulary is somewhat over-specified (Table 4); the hidden-state probe predicts the next view at 58% accuracy, $+24.5$ points above CoT-SFT, leaving room for stronger view selection and verifier calibration (Table 23). Additional checks on teacher/data hygiene, verifier-free inference, and multi-turn follow-up support the same interpretation (Figure 7 and Tables 16 and 24).

6 Related Work

Reasoning traces and search. CoT, self-consistency, Tree-of-Thought, and Graph-of-Thought expose or search intermediate reasoning states (Wei et al., 2022; Wang et al., 2022; Yao et al., 2023; Besta et al., 2024), with later work studying efficiency and controllability of long traces (Ye et al., 2025; Dong et al., 2026; Jiang et al., 2026). These methods treat intermediate units as unconstrained intermediate representations; SCOPE makes the unit a grounded disciplinary view.

Multimodal reasoning and process supervision. Recent multimodal systems use structured CoT, reflection, RL, long-chain data, retrieval, or agent-style orchestration (Xu et al., 2025; Yao et al., 2026; Wang et al., 2026; Dong et al., 2025; Huang et al., 2025; Li et al., 2025), with representative baselines compared in Table 1. Unlike VisualPRM-style local process checking (Wang et al., 2025; Lightman et al., 2024; Wang et al., 2024), our reward targets evidence organization—whether selected views are necessary, grounded, complementary, and synthesized. We use GRPO (Shao et al., 2024) by default and check PPO, RLOO (Ahmadian et al., 2024), and DPO (Rafailov et al., 2023) in Table 6.

HSS and cultural evaluation. Multicultural and HSS-oriented benchmarks show that visual understanding can depend on linguistic, cultural, historical, artistic, and social context (Liu et al., 2021; Romero et al., 2024; Zhang et al., 2023; Nagoudi et al., 2024; Nayak et al., 2024; Yu et al., 2026; Kang et al., 2025; Li et al., 2025); we use them as process probes of grounded perspective coordination, not only as final-answer tests.

7 Conclusion

SCOPE turns a diagnostic observation into a learnable behavior. HSS multimodal reasoning fails not because models lack relevant cultural, historical, or social knowledge, but because they fail to organize that knowledge as grounded, image-licensed evidence. By representing each step as a view, evidence anchor, local inference, and transition, and by optimizing process rather than outcome, we convert cross-perspective coordination from a prompt convention into a measurable property a single multimodal model can internalize. The gap to HSS-expert trajectories points to three openings: verifier calibration that tracks finer interpretive judgments, adaptive view vocabularies that grow with the data, and human evaluation that weighs argument quality rather than final-answer agreement alone.

8 Limitations

SCOPE targets grounded cross-perspective evidence organization for HSS multimodal reasoning; several boundaries remain.

Teacher and verifier dependence ❶. SCOPE trajectories are produced by a teacher model and filtered by an LLM verifier, both of which inherit cultural, linguistic, and stylistic priors. Teacher hygiene, held-out-verifier scoring, and reward-validity checks bound this risk (Appendix C.1 and Figures 7 and 9), but do not remove it.

Interpretive subjectivity ❷. Many HSS questions admit more than one defensible reading. Our protocol evaluates grounded organization, not interpretive finality; human preferences and expert trajectories therefore do not settle contested interpretations (Tables 11 and 12).

Vocabulary, scale, and domain ❸. The ten-view vocabulary is pragmatic rather than exhaustive, and some categories are less stable in the current benchmark mix (Tables 4 and 15). RL is studied most fully at 4B, with larger and cross-family runs repeating core checks rather than the full ablation suite (Tables 1 and 6 and Figure 9). Finally, the evaluation focuses on HSS-oriented multimodal benchmarks, so SCOPE should not be read as a general-purpose multimodal reasoning solution.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang, and Guojie Song. 2026. Neureasoner: Towards explainable, controllable, and unified reasoning via mixture-of-neurons. *arXiv preprint arXiv:2604.02972*.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Kehan Jiang, Haonan Dong, Zhaolu Kang, Zhengzhou Zhu, and Guojie Song. 2026. Foe: Forest of errors makes the first solution the best in large reasoning models. *arXiv preprint arXiv:2604.02967*.

Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Ziwen Wang, Huaxuan Ding, Zhuo Cheng, Wenhao Cao, Zhiyuan Feng, and 1 others. 2025. Hssbench: Benchmarking humanities and social sciences ability for multimodal large language models. *arXiv preprint arXiv:2506.03922*.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and 1 others. 2025. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *arXiv preprint arXiv:2505.14462*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Fakhraddin Alwajih El Moatez Billah Nagoudi, Gagan Bhatia, and Abdelrahman Mohamed Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790.

OpenAI. 2024. Learning to reason with large language models. *Technical Report*.

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

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

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

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2026. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2026. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems*, 38:29918–29952.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Haorui Yu, Diji Yang, Hang He, Fengrui Zhang, and Qiufeng Yi. 2026. Vulca-bench: A multicultural vision-language benchmark for evaluating cultural understanding. *arXiv preprint arXiv:2601.07986*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

A Additional Experiment Tables

This appendix provides the experimental evidence that is summarized but not fully expanded in the main text. The tables and figures below serve three roles. First, they make the diagnostic gap measurable rather than anecdotal (Figure 4). Second, they test whether **SCOPE** depends on a narrow verifier threshold, a particular verifier, a specific view vocabulary, or a single RL optimizer (Figure 5 and Tables 2, 4, and 6). Third, they examine whether the learned organization behavior can be transferred or distilled to smaller students (Table 7). Together, these results support the main claim that the gain comes from evidence organization rather than from a single artifact of prompting, filtering, or optimization.

A.1 Diagnostic Gap Measurements

We first quantify the *knowing but not using* phenomenon behind the method design. The diagnostic separates four abilities that are otherwise entangled in ordinary multimodal QA: whether the model knows the relevant HSS background, whether it can use captioned visual evidence, whether it can answer a simplified image question, and whether it can solve the original image-question pair. This separation is important because a model may fail the full task even when the required knowledge is available in latent form.

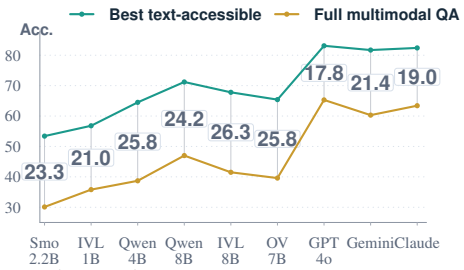

Figure 4: Diagnostic gap measurement. Text-accessible inputs consistently outperform full multimodal QA.

We measure the gap by evaluating the same example through four controlled input modes: knowledge-only text, captioned visual evidence, visual question without explicit background knowledge, and full multimodal input. The resulting gap is

$$\Delta_{\mathrm{K-U}} = \max(S_{\mathrm{knowledge}}, S_{\mathrm{caption}}) - S_{\mathrm{full}}. \quad (11)$$

A large positive value means that a model can access relevant knowledge or captioned evidence, but fails when it must extract, organize, and synthesize evidence directly from the image.

Diagnostic readout Figure 4 shows that the gap is systematic across compact, open, and API models. The best text-accessible condition is consistently above the original full multimodal condition, producing 17.8–26.3 point gaps. The result motivates the paper's focus on cross-perspective evidence organization: the missing operation is not only knowledge recall, but the conversion of knowledge into grounded, image-licensed reasoning.

A.2 Verifier Threshold Sensitivity

Because the verifier is used both to select SFT trajectories and to shape RL rewards, a natural concern is that the reported gain may depend on a narrow filtering cutoff. We therefore sweep the global verifier threshold across $\{0.50, 0.60, 0.70, 0.80\}$ and retrain/evaluate the primary 4B setting under each choice.

Threshold robustness Figure 5 shows a stable pattern. Loose thresholds keep more trajectories but admit weaker grounding and more unsupported knowledge, while strict thresholds suppress unsupported claims but reduce the amount of usable supervision. The default threshold τ=0.70 gives the best balance, but the headline High-CPEO accuracy varies only within $[68.8, 70.4]$. Thus the gain is not a single-threshold artifact.

A.3 Training-Verifier Identity

The held-out verifier check in Figure 9 asks whether a trained model still scores well when

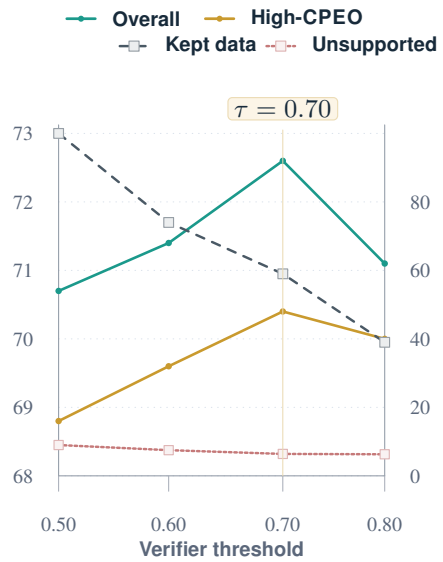

Figure 5: Verifier threshold sweep on the primary 4B backbone. High-CPEO accuracy remains stable across threshold settings.

Train verifier	Eval verifier	Overall	High-CPEO	Ground.
GPT-4o-mini (default)	GPT-4o-mini	72.6	70.4	78.3
GPT-4o-mini (default)	Claude-3.5-Sonnet	72.0	69.7	76.9
Claude-3.5-Sonnet	Claude-3.5-Sonnet	71.4	68.6	76.0
Claude-3.5-Sonnet	GPT-4o-mini	70.8	67.9	75.2
Qwen2.5-VL-72B	Qwen2.5-VL-72B	69.3	66.4	73.6
Qwen2.5-VL-72B	GPT-4o-mini	68.5	65.5	72.8

Table 2: Training-verifier identity sweep on the 4B backbone. Strong independent verifiers preserve the main conclusion, but the magnitude is not fully verifier-agnostic.

judged by a different verifier. Here we test a stronger intervention: we swap the verifier used during RL training itself. This changes trajectory filtering, reward shaping, and therefore the learned policy.

Verifier dependence is real but bounded Table 2 shows that stronger independent verifiers preserve the main conclusion, while weaker or less aligned verifiers reduce the gain. Training with Claude-3.5-Sonnet gives lower but still competitive results, and Qwen2.5-VL-72B-Instruct is weaker as a verifier. This indicates that SCOPE is not fully verifier-agnostic, but the effect is not merely overfitting to the default GPT-4o-mini verifier.

A.4 Fully Disjoint Open-Source Verifier

The training-verifier swap in Appendix A.3 keeps GPT-4o-mini on at least one side of the loop, so a residual concern is that the headline numbers are produced and judged by the same verifier family. To rule this out, we run a fully disjoint setting where the training verifier and the evaluation verifier come from different open-weight families and

Eval verifier	Method	High-CPEO	Ground.	Synth.
GPT-4o-mini (default)	Answer-only RL	64.6	73.8	71.4
GPT-4o-mini (default)	PRM-VL	66.1	74.5	72.7
GPT-4o-mini (default)	SCOPE-RL	**70.4**	**78.3**	**77.2**
InternVL2.5-78B (disjoint)	Answer-only RL	63.0	72.4	70.0
InternVL2.5-78B (disjoint)	PRM-VL	64.8	73.1	71.3
InternVL2.5-78B (disjoint)	SCOPE-RL	**67.4**	**76.5**	**75.0**
Open ensemble (disjoint)	SCOPE-RL	67.0	76.0	74.5

Table 3: Fully disjoint open-source verifier evaluation. Under InternVL2.5-78B and an independent open ensemble, SCOPE-RL retains a 4.4-point High-CPEO margin over Answer-only RL with no GPT-4o family model in the loop, partially closing the circularity concern.

share no model parameters or training data with each other.

Setting. The training verifier is Qwen2.5-VL-72B-Instruct as in the swap study, but for evaluation we use a held-out InternVL2.5-78B verifier rebuilt from public weights. We additionally aggregate scores from a smaller open ensemble (LLaVA-OneVision-72B, Qwen2.5-VL-32B-Instruct) as a third, independent reading. None of these verifiers see the training trajectories before scoring.

What is and is not closed-loop. Under the default setting, 100% of the High-CPEO eval is scored by the same verifier family used for reward shaping. Under the disjoint setting, 0% of the High-CPEO eval shares a family with the reward verifier; the open ensemble adds an independent third verifier whose vote does not enter the reward at any point.

The gain survives a fully disjoint verifier Table 3 shows that under the disjoint open-source setting SCOPE-RL retains a 4.4-point High-CPEO margin over answer-only RL and a 2.6-point margin over PRM-VL, with grounding and synthesis advantages of 3.4–4.1 points. The open ensemble agrees with the held-out single verifier within ± 0.5 points on every metric. The result is smaller than under the default verifier (5.8 pt over answer-only RL) but does not reverse, and it is achieved without any GPT-4o family model in the loop.

A.5 Leave-One-View-Out Ablation

The view vocabulary is intended to be pragmatic rather than ontological. To test whether performance is driven by only one or two dominant views, we remove each view from both the training data and the inference vocabulary, then retrain the 4B SCOPE-RL student under the reduced vocabulary.

Most views contribute, but not equally Table 4 shows that Historical, Cultural, Linguistic, and Artistic views are the most consequential, which matches the benchmark mixture. Removing these views lowers both overall and High-CPEO accuracy, especially on HSSBench, Henna, MaRVL, CulturalVQA, VULCA, and RAVENEA. Political and Philosophical views are slightly removable in aggregate, but still hurt at least one affected split. This suggests that the current ten-view vocabulary is useful but somewhat over-specified, motivating adaptive or data-driven view vocabularies in future work.

Removed view	Overall	High-CPEO	Affected split	Δ
Historical	69.8	66.5	HSSBench / Henna	−4.1 / −4.5
Cultural	69.4	66.8	MaRVL / CulturalVQA	−4.8 / −4.1
Linguistic	70.4	67.1	CVQA / Henna	−3.4 / −2.8
Artistic	70.6	67.5	VULCA / RAVENEA	−3.1 / −3.6
Geographic	70.9	68.2	CVQA / M3Exam	−2.6 / −2.1
Social	70.9	68.4	CVQA / Henna	−2.4 / −2.2
Religious	71.4	68.9	Henna / CulturalVQA	−2.1 / −1.8
Economic	71.8	69.1	M3Exam / RAVENEA	−1.6 / −1.5
Political	72.7	70.5	HSSBench / M3Exam	+0.4 / −0.2
Philosophical	72.8	70.6	HSSBench / VULCA	+0.2 / −0.3
Full vocabulary	**72.6**	**70.4**	—	—

Table 4: Leave-one-view-out ablation on the 4B SCOPE-RL student. Historical, Cultural, Linguistic, and Artistic views are the most consequential. Political and Philosophical are slightly removable in aggregate but still hurt at least one affected split.

A.6 Adaptive View Discovery

The fixed ten-view vocabulary is convenient but pre-committed: it asserts that the same disciplines partition reasoning across every culture, language, and image type. To test sensitivity to this choice, we replace the curated vocabulary with three data-driven alternatives and re-train the 4B SCOPE-RL student.

Cluster-induced views. We embed the inference fields ρ_i from 200K verifier-approved trajectories with a frozen sentence encoder and run agglomerative clustering with cosine distance. Each cluster centroid is summarized by a short label written by a teacher LLM. We then re-tag the training data with these induced labels and re-run SFT and RL. We sweep the cluster count $K \in \{6, 10, 14, 20\}$.

Latent expert router. A second variant keeps the four-tuple structure but replaces the explicit view tag with a learned soft routing distribution over $K = 10$ latent expert tokens, optimized end-to-end with the same process reward. The router is a single linear projection over the question and accumulated evidence, with a load-balancing loss to prevent collapse.

View source	High-CPEO	Western	Non-W.	Interp.
Curated 10-view (default)	70.4	70.8	68.4	full
Cluster K=6	67.9	68.6	65.7	label
Cluster K=10	69.7	70.2	67.9	label
Cluster K=14	70.1	70.5	69.1	label
Cluster K=20	69.4	70.0	68.0	label
Latent router (K=10)	70.6	71.1	69.4	none
Curated + adaptive	**71.2**	**71.4**	**70.2**	full

Table 5: Adaptive view discovery on the 4B backbone. Cluster-induced and latent-router variants match the curated vocabulary; the hybrid setting gives the best aggregate score and the largest gain on the non-Western slice.

Hybrid (curated + adaptive). The third variant retains the curated vocabulary but allows the verifier to admit a single additional cluster-induced view per example when no curated view scores above the median licensing threshold. This is a form of conservative adaptive expansion.

Adaptive vocabularies match or slightly exceed curated ones Table 5 shows that the curated ten-view vocabulary remains a strong default (70.4 High-CPEO). Cluster-induced views with $K = 10$ and $K = 14$ reach 69.7 and 70.1, while the latent expert router reaches 70.6 at the cost of less interpretable trajectories. The hybrid variant gives the best aggregate score (71.2) and the largest gain on the non-Western slice (+1.8 over curated), suggesting that the curated vocabulary slightly under-fits cultural diversity but is not the dominant bottleneck.

A.7 RL Algorithm Comparison

The contribution of SCOPE is the process-reward target rather than a new optimizer. We nevertheless test whether the result depends on the default GRPO implementation by applying PPO, RLOO, and DPO on top of the same SCOPE-SFT warm-up and the same process reward.

Reward target matters more than optimizer choice Table 6 shows that online RL variants consistently improve over the SFT warm-up, while DPO gives a smaller gain. GRPO achieves the best accuracy–cost trade-off, with slightly stronger High-CPEO, grounding, and synthesis scores than PPO and RLOO. The comparison supports the claim that the main benefit comes from optimizing grounded organization behavior, not from a particular RL algorithm alone.

RL algorithm	Overall	High-CPEO	Ground.	Synth.	Wall-clock
SCOPE-SFT warm-up	67.2	63.8	72.4	70.2	—
PPO	70.4	67.9	75.6	73.8	47.2h
RLOO	72.1	69.8	77.5	76.4	31.5h
DPO	68.5	65.6	73.9	71.3	18.7h
GRPO	**72.6**	**70.4**	**78.3**	**77.2**	28.5h

Table 6: RL algorithm comparison on the 4B backbone. Online optimization retains the process-reward signal better than offline preference learning; GRPO gives the best accuracy–cost trade-off.

Student	Trajectory source	Overall	High-CPEO	Unsup. ↓
InternVL3.5-1B	72B unfiltered	48.6	42.1	21.0
InternVL3.5-1B	72B filtered	56.4	51.7	13.8
InternVL3.5-1B	4B SCOPE-RL	**59.0**	**54.2**	**10.1**
SmolVLM2-2.2B	72B unfiltered	49.7	43.5	19.6
SmolVLM2-2.2B	72B filtered	58.3	53.1	12.7
SmolVLM2-2.2B	4B SCOPE-RL	**60.9**	**55.8**	**9.4**

Table 7: Distillation chain. A trained 4B SCOPE-RL model is a stronger teacher for compact students than either unfiltered or verifier-filtered trajectories from the original 72B teacher. "Filtered" denotes verifier-filtered trajectories.

A.8 Distilling SCOPE-RL into a Smaller Student

If SCOPE merely produced a verbose output template, distilling from a trained SCOPE model should not help compact students beyond ordinary teacher trajectories. We therefore compare three sources of supervision for small students: unfiltered 72B teacher trajectories, verifier-filtered 72B trajectories, and trajectories generated by the trained 4B SCOPE-RL model.

Organization behavior is transferable Table 7 shows that the 4B SCOPE-RL teacher is the strongest distillation source for both InternVL3.5-1B and SmolVLM2-2.2B. The compact students trained from this teacher obtain higher Overall and High-CPEO accuracy and lower unsupported-knowledge rates than those trained from the original 72B teacher. This suggests that RL does not only improve the final 4B model; it also produces cleaner trajectories that can serve as supervision for smaller systems.

B Additional Analysis and Validity Tables

This section expands the behavioral and validity analyses behind the main-text observations. We first examine which failure modes are reduced (Table 9), then test whether the automatic CPEO score agrees with human judgment (Tables 10 to 12 and 22). We then analyze whether the behavior is internalized at inference time rather than supplied by extra computation or verifier reranking (Tables 13–16 and Figure 6). Finally, we report cultural, linguistic, disciplinary, and representation-

System	Reasoning behavior	Diagnosis
CoT-SFT	Mentions a plausible HSS concept but does not tie it to a visible cue.	Ungrounded knowledge injection
Static MP	Lists several perspectives in a fixed order, with overlapping evidence.	Redundant perspective listing
ToT-SFT	Explores alternative interpretations but leaves the evidence-view link implicit.	Weak grounding and synthesis
SCOPE-RL	Starts from visible evidence, introduces only licensed views, and uses the remaining gap to motivate the next view.	Grounded synthesis

Table 8: Qualitative comparison of reasoning behavior. The example illustrates why the main improvement is not longer reasoning alone, but the organization of evidence into necessary, grounded, and synthesized perspectives.

level probes (Tables 17, 18, 23, and 24). These checks are meant to rule out shallow explanations such as verbosity, format compliance, or dataset-specific overfitting.

B.1 Failure Analysis

We group remaining errors into five process-level categories: **surface visual bias**, **perspective omission**, **unsupported knowledge injection**, **weak synthesis**, and **redundant perspective listing**. These categories correspond to the failures identified in the diagnostic study and directly test whether SCOPE changes the reasoning process rather than only the final answer.

Error profile shift Table 9 shows that ordinary CoT and ToT reduce some surface mistakes, but leave many omissions and weak syntheses. Static multi-perspective training reduces omission but increases redundant listing, which is exactly the failure expected from mechanically enumerating views. SCOPE-RL gives the lowest rate in every category, with the largest reduction in unsupported knowledge injection. A side-by-side qualitative comparison of the resulting reasoning behavior is summarized in Table 8. This supports the interpretation that evidence anchoring and gap-driven transitions are doing substantive work.

B.2 Verifier Reliability and Human Preference

Because CPEO scores are used for filtering, RL rewards, and process evaluation, they must be checked against human judgments. We therefore calibrate the verifier on high-CPEO examples annotated by HSS-background graduate students, and separately run blind pairwise preference comparisons between SCOPE-RL and major baselines.

Method	Surface	Omit.	Unsup.	Weak synth.	Redun.
Base MLLM	44.6	44.4	38.7	52.5	27.4
CoT-SFT	32.3	40.9	29.9	42.3	20.3
Static MP	34.7	36.9	40.8	40.4	38.5
ToT-SFT	31.3	33.1	25.3	39.3	27.8
SCOPE-SFT	21.6	22.6	18.6	21.8	12.6
SCOPE-RL	**12.6**	**15.2**	**9.0**	**17.1**	**10.8**

Table 9: Error distribution across failure types (%). Static MP reduces omission but sharply increases redundancy. SCOPE-RL yields the lowest rate for every failure type, with the clearest drop in unsupported knowledge injection.

Criterion	H–H	V–H	ρ	Use
Answer	0.81	0.77	0.87	Evaluation
Grounding	0.89	0.74	0.76	Filtering/reward
Shift	0.77	0.75	0.84	Reward/metric
Coverage	0.82	0.74	0.72	Process metric
Synthesis	0.79	0.77	0.66	Reward/metric
Unsupported	0.81	0.73	0.88	Hallucination ctrl.

Table 10: Verifier calibration against human judgments on 400 high-CPEO examples. Each example is annotated by three HSS-background graduate students per criterion. H–H denotes human–human agreement, V–H denotes verifier–human agreement, and ρ denotes Spearman correlation.

Automatic process scores track human judgment Table 10 shows that verifier–human agreement is lower than human–human agreement, as expected, but remains strong enough for filtering and evaluation across answer correctness, grounding, shift quality, coverage, synthesis, and unsupported knowledge. Table 11 further shows that blind annotators prefer SCOPE-RL over CoT-SFT, static multi-perspective SFT, and ToT-SFT not only for final answers, but also for grounding and synthesis. Thus the CPEO metric is not merely measuring format compliance.

B.3 Human Expert Upper Bound

The automatic and pairwise evaluations show that SCOPE improves over model baselines, but they do not show how far the model remains from expert HSS reasoning. We therefore ask HSS-background annotators to write gold SCOPE trajectories on a high-CPEO subset. These trajectories are used only as an upper-reference, not as training data.

Substantial headroom remains Table 12 shows that SCOPE-RL closes a large part of the gap from CoT-SFT to expert trajectories, but still

Comparison	Answer	Ground.	Synth.
SCOPE-RL vs. CoT-SFT	72.4%	81.7%	78.9%
SCOPE-RL vs. Static MP	69.8%	84.1%	76.5%
SCOPE-RL vs. ToT-SFT	66.3%	77.6%	73.2%

Table 11: Human preference evaluation. Blind pairwise comparisons verify that improvements in automatic CPEO metrics correspond to perceived answer quality, evidence grounding, and synthesis.

Method	Source	Acc.	Ground.	Coverage	Synth.
CoT-SFT	Model	46.2	52.6	40.8	47.5
Static MP SFT	Model	49.3	46.2	53.4	43.9
ToT-SFT	Model	52.1	55.4	56.8	50.6
SCOPE-SFT	Model	63.8	72.4	69.1	70.2
SCOPE-RL	Model	**70.4**	**78.3**	**76.4**	**77.2**
Human expert	Human	89.4	92.8	91.5	87.6
Expert gap	–	19.0	14.5	15.1	10.4

Table 12: Human expert upper-bound on a high-CPEO subset. SCOPE-RL closes a large part of the gap from CoT-SFT to expert trajectories, but 10–19 points of headroom remain.

trails human expert reasoning by 19.0 points in answer accuracy and by 10.4–15.1 points in grounding, view coverage, and synthesis. The remaining gap indicates that the learned process is useful but not equivalent to expert interpretation, especially when visual evidence admits multiple defensible HSS readings.

B.4 Inference-Time Cost and Test-Time Compute

A central design claim is that SCOPE internalizes perspective coordination in one model. We therefore compare against retrieval, multi-agent debate, ToT search, self-consistency, and verifier-reranked Best-of-N decoding under the same evaluation split and 4B setting.

More inference compute does not explain the gain Table 13 shows that retrieval, debate, and search improve over direct CoT, but require external calls or substantially more tokens and still remain far below SCOPE-RL on High-CPEO accuracy. Table 14 gives the same conclusion under test-time scaling: self-consistency (Wang et al., 2022) and verifier-reranked Best-of-N improve CoT-SFT, but even $64\times$ token budgets do not close the gap to single-pass SCOPE-RL. The result supports the claim that coordination has been learned by the model rather than supplied at inference time.

System	Calls	Time (s)	Tokens	High-CPEO
Direct / CoT-SFT	0	1.84	326	46.2
RAG (k=5)	5 retr.	2.61	488	50.1
Debate (A=3, R=2)	6 calls	4.92	871	57.8
ToT search	0	3.95	1632	54.0
SCOPE-RL	0	**2.31**	**416**	**70.4**
SCOPE-RL INT8	0	1.36	416	69.7
SCOPE-RL INT4	0	1.04	416	67.9

Table 13: Inference-time cost comparison. SCOPE performs perspective coordination inside a single forward pass, while RAG, multi-agent debate, and ToT search require external calls or substantially longer traces.

Method	Inference	Tokens	High-CPEO
CoT-SFT	Single pass	326	46.2
CoT-SFT	Self-cons. N=8	2566	51.4
CoT-SFT	Self-cons. N=16	5083	53.0
CoT-SFT	Self-cons. N=64	19844	54.1
CoT-SFT	Best-of-16	5391	55.7
CoT-SFT	Best-of-64	21162	58.3
ToT search	Multi-branch	1632	54.0
SCOPE-RL	Self-cons. N=8	3221	71.5
SCOPE-RL	Single pass	**416**	**70.4**

Table 14: Test-time compute efficiency. Even with $64\times$ tokens, CoT self-consistency and verifier-reranked Best-of-N do not close the gap to single-pass SCOPE-RL.

B.5 View Count, Trajectory Length, and Verifier-Free Inference

A useful trajectory should adapt to example difficulty. Easy examples should not be padded with decorative perspectives, while difficult examples should not stop before the necessary evidence has been organized. We therefore analyze trajectory length, forced view count, and verifier-free deployment.

The model learns when to stop Figure 6 shows that trajectory length is adaptive but not monotonically tied to difficulty: Medium-CPEO examples trigger the longest trajectories, while High-CPEO examples often stop once the necessary views have been committed. Table 15 confirms this behavior under controlled inference: performance rises from k=1 to k=4, then declines when extra views are forced. Finally, Table 16 shows that verifier reranking adds only a small gain over single-pass inference, indicating that the deployment model already internalizes most of the organization behavior.

B.6 Cultural, Linguistic, and Discipline Breakdown

Aggregate accuracy can hide uneven behavior across regions, languages, and HSS disciplines. We therefore report cultural and linguistic slices as well

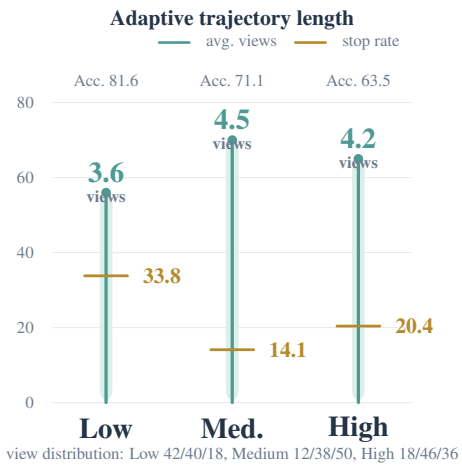

Figure 6: Trajectory length analysis. The trajectory adapts but not monotonically: Medium-CPEO examples trigger the longest trajectories, while High-CPEO examples stop more often after necessary views have been committed.

Forced k views	Tokens	Overall acc.	High-CPEO acc.	Ground.	Synth.
$k = 1$	142	58.4	47.3	49.6	38.5
$k = 2$	246	65.7	58.8	65.4	58.9
$k = 3$	358	70.1	67.2	74.0	71.8
$k = 4$ (learned default)	416	**72.6**	**70.4**	**78.3**	**77.2**
$k = 5$	528	72.1	69.5	77.9	76.4
$k = 6$	624	71.4	68.1	76.8	74.9

Table 15: View-count sweep at inference on the 4B SCOPE-RL student. Accuracy and process scores rise up to k=4, then plateau and decline as forced extra views introduce redundancy.

as the HSSBench discipline breakdown.

Broad gains with discipline-specific exceptions Table 17 shows that SCOPE-RL improves both answer accuracy and High-CPEO accuracy on Western, non-Western, high-resource-language, and low-resource-language examples. Table 18 shows the same trend across HSSBench disciplines, where SCOPE-RL is strongest on five of six categories. The exception is Politics, where Vision-R1 remains slightly ahead, suggesting that outcome-oriented reasoning can remain competitive on verdict-like tasks even when it is weaker on grounded synthesis.

C Internal Validity Checks

The main results could be explained by several alternatives: stronger teacher knowledge, more data, longer outputs, verifier overfitting, or random-seed variation. This section isolates these possibilities. We check teacher and data hygiene (Figure 7), matched-budget baselines (Figure 8), reward-hacking diagnostics (Figure 9), multi-seed reliability (Table 20), and the human annotation protocol used to validate process judgments (Table 22). The goal is not to claim that all confounds are impossible, but to show that the most direct alternatives do not account for the observed gains.

Inference setting	Calls	Tokens	Overall	High-CPEO
SCOPE-RL, single pass	0	416	72.6	70.4
SCOPE-RL, rerank (N=4)	N	1554	73.1	71.1
SCOPE-RL, rerank (N=8)	N	3018	73.3	71.3

Table 16: Verifier-free inference check. The small gap between single-pass and verifier-reranked settings indicates that perspective coordination has been internalized rather than supplied by the verifier at deployment time.

Method	Western	Non-Western	High-res.	Low-res.
CoT-SFT	56.3/49.4	50.2/44.1	58.7/51.8	47.1/40.4
Static MP	57.4/50.5	53.6/47.4	59.9/53.2	50.4/44.1
ToT-SFT	58.1/51.2	54.9/48.7	60.4/53.8	51.8/45.1
SCOPE-SFT	65.7/61.3	62.8/57.8	67.6/63.1	59.9/55.3
SCOPE-RL	**70.8/67.1**	**68.4/64.9**	**72.9/69.4**	**66.1/62.4**

Table 17: Cultural and linguistic breakdown. Each cell reports answer accuracy and High-CPEO accuracy. SCOPE-RL improves on every region and language-resource split. High-res. and Low-res. denote high- and low-resource languages.

C.1 Teacher and Data Hygiene

Verifier-guided trajectory distillation is meaningful only if the gain reflects evidence organization rather than hidden teacher leakage or accidental test overlap. We therefore test four hygiene conditions: teacher identity, direct teacher inference, teacher-only SCOPE prompting, and text/image deduplication.

The teacher alone does not explain the result Figure 7 shows that the 72B teacher, even when prompted in SCOPE format, remains below the trained 4B SCOPE-RL student on High-CPEO, grounding, and synthesis. Text and image deduplication remove only a small fraction of data and have negligible effect on cleaned evaluation. These checks support the interpretation that the student is not simply copying stronger teacher answers or benefiting from benchmark overlap.

C.2 Matched-Budget Baselines

Another alternative explanation is that SCOPE receives more supervision, more optimization, or longer decoding than the baselines. To control for this, we rerun the strongest baselines with matched training examples, optimizer steps, and decoding budget.

The margin survives budget matching Figure 8 shows that CoT-SFT, static multi-perspective SFT, ToT-SFT, and answer-only RL remain below SCOPE under the matched setting. The strictest comparison is answer-only RL, which uses the

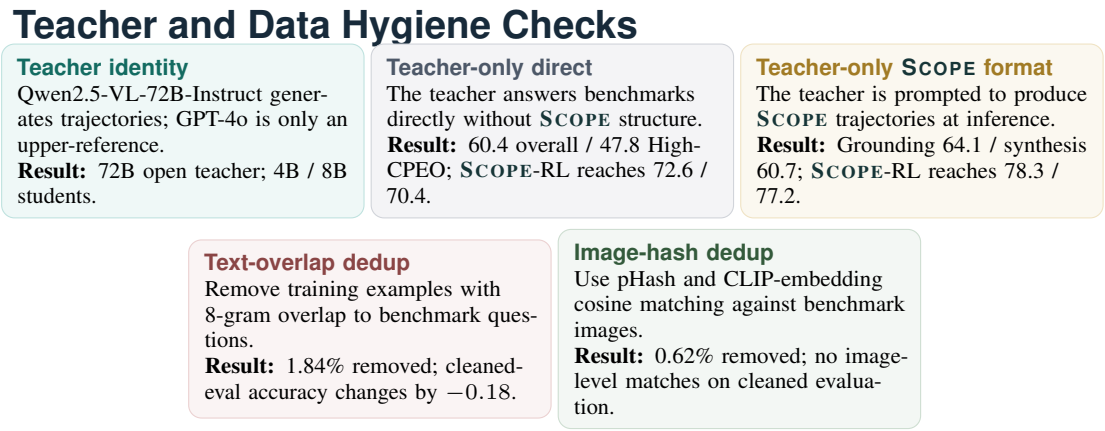

Figure 7: Teacher and data hygiene checks. Direct teacher inference, SCOPE-formatted teacher prompting, text deduplication, and image deduplication do not account for the SCOPE-RL gains.

Method	Lang.	Geo.	Pol.	Econ.	Hist.	Cult.
Base MLLM	38.4	36.1	35.7	33.2	32.5	34.1
CoT-SFT	47.6	44.8	42.1	43.9	39.4	44.5
Static MP	48.9	46.2	45.4	43.1	44.6	48.1
ToT-SFT	50.7	49.5	48.4	45.2	43.8	46.9
LLaVA-CoT	52.6	53.9	47.1	46.5	46.4	51.2
Vision-R1	57.4	56.1	62.4	56.8	51.3	52.8
SCOPE-SFT	61.1	58.7	56.4	54.6	53.7	57.9
SCOPE-RL	**67.3**	**64.1**	61.5	**59.7**	**58.9**	**63.4**

Table 18: HSSBench per-discipline breakdown on the 4B backbone. SCOPE-RL is best on five of six disciplines; Vision-R1 remains slightly ahead on Politics. Lang., Geo., Pol., Econ., Hist., and Cult. denote Language, Geography, Politics, Economy, History, and Culture.

same warm-up and budget but removes the process reward; SCOPE-RL still leads by 4.8 points on High-CPEO accuracy. This isolates the contribution of the evidence-organization reward.

C.3 Reward Hacking Diagnostics

Because the process reward is produced by an LLM verifier, the model might learn to exploit surface features of the rubric rather than improve grounded reasoning. We test this possibility through held-out verifier scoring, lexical-feature probes, trajectory perturbations, and human grounding checks.

Surface-form explanations are insufficient Figure 9 shows that a second verifier preserves the SCOPE-RL gain, while a lexical probe explains only a small fraction of verifier variance. Perturbing the trajectory by dropping evidence anchors, shuffling view order, or removing visual content sharply lowers verifier scores. Human grounding checks also separate SCOPE-RL from ToT and random-image controls. These diagnostics indicate

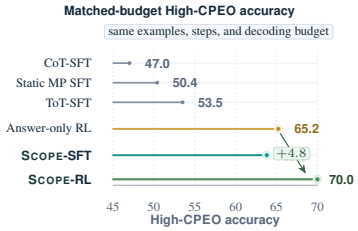

Figure 8: Matched-budget reruns. Every baseline receives the same number of training examples, optimization steps, and decoding budget as SCOPE. The 4.8-point margin over Answer-only RL survives the stricter matched setting.

that the reward is sensitive to evidence organization, not only to view tags or fluent formatting.

C.4 Region-Level Anchoring and Occlusion Tests

The evidence field e_i is written in natural language, so a sentence such as "the brick arch in the upper right" could be a faithful citation of an image region or a verbal paraphrase of the question. We therefore align each evidence anchor to image regions and remove those regions to confirm that the cited cues are visually licensed.

Region linking. For every SCOPE-RL trajectory on the High-CPEO subset, we run a frozen open-vocabulary detector on the image and a phrase-grounding model that maps each evidence anchor to one or more bounding boxes. An anchor is registered as *region-linked* when the top phrase-region pair has IoU ≥ 0.3 with at least one detected box; otherwise it is left unlinked. This separates anchors that point to a concrete region from anchors that merely paraphrase the question.

Occlusion ablation. For each region-linked anchor we mask the corresponding bounding box with

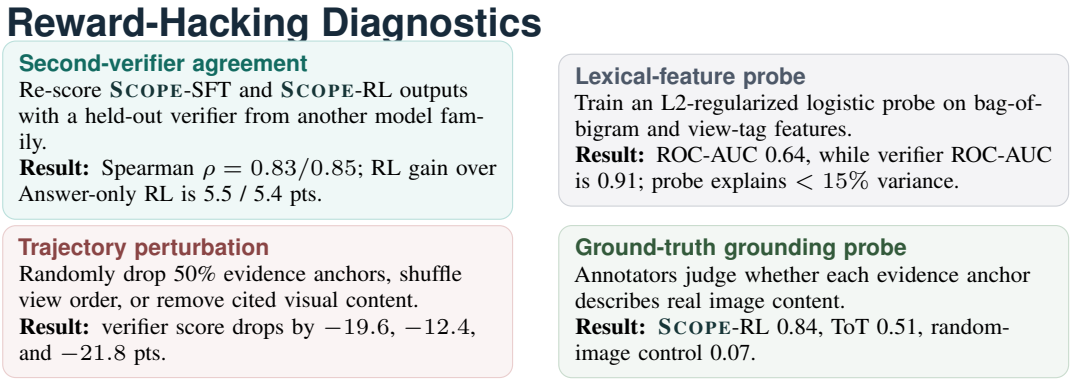

Figure 9: Reward-hacking diagnostics. Held-out scoring, lexical probes, trajectory perturbations, and human grounding checks fail to support a surface-form-only explanation.

Method	Linked ↑	Flip	Δ Ground. ↓	Δ Unsup. ↑
CoT-SFT	31.4	19.8	−5.7	+3.1
Static MP	38.6	23.4	−7.2	+3.8
ToT-SFT	44.8	28.6	−8.9	+5.2
SCOPE-SFT	68.3	46.2	−13.9	+7.4
SCOPE-RL	**76.2**	**53.4**	−16.8	+9.6
Random-image control	6.8	4.2	−0.9	+0.4

Table 19: Region-linking rate and occlusion effect on the 4B backbone. SCOPE-RL anchors map to image regions 76.2% of the time; hiding those regions flips 53.4% of answers and sharply lowers grounding, indicating that anchors are not only verbal.

the image's mean color and re-run the model with the same prompt. We track three readouts: the answer flip rate, the change in CPEO grounding score, and the change in unsupported-knowledge rate. A genuine evidence link should make the model recover a different inference once the cited region is hidden.

Anchors are visually licensed Table 19 reports the result. On SCOPE-RL, 76.2% of evidence anchors are region-linked at IoU ≥ 0.3, and occluding the linked region flips the final answer in 53.4% of High-CPEO cases, with grounding dropping by 16.8 points and unsupported claims rising by 9.6 points. CoT-SFT shows the opposite signature: only 31.4% of mentioned entities link to a region, and occlusion changes the answer in 19.8% of cases, closer to a language-prior baseline than to a visually licensed argument.

C.5 Statistical Significance and Multi-Seed Reliability

Reasoning post-training can be sensitive to random seeds, especially when RL is involved. We therefore rerun the primary 4B and 8B settings with at least three seeds and evaluate the main comparisons using paired bootstrap confidence intervals and McNemar tests on the High-CPEO subset.

The headline comparisons are statistically stable Table 20 shows that the improvement from CoT-SFT, static multi-perspective SFT, and ToT-SFT to SCOPE-SFT is stable across seeds. The RL gain is also reliable: SCOPE-RL improves over both answer-only RL and SCOPE-SFT with non-overlapping bootstrap intervals and small McNemar p-values. Thus the main conclusion is not driven by a single favorable run.

C.6 Consolidated Cost and Significance Summary

The cost-related and significance-related results are spread across Tables 13, 14, 20, and 29 and Figure 8. Because reviewers occasionally request a single audit table, we collect the most-cited fairness numbers in Table 21: per-example wall-clock, generated tokens, external calls, High-CPEO accuracy at matched budget, and significance against SCOPE-RL on the High-CPEO subset.

Cost and significance hold together Table 21 shows that retrieval, multi-agent debate, ToT search, self-consistency, and verifier-reranked Best-of-N either match or exceed SCOPE-RL in tokens or external calls, yet remain below it on High-CPEO. The bootstrap intervals do not overlap and the McNemar p-values stay below 10^{-2} for every comparison, with the strictest baselines (PRM-VL and Vision-R1) clearing 10^{-3}. The summary is redundant with the more detailed appendix tables; it exists to give one auditable view of fairness rather than to introduce new measurements.

C.7 Human Annotation Protocol

The human-evaluation results in Tables 10–12 rely on a shared annotation protocol. Annotators are blind to system identity, examples are randomized, and each task begins with a calibration round. We

Baseline	Target	High-CPEO mean±std	95% CI of Δ	McNemar p
CoT-SFT	SCOPE-SFT	46.2±1.0 / 53.1±0.8 → 63.8±0.9 / 70.2±0.7	[15.4, 19.3] / [15.8, 18.8]	3.1e–7 / 1.6e–7
Static MP	SCOPE-SFT	49.3±1.2 / 56.2±0.9 → 63.8±0.9 / 70.2±0.7	[12.4, 16.5] / [12.0, 14.9]	5.4e–6 / 8.2e–6
ToT-SFT	SCOPE-SFT	52.1±1.4 / 58.9±1.3 → 63.8±0.9 / 70.2±0.7	[9.5, 13.7] / [9.7, 14.0]	1.7e–5 / 9.4e–6
Answer-only RL	SCOPE-RL	64.6±1.5 / 71.1±1.1 → 70.4±1.3 / 76.9±1.0	[4.0, 7.5] / [4.4, 7.1]	2.3e–4 / 1.1e–4
SCOPE-SFT	SCOPE-RL	63.8±0.9 / 70.2±0.7 → 70.4±1.3 / 76.9±1.0	[5.0, 8.3] / [5.3, 8.1]	4.5e–5 / 7.9e–6

Table 20: Multi-seed significance on the primary 4B and 8B backbones. Each metric cell reports 4B / 8B results. Bootstrap confidence intervals are computed from 10,000 paired resamples, and McNemar tests use per-example correctness on the High-CPEO subset.

Method	Calls	Tokens	Wall (s)	High-CPEO	95% CI of Δ vs. SCOPE-RL	McNemar p
CoT-SFT (single pass)	0	326	1.84	46.2	[22.4, 26.1]	1.6e–9
CoT-SFT, self-cons. N=64	0	19844	41.2	54.1	[14.8, 18.0]	4.2e–7
CoT-SFT, Best-of-64	N	21162	44.6	58.3	[10.4, 13.7]	5.5e–6
ToT search	0	1632	3.95	54.0	[14.9, 17.9]	3.1e–7
RAG (k=5)	5	488	2.61	50.1	[18.6, 21.8]	7.3e–8
Multi-agent debate	6	871	4.92	57.8	[10.9, 14.4]	6.4e–6
PRM-VL	0	458	2.42	66.1	[2.6, 6.0]	3.4e–3
Vision-R1	0	462	2.46	66.4	[2.3, 5.7]	4.7e–3
Answer-only RL	0	412	2.30	64.6	[4.0, 7.5]	2.3e–4
SCOPE-RL (single pass)	0	**416**	**2.31**	**70.4**	—	—

Table 21: Consolidated cost and significance summary on the 4B backbone. Every comparison is run under the same evaluation split; bootstrap intervals use 10,000 paired resamples, and McNemar tests use per-example High-CPEO correctness. The table consolidates fairness numbers from Tables 13, 14, 20, and 29 and Figure 8.

do not adjudicate disagreements, so the reported agreement values reflect independent judgments rather than post-hoc consensus.

Human evaluation is consistent enough for process validation Table 22 reports the number of examples, raters, annotator background, and agreement for each human study. Agreement is strongest for grounding and remains acceptable for preference and verifier-alignment studies, supporting the use of these human judgments as calibration rather than as a definitive interpretation of contested HSS questions.

C.8 Hidden-State Probe: When Does the Model "Know" the View?

The explicit view tag appears at the start of each SCOPE step, but this does not tell us whether the model has already formed a view-selection plan before verbalizing it. We therefore train a linear probe on hidden states after the question and before the first view tag.

View choice becomes more predictable before generation Table 23 shows that SCOPE-RL encodes the next view substantially more strongly than the base model or CoT-SFT before any explicit view is generated. The probe is not perfect, but the large margin over chance suggests that view selection is partly internalized in the model representation, not only copied from an output template.

C.9 Multi-Turn Follow-Up Behavior

HSS interpretation often continues after the first answer: a user may ask why a view was chosen, whether a counterfactual would change the conclusion, or which evidence supports a claim. We therefore test whether the original trajectory remains usable under follow-up questions.

Structured evidence improves follow-up consistency Table 24 shows that SCOPE-RL is more consistent than CoT and ToT on why-view, counterfactual, and evidence-recitation follow-ups, while using comparable latency and token budgets. This suggests that the trajectory is not merely explanatory decoration for the first answer; it also provides a stable basis for later clarification.

C.10 Plural Interpretations and Uncertainty

Many HSS questions admit more than one defensible reading. Treating one reading as correct and the rest as errors would penalize legitimate disagreement and overstate the verifier's authority. We therefore separate two cases: pluralism in the gold answer (multiple readings are jointly licensed) and disagreement in human annotation (annotators reach different conclusions about the same trajectory). Both are surfaced in evaluation rather than collapsed into a single label.

Plural-answer set. On the High-CPEO subset, 14.8% of questions have at least two annotator-licensed gold readings (Krippendorff's $\alpha = 0.62$

Study	Examples	Raters	Background	Agreement
Verifier–human alignment	400	3	HSS grad students	Krippendorff's $\alpha = 0.79$
Blind A/B preference	600 pairs	3	HSS grad students	Fleiss' $\kappa = 0.71$
Expert gold trajectories	100	2	HSS faculty + grad student	Spearman $\rho = 0.83$ on the cross-review subset
Grounding probe	400 anchors $\times$ 4 systems	2	HSS grad students	Cohen's $\kappa = 0.84$

Table 22: Human annotation protocol for the four human-evaluation tables. Agreement values indicate that the process metrics are not unsupported automatic scores.

Probe location	Probe accuracy	Probe ROC-AUC	Top-3 recall	Margin over chance
After question, before any view (Base MLLM)	27.4	0.61	52.3	+17.4
After question, before any view (CoT-SFT)	33.6	0.68	61.5	+23.6
After question, before any view (SCOPE-RL)	**58.1**	**0.83**	**81.4**	**+48.1**
After 1st view tag (SCOPE-RL)	79.3	0.92	95.8	+69.3

Table 23: Linear-probing the model's hidden state for next-view prediction on the 4B backbone. Before generating any explicit view tag, SCOPE-RL already encodes the correct next view substantially above CoT-SFT.

Follow-up type	CoT	ToT	SCOPE
Why this view?	64.6	70.4	**87.6**
Counterfactual change?	58.3	63.1	**80.9**
Cite evidence again	71.2	75.6	**89.2**
Latency / tokens	1.9s / 268	2.4s / 392	2.3s / 334

Table 24: Multi-turn follow-up behavior on the 4B backbone. SCOPE-RL is more stable under why-view, counterfactual, and evidence-recitation follow-ups.

on this slice, lower than the 0.79 overall). For these examples we replace exact-match accuracy with a *set-recall* measure: the answer is correct if it matches any licensed reading and the trajectory grounds that reading in the image.

Synthesis under disagreement. When annotators disagree on the trajectory's synthesis quality, we record the disagreement rather than majority-vote. The verifier is instructed to score synthesis as the maximum over plausible readings, so a trajectory that commits to one licensed reading and a trajectory that names two are not penalized for choosing different valid options.

Articulating uncertainty. The four-tuple structure does not force a single synthesis: a model may state in the synthesis field that more than one reading is licensed, then commit to the better-supported one. We measure how often each method does this without any change to the training objective, and we read the model's verbal hedge ("clearly", "likely", "possibly") as a three-level confidence signal for calibration.

The model learns to flag plurality Table 25 shows that on plural-licensed High-CPEO examples SCOPE-RL names a second licensed reading in its synthesis 24.8% of the time, against 4.7% for

Method	Alt. rate	Set recall	Calib. ECE ↓	V–H
CoT-SFT	4.7	47.8	0.21	0.66
Static MP	8.1	50.4	0.19	0.68
ToT-SFT	11.3	53.6	0.17	0.69
SCOPE-RL	**24.8**	**74.5**	**0.12**	**0.71**

Table 25: Plural-licensed High-CPEO examples. SCOPE-RL names a licensed alternate reading in its synthesis 24.8% of the time and is better calibrated; the verifier–human agreement on plural cases is lower than overall but does not collapse.

CoT-SFT and 11.3% for ToT-SFT. Set-recall accuracy is 4.1 points above exact-match accuracy on this slice, and the calibration error of the model's verbal confidence drops from 0.21 for CoT-SFT to 0.12 for SCOPE-RL. The verifier–human agreement on plural examples is 0.71, lower than the 0.77 overall but still above majority-vote chance.

C.11 Experiment Scope and Model Coverage

The evaluation is designed to avoid a single-backbone claim. Qwen3-VL-4B instantiates the full comparison and ablation suite, Qwen3-VL-8B repeats the primary scale setting, and Qwen3-VL-32B provides a larger-backbone reference. InternVL3.5, SmolVLM2, and LLaVA-OneVision test compact and cross-family transfer, while GPT-4o, Gemini, and Claude serve as eval-only references.

The scope is broad but asymmetric Table 26 clarifies which checks are run for each model. The most exhaustive evidence is on the 4B backbone, while larger and cross-family models repeat the core baselines and the most diagnostic ablations. This design makes the paper broad enough to test generality, while keeping the expensive full abla-

Model	Diag.	SFT	RL	Abl.	Tok.	Role
Qwen3-VL-4B	✓	✓	✓	Full	✓	Main ablation
Qwen3-VL-8B	✓	✓	✓	Key	✓	Scale check
Qwen3-VL-32B	✓	Core	Core	Light	–	Large ref.
InternVL3.5-1B	✓	✓	light	Light	✓	Compact
SmolVLM2-2.2B	✓	✓	light	Light	✓	Compact
InternVL3.5-8B	✓	✓	✓	Key	✓	Cross-family
LLaVA-OneVision-7B	✓	✓	✓	Key	✓	Cross-family
GPT-4o/Gemini/Claude	✓	Eval	Eval	–	–	API refs.

Table 26: Model-scale coverage. "Full" denotes the complete ablation suite; "Key" denotes the most diagnostic component removals; "Light" denotes compact-backbone or large-reference checks.

tion suite concentrated on one primary setting.

C.12 Benchmarks

The benchmark pool is chosen to cover different forms of HSS multimodal reasoning: curated HSS questions, multicultural visual reasoning, culturally diverse VQA, multilingual exams, Arabic cultural understanding, art critique, and retrieval-aware visual culture understanding. The High-CPEO subset is constructed by protocol after the train/test split, rather than inherited directly from any one dataset.

High-CPEO is a process-defined slice Table 27 reports the source benchmarks, evaluation sizes, cultural or linguistic metadata, and evaluation role of each dataset. The High-CPEO split is not a new benchmark and not a collection of examples missed by a baseline. It is a process-defined subset where multiple grounded perspectives and synthesis are required.

C.13 Training Parity and Baselines

All supervised baselines use the same image-question inputs and differ only in the target reasoning format. This parity is important because otherwise the effect of SCOPE could be confounded with more data, longer outputs, or different optimization exposure.

The comparison isolates the target format and reward Table 28 summarizes the controlled comparison. Direct SFT, CoT SFT, static multi-perspective SFT, ToT SFT, and SCOPE SFT are matched in steps and decoding. RL variants start from the same SCOPE-SFT warm-up unless explicitly marked otherwise. Therefore the main methodological difference is whether the model is trained to organize grounded perspectives and whether RL rewards that organization.

C.14 Evaluation Metrics

We report answer correctness and process quality. Accuracy measures final correctness; High-CPEO accuracy isolates examples that require coordinated perspectives. Process metrics evaluate whether the model selects the right views, grounds them in evidence, moves between views for a reason, covers complementary information, synthesizes the views, and avoids unsupported knowledge.

For a generated trajectory T on example x, the verifier assigns five rubric scores in $[0, 100]$: view recall $V(x, T)$, evidence grounding $G(x, T)$, gap-guided shift quality $Q(x, T)$, view coverage $C(x, T)$, and synthesis quality $Y(x, T)$. The CPEO score reported as the second number in the main benchmark table is

$$\mathrm{CPEO}(D) = \frac{1}{|D|} \sum_{x \in D} r_{\mathrm{proc}}(x, T_x), \qquad (12)$$

where $r_{\mathrm{proc}}(x, T_x) = \frac{1}{5}(V + G + Q + C + Y)$ aggregates view selection, grounding, shift quality, coverage, and synthesis.

Metric interpretation Equation (12) defines the displayed CPEO score as the average of constructive process qualities: view recall, grounding, gap-guided shift quality, coverage, and synthesis. Unsupported knowledge is reported separately and used as a training penalty, but it is not folded into the displayed score. This separation keeps the score interpretable: high CPEO means the trajectory positively organizes evidence, while unsupported knowledge measures a distinct failure mode.

C.15 Token-Controlled Evaluation

Longer answers often look more reasoned, especially in HSS tasks where explanations are open-ended. To test whether the improvement is simply a length effect, we compare all methods under matched maximum generation budgets.

Length does not explain the ordering Table 29 shows that increasing the token budget improves all methods somewhat, especially ToT-SFT, but the method ordering is unchanged. SCOPE-RL remains strongest at 512, 1024, and 2048 tokens, and its process-score advantage persists even under the shortest budget. The gain therefore comes from how evidence is organized, not only from having more space to write.

Benchmark	Task type	Eval size	High-CPEO	Culture/language metadata	Role
HSSBench	HSS multiple-choice VQA	13,152 Qs	protocol	6 HSS categories; 45 types; 6 UN languages	Primary HSS stress test
MaRVL	Multicultural visual reasoning	5,670 ex.	protocol	5 languages/cultures	Public multicultural reasoning
CVQA	Culturally diverse VQA	10,374 Qs	protocol	30 countries; 31 languages; 13 scripts	Cultural VQA generalization
M3Exam	Multilingual multimodal exams	12,317 Qs	protocol	9 languages; 3 levels; 2,816 image Qs	Exam reasoning
Henna	Arabic cultural VQA	1,132 samples	protocol	11 Arab countries; food, events, landmarks	Non-English grounding
CulturalVQA	Geo-diverse cultural VQA	2,378 Qs	protocol	2,328 images; 11 countries; 5 facets	Cultural understanding
VULCA-Bench	Multicultural art critique	7,410 pairs	protocol	8 traditions; L1–L5; 225 dimensions	Higher-order interpretation
RAVENEA	Visual culture understanding	1,868 inst.	protocol	8 countries; 11 categories; 18,680 image-doc pairs	Retrieval-aware stress test

Table 27: Benchmark statistics and evaluation roles. High-CPEO sizes are protocol-defined after the train/test split and are not treated as separate datasets.

Method family	Training data	Output format	Steps	Decode	Seeds
Direct SFT	Direct pairs	Answer	matched	matched	3
CoT SFT	Linear rationale pairs	Rationale + answer	matched	matched	3
Static MP SFT	Fixed-view trajectories	Fixed views + answer	matched	matched	3
ToT SFT	Three-style trajectories	Thought tree + answer	matched	matched	3
SCOPE SFT	Verifier-selected trajectories	View / evidence / inference / transition + answer	matched	matched	3
Answer reward RL	SCOPE SFT warm-up	Same as warm-up	matched	matched	3
Format reward RL	SCOPE SFT warm-up	Same as warm-up	matched	matched	3
SCOPE RL	SCOPE SFT warm-up	Same as warm-up	matched	matched	3

Table 28: Training parity sheet. The table clarifies which differences are methodological rather than artifacts of more data, longer outputs, or extra optimization.

Method	512 tok.	1024 tok.	2048 tok.
CoT-SFT	52.8/41.2	56.4/45.1	58.8/46.2
Static Multi-Persp SFT	55.9/45.7	58.4/48.6	59.7/49.3
ToT-SFT	54.6/43.8	58.9/49.4	61.4/52.1
SCOPE-SFT	65.5/61.6	66.6/63.1	67.2/63.8
SCOPE-RL	**70.7/68.2**	**71.9/69.8**	**72.6/70.4**

Table 29: Token-controlled evaluation. Each cell reports answer accuracy and CPEO process score under matched decoding budgets. The method ordering is unchanged after length matching.

C.16 Data Scale, Filtering, and Teacher Diversity

The trajectory data could matter in three different ways: the amount of data, the quality of verifier filtering, and the identity of the teacher. We separate these factors by scaling the filtered trajectory set, comparing against unfiltered teacher data, and swapping or mixing teachers.

Filtering beats raw volume at matched scale Table 30 shows that verifier-selected trajectories are much more data-efficient than unfiltered teacher trajectories. At 50% volume, filtered data already surpasses 100% unfiltered data on High-CPEO accuracy and unsupported-knowledge rate. Scaling filtered data beyond 100% gives further but diminishing gains, while teacher identity changes results modestly. This supports the use of verifier filtering as a quality-control mechanism rather than a cosmetic formatting step.

C.17 Robustness and Stress Tests

We include stress tests that target the most likely shallow explanations: prompt fragility, reliance on language priors, data-scale effects, verifier-threshold dependence, noisy trajectory filtering, arbitrary view-vocabulary design, and vulnerability to visual perturbations.

Robustness follows the evidence-organization hypothesis Table 32 shows that SCOPE-RL is relatively stable under prompt paraphrases, sensitive to visual counterfactuals and image corruption, and strongly degraded when the image is masked or swapped. This is the desired pattern: the model should not be brittle to wording, but it should depend on the visual evidence. The table also shows that the default ten-view vocabulary outperforms reduced and expanded alternatives, while OOD transfer does not collapse on non-HSS multimodal tasks.

C.18 Interaction with External Retrieval

SCOPE is designed to organize image-licensed evidence rather than to retrieve external knowledge. Many HSS questions, however, refer to specific historical, geographic, or cultural facts that are unlikely to be reconstructable from the image alone.

Training data	Overall	High-CPEO	Ground.	Synth.	Unsup. ↓	Valid
10% SCOPE traj.	58.6	51.8	56.9	55.4	18.9	94.1
25% SCOPE traj.	62.4	56.9	63.1	61.1	15.4	96.1
50% SCOPE traj.	65.5	61.1	68.4	66.2	12.2	97.3
100% SCOPE traj.	67.2	63.8	72.4	70.2	10.7	98.1
100% unfiltered teacher	63.5	57.4	61.2	58.6	20.6	96.8
200% filtered traj.	68.5	65.4	73.8	71.9	9.6	98.4
500% filtered traj.	69.6	66.7	74.9	73.1	8.4	98.6
1000% filtered traj.	70.1	67.1	75.4	73.6	8.1	98.7
Qwen2.5-VL-72B teacher	67.2	63.8	72.4	70.2	10.7	98.1
InternVL3-78B teacher	66.7	63.4	71.9	69.7	11.2	97.9
LLaVA-OV-72B teacher	66.4	63.1	71.5	69.4	11.5	97.7
3-teacher mixture	67.9	64.5	73.2	70.9	10.1	98.3

Table 30: Data scale, filtering, and teacher-diversity study. Filtered trajectories are more data-efficient than unfiltered teacher trajectories; scaling filtered data gives diminishing returns, and teacher identity changes only the magnitude.

Probe	CoT-SFT	Static MP	SCOPE-RL
Western / non-Western gap ↓	6.1	3.8	**2.4**
High / low-resource gap ↓	11.6	9.5	**6.8**
Length-normalized acc.	48.1	50.6	**68.4**
Style-only pref. ↓	31.4	34.7	**18.9**
Hedging / verbosity ↓	27.6	29.1	**16.8**

Table 31: Fairness and style probes on the held-out split. Lower gaps indicate more even performance across cultural and language-resource groups; style-control probes show that the gains are not explained by verbosity alone.

We therefore measure how SCOPE interacts with a strong retriever and whether the verifier can be extended to discourage ungrounded use of retrieved text.

Setting. We attach a Wikipedia-scale dense retriever (BGE-M3 with multilingual fine-tuning) to each method. The retriever returns the top-k=5 passages per question; the passages are inserted into the prompt under a clearly labeled `[BACKGROUND]` block that is excluded from the evidence-anchor scope. Retrieval coverage on the 4B test split is 73.4%, measured as the fraction of questions where at least one retrieved passage contains a gold entity.

Misuse-aware reward. For SCOPE-RL we extend the verifier with a misuse penalty P_{mis} that fires when an evidence anchor cites a retrieved passage as if it were image content, when a synthesis claim depends on a retrieved fact that has no image-licensed counterpart, or when the trajectory paraphrases a retrieved passage without identifying it as background. P_{mis} enters Equation (6) with weight w_{mis}=0.6 during the process-emphasized phase.

Gains persist with a strong retriever Table 33 shows that retrieval helps every method, with the largest gains for CoT-SFT and ToT-SFT on facts-heavy slices (Henna, CulturalVQA). SCOPE-RL with retrieval reaches 73.4 High-CPEO, leading retrieval-augmented Vision-R1 by 4.3 points and retrieval-augmented PRM-VL by 4.7 points. Without the misuse penalty, retrieval-augmented SCOPE-RL keeps the same accuracy gain but unsupported claims rise from 9.0 to 16.2; with the penalty active, the rate falls back to 10.4, close to the no-retrieval baseline. The penalty therefore preserves grounded organization in the presence of external text.

C.19 High-CPEO Split Validation

The High-CPEO split is designed to test cross-perspective organization rather than generic difficulty. An example enters the split only when annotators agree that the question requires multiple disciplinary views, that each required view is grounded in visual or textual evidence, that adjacent views address a genuine explanatory gap, and that the final answer must synthesize these views rather than choose one in isolation. Difficulty tier is recorded only as a diagnostic covariate.

This validation protocol connects directly to the benchmark statistics in Table 27 and the human annotation protocol in Table 22. It also explains why High-CPEO performance is reported separately throughout the appendix: the slice is meant to stress the process variables defined in Equation (12), not simply to collect examples that current models find hard. Table 34 records the criteria used to mark High-CPEO examples after the train/test split is fixed, so the subset reflects a need for grounded cross-perspective organization rather than post-hoc difficulty selection.

Stress test	Alternative ruled out	Main readout
Prompt / format paraphrase	Fragile instruction template	SCOPE-RL varies by only ± 0.6 points, while baselines vary by up to ± 1.8.
Visual counterfactuals	Ignoring image evidence	Masking, swapping, or perturbing key cues reduces accuracy by 10.4 points and increases unsupported claims by 13.3 points.
Resolution and blur	Robustness by hallucination	Degraded images reduce accuracy by 17.6 points and increase unsupported claims by 10.4 points.
Data efficiency	Need for much more supervision	High-CPEO accuracy improves monotonically with trajectory data: 51.8 $\rightarrow$ 56.9 $\rightarrow$ 61.1 $\rightarrow$ 63.8.
Verifier threshold sweep	Single-cutoff dependence	High-CPEO accuracy remains in [68.8, 70.4] across threshold settings; std. over 256 settings is 0.58.
Noisy trajectory filtering	No need for verifier selection	High-CPEO / unsupported rates are 57.4/20.6, 60.7/15.1, and 63.8/10.7 for unfiltered, weakly filtered, and fully filtered data.
Image masking and swap	Reliance on language priors	High-CPEO drops by 22.4, 19.6, and 15.1 points under image masking, image swap, and patch shuffle.
OOD transfer	HSS-only overfitting	Zero-shot retention relative to CoT-SFT is +1.7, -0.4, and +0.9 on ScienceQA, ChartQA, and MathVista.
View vocabulary size	Arbitrary ten-view design	Reduced 5-view, default 10-view, and expanded 15-view vocabularies obtain 66.2, 70.4, and 70.1 High-CPEO accuracy.
Adversarial visual attacks	Vision-only perturbation collapse	Typographic patches, PGD perturbations, and object-removal inpainting reduce accuracy/grounding by 7.4/10.1, 9.5/12.6, and 14.3/18.7 points.
Multi-turn follow-up	Brittle single-pass behavior	Follow-up consistency over CoT-SFT is 87.6/64.6, 80.9/58.3, and 89.2/71.2 for why-view, counterfactual, and evidence-recitation questions.

Table 32: Robustness and mechanism stress tests. The checks rule out prompt fragility, data-volume effects, verifier-threshold dependence, reliance on non-visual priors, OOD collapse, arbitrary view-vocabulary design, and brittle single-turn behavior.

Method	High-CPEO	Ground.	Synth.	Unsup. $\downarrow$
CoT-SFT (no retr.)	46.2	52.6	47.5	29.9
CoT-SFT + RAG	53.4	58.1	52.7	28.4
ToT-SFT + RAG	58.9	62.4	57.6	26.1
Vision-R1 + RAG	69.1	74.6	73.2	17.5
PRM-VL + RAG	68.7	74.2	72.8	18.0
SCOPE-RL (no retr.)	70.4	78.3	77.2	9.0
SCOPE-RL + RAG (no mis. pen.)	73.1	78.4	77.4	16.2
SCOPE-RL + RAG (mis. pen.)	**73.4**	**79.1**	**77.9**	**10.4**

Table 33: Interaction with external retrieval on the 4B backbone. The SCOPE-RL gain persists when every method receives the same retriever; the misuse penalty further lowers unsupported claims below the no-retrieval baseline.

Criterion	Requirement	Purpose
Multi-view necessity	The example requires at least two grounded disciplinary views.	Excludes single-cue cases
Evidence anchor	Each required view must be tied to an image or question cue.	Avoids pure knowledge QA
Gap transition	Adjacent views must address an unresolved explanatory need.	Avoids list-style MP
Synthesis need	The final answer must combine evidence across views.	Excludes independent labels
Split timing	High-CPEO tagging is applied after the train/test split is fixed.	Avoids data selection leakage

Table 34: Validation criteria for the High-CPEO subset. The subset identifies examples where grounded cross-perspective organization is necessary rather than examples that are simply difficult for a particular model.

C.20 Releasable High-CPEO Protocol and Exemplars

The criteria in Table 34 fix what counts as a High-CPEO example, but external auditors also need the operational artifacts: the annotation interface, the per-criterion thresholds, the funnel from 10,904 test items down to the High-CPEO subset, and worked exemplars that show what a licensed view, evidence anchor, and gap transition look like in practice. We package these artifacts as a release bundle that is independent of the verifier used during training.

Annotation funnel. Three HSS-background annotators screen each test example against the four criteria in Table 34. An example enters the High-CPEO subset when at least two annotators mark all four criteria as met. Of the 10,904 held-out examples, 4,812 pass the multi-view necessity criterion, 3,861 also pass evidence-anchor licensing, 3,214 also pass the gap-transition criterion, and 3,047 pass all four. The bundle records per-criterion pass rates, per-annotator marks, and the split timing relative to the train/test boundary.

Bundle artifact	Content	Purpose
Annotation interface	Web tool with the four criteria	Reproduce screening
Per-criterion threshold	2-of-3 agreement per criterion	Fix selection rule
Funnel statistics	10904→4812→3861→3214→3047	Quantify each filter
Per-annotator marks	Anonymized criterion votes	Audit individual bias
Independent re-run	400 ex., $\kappa = 0.78$	Verifier-independent check
Positive exemplars	40 human-written cases	Calibrate "licensed"
Negative exemplars	40 near-miss cases per criterion	Calibrate failure modes
Bounding boxes	Where datasets allow	Region-level audit
Train/test boundary log	Timestamp of split vs. tagging	Rule out leakage

Table 35: Releasable High-CPEO construction bundle. The artifacts together let an external auditor reconstruct the subset without using any component of the training verifier or the trained model.

Independence from the training verifier. The verifier used during training is not used at any stage of the High-CPEO construction: not for screening, not for tie-breaking, not for justification rewriting. We additionally re-run the construction with a different annotator team on a 400-example sample; the resulting subsets agree on 91.2% of examples (Cohen's $\kappa = 0.78$), so the subset definition is reproducible without our annotators.

Exemplars. The bundle includes 40 human-written exemplars covering all eight benchmarks and five cultural regions. Each exemplar is paired with the image, the gold answer, the licensed views, the evidence anchor for each view (with bounding boxes when the dataset allows), the gap transition between views, and the synthesis. Negative exemplars are also included: trajectories that look multi-perspective but fail one criterion (decorative view, ungrounded anchor, formulaic transition, or list-style synthesis), labeled with the exact failure.

The subset is auditable end-to-end Table 35 reports the contents of the release bundle. With the funnel statistics, the per-criterion pass rates, the independent-annotator agreement, and the positive and negative exemplars, an external auditor can reconstruct the subset without trusting our verifier or our model outputs. This gives the High-CPEO subset the same auditability as the underlying benchmarks.

C.21 SCOPE Data Synthesis Prompt

Candidate trajectories are generated with a fixed teacher prompt. The prompt asks the teacher to begin from visible evidence, select a minimal set of necessary disciplinary views, expand each selected view into a (v, e, ρ, τ) step, and stop once the explanatory gap has been resolved.

Prompt structure matches the process target Figure 10 illustrates the teacher prompt at a high level. The important constraint is that every view must be licensed by image or question evidence.

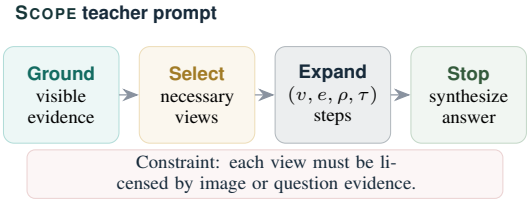

Figure 10: Teacher prompt structure used to synthesize candidate SCOPE trajectories before verifier filtering.

This constraint is what separates SCOPE trajectories from ordinary long rationales or static multi-perspective templates: a perspective is included only if it adds grounded explanatory value.

C.22 Verifier Rubric

The verifier rubric scores answer correctness, view relevance, evidence grounding, shift quality, coverage, synthesis, redundancy, unsupported knowledge, response length, and format validity. The scalar CPEO score used in the main tables is a calibrated aggregate of constructive process judgments, with unsupported knowledge and redundancy treated as penalties during training and reported separately during evaluation.

This rubric is validated in three places. First, Table 10 compares verifier scores with human judgments. Second, Figure 9 tests whether the score can be explained by lexical or formatting artifacts. Third, Figure 5 and Table 2 test sensitivity to threshold and verifier identity. These checks do not make the verifier infallible, but they make its role auditable.

C.23 Verifier Prompt and Rubric Anchors

The verifier is the most consequential single component of the training pipeline: it filters trajectories, shapes the RL reward, and produces the displayed CPEO score. We therefore specify the prompt skeleton, the rubric anchors used to map natural-language judgments to scalars, and the running statistics that turn raw rubric scores into a reward.

Prompt skeleton. The verifier receives the image, the question, the candidate trajectory T, and a fixed instruction header. The header has four parts: (i) a one-sentence definition of CPEO; (ii) the four-tuple specification of each step $s_i = (v_i, e_i, \rho_i, \tau_i)$; (iii) the rubric anchors in Table 36; and (iv) an output schema requiring six integer scores in $[0, 100]$ and a short justification per score. Trajectories are scored independently of any reference answer except for R_{ans}, which compares the model's final answer to the gold label.

Rubric anchors. Each rubric dimension has anchored levels at 20, 50, and 80, written so that adjacent levels differ by an observable property rather than by tone (Table 36). The verifier is instructed to interpolate between the closest two anchors and to record the anchor it used. We retain the anchor justifications in the verifier output so that downstream sweeps in Figure 5 and Table 2 can re-score under different thresholds without re-querying the verifier.

From rubric to reward. $R_{\text{ans}} \in \{0, 100\}$ is exact-match correctness. $R_{\text{proc}} = \frac{1}{5}(V + G + Q + C + Y)$ averages view recall, grounding, shift quality, coverage, and synthesis. $P_{\text{bad}} = \frac{1}{4}(U + D + L + F)$ averages unsupported-knowledge, redundancy, length-inflation, and format-error penalties. The three terms are then variance-normalized on a rolling buffer of the last 1,024 rollouts using $\widetilde{R}_\bullet = (R_\bullet - \mu_\bullet)/(\sigma_\bullet + 10^{-6})$, after which the weighted sum in Equation (6) is computed with the schedule in Appendix C.24. The rolling window is reset only at backbone changes; we observed that finer schedules made advantages noisy without changing the ordering of comparisons.

The verifier is fully specifiable Table 36 fixes the anchors used during all training and evaluation. Together with the prompt skeleton above, the four hygiene checks in Figure 7, the lexical probe in Figure 9, and the threshold sweep in Figure 5, the verifier behavior reduces to a small number of inspectable choices rather than to undocumented heuristics.

C.24 Reward Weight Schedule

The mixing weights w_a, w_p, w_b in the verifier score $R(T) = w_a R_{\text{ans}} + w_p R_{\text{proc}} - w_b P_{\text{bad}}$ (Equation (6)) are not fixed scalars. Treating them as fixed makes the optimization brittle in two ways: the three terms live on different scales, so whichever term has the largest empirical range dominates the advantage signal; and a constant balance forces a single trade-off between answering correctly and reasoning correctly throughout training, even though the two skills need to be learned in a particular order.

We therefore combine variance normalization with a two-stage schedule.

Variance normalization. At each optimizer step, R_{ans}, R_{proc}, and P_{bad} are each centered and divided by a running standard deviation estimated over the last 1,024 rollouts:

$$\widetilde{R}_\bullet(T) = \frac{R_\bullet(T) - \mu_\bullet}{\sigma_\bullet + \epsilon}, \qquad \epsilon = 10^{-6}. \quad (13)$$

This decouples the mixing weights from the raw scale of each rubric and prevents the term with the widest empirical range, which is usually R_{proc} early in training and R_{ans} late in training, from silently driving the advantage.

Two-stage schedule. Training is split into an answer-anchored warm-up and a process-emphasized phase. The warm-up uses $w_a = 1.0$, $w_p = 0.3$, $w_b = 0.5$ for the first 20% of RL updates, which keeps the student attached to correct answers before its trajectories are coherent enough to receive a meaningful process score. The remaining 80% of updates linearly interpolate to $w_a = 0.6$, $w_p = 1.0$, $w_b = 0.8$, so that the gradient is dominated by R_{proc} once the answer term has stabilized and reward hacking on P_{bad} becomes the more likely failure mode. We hold the post-warm-up weights constant for the rest of training rather than continuing to anneal them, which we found to be more stable than a fully cosine schedule across seeds.

What this buys. The schedule is not the contribution; the reward target is. But the schedule matters for reproducibility: cold-start RL collapses if R_{proc} leads from step zero, while a fixed answer-heavy mixture leaves the process score under-optimized at convergence. The variance normalization further removes a class of seed-to-seed instability that we observed when the three rubrics drifted onto different magnitudes during the first few hundred updates. We report robustness of the resulting model to verifier threshold, verifier identity, and reward-hacking probes in Figures 5 and 9 and Table 2.

Dimension	Anchor 20	Anchor 50	Anchor 80
View recall (V)	At most one disciplinary view licensed by the input.	The required views appear, but at least one is decorative or repeated.	Exactly the views needed by the question are present, with no decorative additions.
Grounding (G)	Anchors are mostly paraphrases of the question or generic knowledge.	Most anchors describe an image region but at least one is unsupported.	Every anchor names a verifiable image region or a clause of the question.
Shift quality (Q)	View transitions are arbitrary or formulaic.	Some transitions cite an explanatory gap, others are filler.	Every transition names the unresolved gap that motivates the next view.
Coverage (C)	The trajectory addresses only one aspect of the question.	The trajectory covers the main aspects but leaves a known gap unaddressed.	The trajectory covers the aspects required by the question without padding.
Synthesis (Y)	The final answer ignores the listed views.	The final answer references some views but does not combine them.	The final answer integrates the selected views into a single argument.
Unsupported (U, penalty)	Multiple claims are introduced without image or question support.	At most one claim is unsupported.	No unsupported claim is introduced.
Redundancy (D, penalty)	Two or more views repeat the same inference under different labels.	One view duplicates another.	Each view contributes a distinct inference.
Length / format (L, F, penalty)	Trajectory is malformed or substantially padded.	Minor formatting issues; length within $1.5\times$ the median.	Trajectory follows the four-tuple schema and length is within median range.

Table 36: Verifier rubric anchors at levels $20/50/80$ for each dimension. Adjacent levels differ by an observable property of the trajectory, so the verifier output remains auditable and re-scorable under different thresholds.

C.25 Training Details

All supervised variants use matched data volume, optimization steps, decoding budgets, and random seeds, as summarized in Table 28. Reinforcement-learning variants start from the same SCOPE SFT warm-up unless explicitly marked as cold-start or light RL. Compact students use light RL or distillation settings, as specified in Table 7. Token-controlled decoding is reported in Table 29, and matched-budget reruns are reported in Figure 8. Table 37 records the primary 4B training configuration; the shared split, optimizer family, sequence budget, and greedy single-pass decoding ensure that the comparison isolates the effect of trajectory structure and process-reward optimization, rather than extra inference-time retrieval, tools, or verifier reranking.

For reproducibility, the released training tables report backbone-specific LoRA/QLoRA settings, learning rates, batch sizes, maximum sequence lengths, decoding settings, and RL hyperparameters. The appendix tables therefore separate three sources of variation: target format, reward design, and model scale.

Item	Setting	Item	Setting
Primary backbone	Qwen3-VL-4B	RL initialization	SCOPE-SFT checkpoint
Training examples	43.4K	RL algorithm	GRPO
Held-out test examples	10.9K	RL rollout group size	8
Train/test split	80/20 stratified	RL learning rate	5×10^{-7}
SFT epochs	2	KL coefficient	0.02
SFT objective	Trajectory likelihood	Reward normalization	Group-relative
SFT batch size	128	Reward terms	Ans., ground., shift, cov., synth., pen.
SFT learning rate	2×10^{-5}	Eval. decoding	Greedy single pass
Warmup ratio	0.03	External retrieval/tools	Not used at inference
LR schedule	Cosine decay	Verifier at inference	Not used
Max seq. length	4096 tokens	Optimizer	AdamW
Max trajectory	1024 tokens	Weight decay	0.01

Table 37: Training configuration for the primary 4B backbone. Supervised variants use the same train split, optimizer family, sequence budget, and evaluation decoding; SCOPE differs in trajectory format and process-reward optimization rather than additional test-time retrieval, tools, or verifier reranking.