REWARDING NON-VERIFIABLE INTERMEDIATE STEPS FOR PHOTOREALISTIC MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning for reasoning has shown strong results in mathematics and code generation, where intermediate steps can often be checked using deterministic rules or exact outcomes. In photorealistic multimodal reasoning, structured intermediate steps may also be needed, but they often lack a single exact answer and must instead be judged from visual and linguistic context. This makes it difficult to design reliable step-wise rewards for photorealistic multimodal reasoning. We propose a staged GRPO framework that first optimizes format and global accuracy rewards, then introduces step-wise rewards for intermediate reasoning. We consider two ways to specify the reasoning structure: one guided by human strategies and one proposed by a model from task examples. We introduce MOSAIC (Multimodal Object-Specific Ambiguity Inference from Context), a multilingual visual word inference testbed over photorealistic images that requires integrating visual and linguistic information and naturally elicits a structured reasoning process, which we decompose into four human-inspired steps. Experiments across 12 languages show improvements over prompt-based, supervised fine-tuning, and vanilla GRPO baselines, with the largest gains in lower-resource languages. Gains extend across Qwen and InternVL model families, and the larger models improve on three languages unseen during RL training. We further ask whether the reasoning structure must be manually specified. On SpatiaLQA, a distinct spatial-reasoning and action-planning task, we use model-proposed reasoning steps and evaluation criteria, yielding a 4.50-point improvement in F1-score over the baseline.

1 INTRODUCTION



Figure 1: Challenge of rewarding intermediate steps in photorealistic multimodal reasoning compared with mathematical reasoning.

Reasoning in real-world multimodal settings often requires resolving ambiguity from incomplete or unfamiliar language. When a sentence contains an unfamiliar or missing term, humans can often infer its intended meaning by combining linguistic context with what is visible in the scene. Doing so requires perceptual and logical inferences rather than a single lookup, illustrating how photorealistic multimodal reasoning can involve a structured intermediate process. AI models would face similar challenges when deployed in diverse cultures, as the models may encounter unfamiliar terms, unusual objects, or objects referred to with unique regional names. To succeed in responding to user queries, they would need to use an intermediate reasoning structure (perceptual and logical inferences) to resolve the ambiguity (or handle the missing information).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

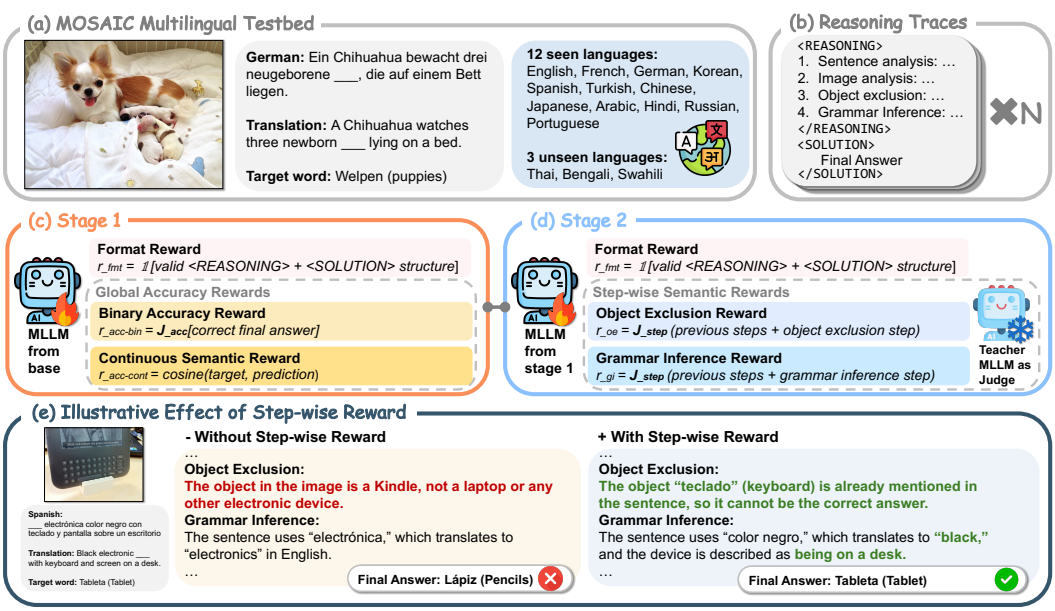


Figure 2: Overview of our staged GRPO framework instantiated for the human-inspired MOSAIC task. **(a)** MOSAIC. **(b)** Human-inspired reasoning decomposition. **(c)** Stage 1 with format and global final-answer rewards. **(d)** Stage 2 with format and step-wise reasoning rewards. **(e)** Effect of step-wise rewards on reasoning and final ambiguity resolution.

Yet this intermediate reasoning structure is rarely used as an explicit training signal in multimodal reasoning. Our goal is to improve final-answer performance by explicitly rewarding structured intermediate reasoning. Many multimodal benchmarks evaluate a mixture of perception, knowledge, and reasoning (Lu et al., 2022; Hudson & Manning, 2019; Goyal et al., 2017; Schwenk et al., 2022; Yue et al., 2024; Lu et al., 2024), but supervision is typically concentrated on final answers rather than the quality of intermediate reasoning. Recent work has begun to explore process rewards in multimodal settings, particularly for mathematical reasoning (Luo et al., 2025). However, rewarding intermediate reasoning in photorealistic settings is more challenging. As illustrated in Fig. 1, mathematical steps can often be checked against exact numerical results, formal rules, or other deterministic criteria. In photorealistic multimodal reasoning, there may be no single exact target for an intermediate step: multiple descriptions can be valid, and a step may be partially correct while omitting relevant information. Extending reinforcement learning approaches such as GRPO to this setting therefore requires rewards that assess the semantic quality of intermediate reasoning, i.e., whether a step is grounded in the image and linguistic context and appropriate for its intended role in solving the task.

A key challenge is to define the intermediate reasoning steps and their evaluation criteria for a given task. We study two approaches: a human-inspired decomposition based on human reasoning strategies, and a model-proposed decomposition derived from the task definition and examples. Our general framework trains with GRPO in two stages: Stage 1 establishes task-level performance using format reward (which enforces the model to include the required steps) and global final-answer reward, while Stage 2 introduces automatically-judged step-wise rewards.

As our human-inspired instantiation, we introduce MOSAIC (Multimodal Object-Specific Ambiguity Inference from Context), where an MLLM must infer a missing object noun from a photorealistic image and an incomplete sentence, as demonstrated in Fig. 2(a). The ambiguity lies in determining which visible object the sentence refers to. Resolving it requires combining sentence meaning with visual candidates, excluding objects that are already mentioned or implausible, and using grammatical cues. Although the final objective resembles standard fill-in-the-blank prediction (Lu et al., 2019; Su et al., 2020; Li et al., 2019), our focus is on the structured multimodal reasoning process required to infer the missing word by integrating visual and linguistic evidence. We construct MOSAIC with 12 seen languages and 3 additional unseen transfer languages, and evaluate six open-source MLLMs.

Performance varies substantially across languages, with lower-resource languages such as Turkish, Arabic, and Hindi remaining challenging.

For MOSAIC, human ambiguity-resolution strategies motivate four interpretable steps: Sentence Analysis, Image Analysis, Object Exclusion, and Grammar Inference. Human annotation shows that current MLLMs perform relatively well on the perceptual steps but struggle more with Object Exclusion and Grammar Inference, motivating us to target these steps with intermediate rewards. On MOSAIC, our framework outperforms prompt-based, supervised fine-tuning, and vanilla GRPO baselines. The largest gains over the zero-shot baseline occur in lower-resource languages, reaching +14.3, +10.7, and +9.3 percentage points in Hindi, Turkish, and Arabic, respectively. For both Qwen2.5-VL-7B and InternVL3.5-8B, the trained models also improve average accuracy on three languages excluded from RL training.

To test whether the framework depends on a human-derived reasoning decomposition, we further instantiate it on SpatiaLQA (Xie et al., 2026), a spatial-reasoning and action-planning task, using model-proposed reasoning steps and evaluation criteria. Our framework improves F1-score by 4.50 percentage points over the baseline. Together, these results show that the framework can support both human-inspired and model-proposed reasoning decompositions across distinct multimodal reasoning tasks. In summary, we contribute:

1. A staged GRPO framework for semantically defined step-wise rewards in photorealistic multimodal reasoning, where intermediate steps often cannot be verified deterministically.
2. MOSAIC, a multilingual visual word inference testbed with 12 seen and 3 unseen languages, together with a four-step reasoning decomposition grounded in human strategies.
3. Improvements over prompting, SFT, and vanilla GRPO on Qwen, with gains extending across Qwen and InternVL model families and to unseen languages for larger models.
4. A model-proposed instantiation on SpatiaLQA, showing that task-specific reasoning steps and evaluation criteria can be obtained without deriving them from human reasoning studies.

2 Related Work

Question-answering tasks. (Antol et al., 2015; Goyal et al., 2017) pose the task of visual question answering, with mixed difficulty which complicates the analysis of the quality of the reasoning steps. Hudson & Manning (2019) requires complex sentence and image parsing, which we find less challenging with modern models. Schwenk et al. (2022); Lu et al. (2022); Yue et al. (2024) require world and scientific knowledge beyond the image (and beyond pure reasoning). More recently, there has been interest in multimodal mathematical tasks (Lu et al., 2024) and logic puzzles (Chia et al., 2024). We instead focus on reasoning over everyday data with a clear reasoning structure. Xie et al. (2026) requires integrating spatial understanding with multi-step logical reasoning to produce ordered steps for everyday tasks; it serves as an additional testbed for our proposed method.

Rewards in language and mathematical reasoning space. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) simplifies training by removing the need for a value model for calculating advantages for individual actions (Schulman et al., 2017), and instead comparing multiple outputs for the same prompt. (Uesato et al., 2022) propose rewards on the model's reasoning process, obtained directly as annotations from humans (Lightman et al., 2024) or indirectly, by considering whether a step in the reasoning process leads to good final outcomes using Monte Carlo (MC) sampling (Wang et al., 2024). Gunjal et al. (2026) prompts an LLM judge with instance-specific rubric criteria and aggregates the criterion-level scores into a scalar reward, but the rubrics score per-response quality aspects, rather than a fixed sequence of named reasoning steps. Process rewards were extended to multimodal reasoning (Luo et al., 2025; Wang et al., 2026; Chen et al., 2025), but largely in mathematical reasoning or structured visual domains such as charts, diagrams, and geometry, rather than everyday photorealistic scenes. Further, steps are typically not named hence no rubric for scoring them can be defined. Verifiable rewards (Guo et al., 2025) have also gained popularity, which are not directly applicable in all multimodal reasoning tasks.

Rewards in broader mathematical reasoning. Recent work has begun to adapt R1-style reinforcement fine-tuning to vision-language models beyond mathematical reasoning. Visual-RFT (Liu et al., 2025b) applies GRPO-style fine-tuning with task-specific verifiable rewards, including classification accuracy, detection IoU, and confidence rewards. Reason-RFT (Tan et al., 2025) combines SFT and

RL for visual reasoning tasks such as counting, structural perception, and spatial transformation, using task-specific global accuracy rewards. These works are closely related to ours in their use of reinforcement fine-tuning for multimodal models, but their rewards remain primarily output-level and verifiable. In contrast, we reward explicitly categorized intermediate reasoning steps in photorealistic multilingual reasoning, where step correctness is semantically defined and not uniquely verifiable. Related chain-of-thought training methods such as ARES (Byun et al., 2024) also train reasoning traces, but do not study structured, categorized step-wise rewards.

3 Semantic Rewards for Photorealistic Reasoning

3.1 Staged GRPO Framework Overview

Prior applications of GRPO (Shao et al., 2024) focus on symbolic or mathematical tasks, where intermediate steps can often be verified deterministically (see Fig. 1). In photorealistic reasoning, such verification is often unavailable, motivating our *semantic* rewards. A key component of our framework is to decompose task reasoning into structured intermediate steps. We study two ways to obtain this decomposition and its evaluation criteria: a human-inspired approach (Sec. 4) and a model-proposed approach (Sec. 5). Given one or more images and a task prompt, the model samples structured responses containing the prescribed reasoning steps and a final answer. We train with GRPO in two stages: Stage 1 optimizes format and global final-answer accuracy rewards (including that the model includes the required reasoning steps), while Stage 2 continues from the Stage 1 checkpoint using the format reward and step-wise rewards. This staged design first establishes task-level performance, then provides supervision for the intermediate reasoning. It incorporates the reasoning structure (steps) in two ways (format reward in Stage 1, then quality of steps in Stage 2).

3.2 Stage 1 Global Reward Design

We use separate global rewards to enforce structured output (format) and response quality (accuracy). For an input $x = (\mathcal{I}, p)$, where $\mathcal{I}$ contains one or more images and p is a task prompt, we sample a group of K responses $\mathcal{G}(x) = \{R_1, \ldots, R_K\}$, where $R_i \sim \pi_\theta(\cdot \mid x)$. Each response should contain a reasoning trace and a final answer, marked by specific delimiters; the *format* reward is defined as:

$$r_i^{\text{fmt}} = \begin{cases} 1, & \text{if } R_i \text{ contains } \texttt{<REASONING>}\ s_1 : \ldots \cdots s_M : \ldots\ \texttt{</REASONING>} \\ & \quad \text{followed by } \texttt{<SOLUTION>}\ \hat{y}_i\ \texttt{</SOLUTION>}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

Here, $s_1, \ldots, s_M$ denote the prescribed reasoning-step headings, whose presence is verified by pattern matching; the task-specific steps are given in Secs. 4 and 5.

For *accuracy* rewards, we extract the final answer $\hat{y}_i$ between the <SOLUTION> tags through pattern matching. We first employ a judge model $\mathcal{J}_{acc}$ to assess whether the extracted answer $\hat{y}_i$ is correct given the reference answer y and input x. The binary accuracy reward is defined as:

$$r_i^{\text{acc-bin}} = \begin{cases} 1, & \text{if } \mathcal{J}_{acc}(\hat{y}_i, y, x) = \text{Yes}, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

In photorealistic reasoning, valid answers may not be unique, and even among incorrect answers, some may be closer to the correct answer than others. We therefore also use a continuous accuracy reward to provide graded signals toward the correct answer:

$$r_i^{\text{acc-cont}} = \mathcal{A}_{\text{cont}}(\hat{y}_i, y, x), \tag{3}$$

where $\mathcal{A}_{\text{cont}}$ is a task-specific scoring function that measures semantic agreement of the final answer.

3.3 Stage 2 Step-wise Reward Design

To capture the quality of non-verifiable intermediate reasoning, we introduce step-wise rewards. Unlike mathematical reasoning, steps may be partially correct, incomplete, or contain minor hallucinations while still contributing useful information. We prompt the model to generate task-specific

reasoning steps and evaluate each using $\mathcal{J}_{step}$ according to task-specific criteria, with all preceding steps as context. The step-wise reward sums the step scores:

$$r_i^{\text{step}} = \sum_{s \in \mathcal{S}} \mathcal{J}_{step}\left(R_i^{(\leq s)}, x\right), \qquad \mathcal{S} \subseteq \{s_1, \ldots, s_M\}. \tag{4}$$

where $R_i^{(\leq s)}$ contains the reasoning component s and all preceding components in response $R_i = (R_i^{(s_1)}, \ldots, R_i^{(s_M)}, \hat{y}_i)$. The set $\mathcal{S}$ specifies which reasoning steps receive rewards. During training, manually scoring the sampled reasoning traces is infeasible. Thus, we use an automatic judge $\mathcal{J}_{step}$, a frozen MLLM, to evaluate each step according to task-specific criteria.

4 Human-Inspired Reasoning for MOSAIC

4.1 MOSAIC: Multimodal Object-Specific Ambiguity Inference from Context

MOSAIC is a multilingual visual word inference testbed in which a model must identify a missing object noun using a photorealistic image and an incomplete sentence. We choose this task because its intermediate reasoning requires visual grounding and contextual interpretation and cannot always be verified against a single exact target. Given an image I and its caption $C^{(L)}$ in language L, we select an object noun W referring to a physical entity in I and replace it with `[blank]` to obtain the incomplete caption $\tilde{C}^{(L)}$. The model receives $x = (I, \tilde{C}^{(L)}, L)$ and generates a structured response containing intermediate reasoning and a final predicted word $\hat{W}$. This instantiates the general framework in Sec. 3 with $y = W$ and $\hat{y} = \hat{W}$.

Exact-match accuracy is too strict for our relatively open-vocabulary task. For example, if the object in the image is a bouquet of roses and the target word is "flower," a prediction of "rose" should also be considered correct. We therefore use GPT-4o-mini (OpenAI, 2024) as a judge. Given the predicted word $\hat{W}$ and the ground-truth word W, we prompt GPT to determine whether they can refer to the same object in the image. We use a fixed prompt (see Appendix G) and temperature 0 to minimize sampling variability. To validate automated evaluation, we compare GPT-based judgments with human annotations, observing 80% agreement on final correctness labels.

We construct our dataset from XM3600 (Thapliyal et al., 2022), which provides human-written captions for German, French, Korean, Spanish, and Turkish. We select target words that correspond to visible object nouns and can be aligned across the captions. To control for differences in which objects speakers of different languages choose to mention (Kádár et al., 2018), we generate an English caption containing the selected target word and translate it into additional languages: Chinese (Simplified), Japanese, Arabic, Hindi, Russian, and Portuguese (quality was checked on a subset of the data with fluent or knowledgeable speakers). MOSAIC contains 12 *seen* languages with a disjoint 80/20 train/test split for staged GRPO training and evaluation. We additionally construct 3 *unseen* language test sets: Thai, Bengali, and Swahili, which are excluded from RL training and used only to evaluate language generalization. Full details are provided in Appendix D.

4.2 Human-Inspired Reasoning Decomposition

Recent work on *human* reasoning (Wang et al., 2025) finds that people use several strategies to infer unknown words from visual context, including selecting visually salient objects, using spelling similarity, excluding objects already recognized in the text, and grammatical cues to infer properties of the target word. We focus on object exclusion and grammar inference because they require integrating visual and linguistic information, while the first two primarily rely on unimodal cues.

To implement these strategies for *model* reasoning, we organize the reasoning process according to its information dependencies. Later steps rely on information established earlier; for example, object exclusion depends on first identifying objects mentioned in the sentence and visible in the image. We therefore use four reasoning steps in the following order: Sentence Analysis (SA), Image Analysis (IA), Object Exclusion (OE), and Grammar Inference (GI), followed by the final answer. This decomposition defines the task-specific reasoning steps used to instantiate the general framework in Sec. 3. Detailed instructions and scoring criteria for each step are provided in Appendix E.



Original sentence: Ein **Chihuahua** bewacht **drei** neugeborene ___, die **auf** einem **Bett** liegen.

In English: A **Chihuahua** looks after **three** newborn ___ who lie **on** a **bed**.

Target word: Welpen (puppies)

1. Sentence Analysis: *Sentence translates to "A **Chihuahua** watches three newborn [blank] lying on a **bed**." Objects are chihuahua and bed.*

Hallucination: 2 (no issues) - Translation: 2 (correct) - Extra inference: 2 (no issues)

2. Image Analysis: ***Chihuahua** and three **pups** on a **bed** covered with a **blanket**. Pups are the target.*

Image description: 2 - Extra inference: 0

3. Object Exclusion: *The sentence already mentions the **chihuahua** and **bed**, so those should not be considered as candidates for the blank. Background objects like **blanket** can be excluded. The most plausible remaining candidate for the blank is **puppies**.*

Preserved right answer: 2 - Excluded mentioned objects: 2 - Excluded irrelevant objects: 2

4. Grammar Inference: *The word "three" specifies the number of objects corresponding to the target. Gender specifies feminine noun. The preposition "on" explains the target lies on another object.*

Prepositions: 2 - Gender: 0 (no gender specified) - Numbers: 2 - Hallucination: 2 (no issues)

5. Final answer: *Welpen (puppies)* Score: C (analysis is incorrect, final answer is correct)

Figure 3: Illustration of the human-inspired reasoning decomposition and its evaluation criteria. Colors indicate how the available context should be processed to earn maximal scores for each step.

Table 1: Zero-shot multilingual accuracy (%) across 6 MLLMs and 12 languages. EN: English, FR: French, ZH: Chinese–Simplified, PT: Portuguese, DE: German, ES: Spanish, JA: Japanese, RU: Russian, KO: Korean, AR: Arabic, TR: Turkish, HI: Hindi.

Model	**Avg**	**EN**	**FR**	**ZH**	**PT**	**DE**	**ES**	**JA**	**RU**	**KO**	**AR**	**TR**	**HI**
Qwen2.5VL 3B (Bai et al., 2025b)	71.7	88.0	80.4	81.0	82.0	79.8	78.0	75.5	73.0	72.2	55.0	60.4	34.5
Qwen2.5VL 7B (Bai et al., 2025b)	81.1	**92.5**	90.0	86.5	**87.5**	**89.4**	**87.0**	**88.5**	**86.5**	**75.8**	67.0	68.5	54.0
Qwen3VL 2B (Bai et al., 2025a)	67.4	89.0	76.9	82.5	80.0	75.8	67.0	71.5	65.0	64.1	46.0	45.2	45.5
Qwen3VL 4B (Bai et al., 2025a)	**82.3**	**92.5**	**91.5**	**87.5**	**87.5**	87.4	84.5	85.5	80.5	72.2	**73.5**	**69.5**	**75.0**
Gemma3 4B (Team et al., 2025)	74.7	81.5	79.9	79.5	81.0	79.8	76.5	74.0	75.0	65.2	68.0	62.9	72.5
Pangea 7B (Yue et al., 2025)	71.8	76.5	74.9	76.5	73.0	71.7	75.5	73.0	73.0	70.2	69.0	65.0	63.0
Avg (Models)	74.8	86.7	82.3	82.3	81.8	80.7	78.1	78.0	75.5	70.0	63.1	61.9	57.4

To assess whether these reasoning steps can be meaningfully distinguished and evaluated, we conduct human annotation (Appendix E and L) on zero-shot reasoning traces generated by Qwen2.5-VL-7B (Bai et al., 2025b). Human scores are higher for the perceptual steps (Sentence Analysis and Image Analysis) but lower for the logical steps (Object Exclusion and Grammar Inference). These findings identify OE and GI as the steps most in need of additional training signal.

Thus, as illustrated in Fig. 3, we retain all four reasoning steps in the generated trace, while applying step-wise rewards to OE and GI, i.e., $\mathcal{S} = \{\text{oe}, \text{gi}\}$. We instantiate the continuous accuracy reward with MetaCLIP2 (Chuang et al., 2025) similarity between the predicted and target words. For MOSAIC, Stage 2 replaces Stage 1 final-answer rewards with step-wise rewards; details are in Appendix M. We use a frozen Qwen2.5-VL-7B-Instruct model as the training judge for both final-answer correctness and step-wise criteria, and require intermediate reasoning to be in English.

4.3 Evaluation

4.3.1 Zero-shot task instruction baseline

We evaluate six open-source multilingual MLLMs using the zero-shot task instruction and GPT-based accuracy. As shown in Tab. 1, average accuracy ranges from 0.674 to 0.823. Performance varies across languages, with particularly low average accuracy on Arabic, Turkish, and Hindi, highlighting the difficulty of multilingual ambiguity inference in MOSAIC.

4.3.2 Staged GRPO versus Supervised Fine-tuning and Vanilla GRPO

We compare our staged GRPO framework with supervised fine-tuning (SFT) on Qwen2.5-VL-7B. The SFT baseline uses the same train/test split as our RL framework, but optimizes a cross-entropy loss on ground-truth final answers, wrapped in `<SOLUTION>` tags, using the standard task instruction

Table 2: Training-based methods on Qwen2.5-VL-7B (acc in %), including Avg across languages.

Method	Avg	EN	FR	ZH	PT	DE	ES	JA	RU	KO	AR	TR	HI
Zero-shot Task Instruction	81.1	92.5	90.0	86.5	87.5	89.4	87.0	**88.5**	86.5	75.8	67.0	68.5	54.0
SFT Baseline	82.2	**93.5**	88.9	**89.3**	85.4	87.9	85.5	87.5	87.4	**82.3**	71.2	68.0	59.3
Vanilla GRPO	83.0	88.4	87.9	85.3	82.4	84.3	85.5	86.5	83.3	81.3	75.3	**79.2**	**76.4**
Ours: Stage 1	84.3	89.9	**92.0**	84.8	**88.9**	91.4	**90.0**	85.0	88.4	80.8	**76.3**	78.2	66.3
Ours: Stage 1 + Stage 2	**85.3**	88.9	**92.0**	87.3	88.4	**92.4**	89.0	87.5	**92.4**	81.8	**76.3**	**79.2**	68.3

Comparison	Paired gain (pp)	One-sided p
Stage 1 − Vanilla GRPO	$+1.34$	0.0472
Stage 2 − Stage 1	$+0.97$	0.0490
Stage 2 − Continued Stage 1	$+1.13$	0.0245
Stage 2 − Zero-shot	$+5.08$	4.29×10^{-10}

as input. We do not train SFT on intermediate reasoning traces because ground-truth traces are unavailable at scale and MLLM-generated traces are imperfect. As shown in Tab. 2, SFT slightly improves average accuracy over the zero-shot baseline (0.822 vs. 0.811), while staged GRPO reaches 0.853. The difference is most pronounced in Turkish, Hindi, and Arabic, where SFT changes accuracy by -0.5, $+5.3$, and $+4.2$ percentage points, compared with $+10.7$, $+14.3$, and $+9.3$ points for staged GRPO. These results suggest that final-answer supervision alone is insufficient.

We further compare our method with Vanilla GRPO. Our Stage 1 enforces the four-step reasoning structure and supplements the binary accuracy reward with a continuous reward (MetaCLIP embedding similarity of the predicted and ground-truth answers). In contrast, the vanilla GRPO baseline uses a generic format reward and a binary final-answer accuracy reward computed by a frozen Qwen2.5-VL-7B-Instruct judge and generates *free-form, unstructured* reasoning followed by the final answer. It uses the same training split and number of training steps as Stage 1. As shown in Tab. 2, vanilla GRPO reaches an average accuracy of 0.830, while our Stage 1 reaches 0.843. Our Stage 2 further improves accuracy to 0.853.

We use one-sided exact McNemar tests for Stage 1 vs. vanilla GRPO and Stage 2 vs. Stage 1. Stage 1 improves over vanilla GRPO by 1.34 points ($p = 0.0472$), and Stage 2 adds another 0.97 over Stage 1 ($p = 0.0490$). To control for additional training, we continue Stage 1 for the same number of additional updates (as Stage 2) while retaining the Stage 1 reward composition. Stage 2 outperforms this Continued Stage 1 control by 1.13 points ($p = 0.0245$). Details are in Appendix M.

Additional results in Appendix O show that prompting (zero-shot structured reasoning) alone does not improve overall performance on our task, and that few-shot demonstrations and self-revision underperform the zero-shot task-instruction baseline (0.811), motivating our staged training approach.

4.3.3 Generalization across Model Sizes and Families

Table 3: Avg accuracy (%) across model sizes and families on the 12 training languages. Gain: Stage 2 - Baseline. Baseline is the Zero-shot Task Instruction method.

Model	Baseline	Stage 1	Stage 2	Gain (pp)	One-sided p
Qwen2.5-VL-3B	71.7	77.6	**78.7**	$+7.80$	1.63×10^{-17}
Qwen2.5-VL-7B	81.1	84.3	**85.3**	$+4.23$	4.29×10^{-10}
InternVL3.5-8B	77.5	79.3	**81.7**	$+4.20$	1.13×10^{-7}

We next demonstrate that the gains from our staged GRPO hold across model sizes and families. We train Qwen2.5-VL-3B/7B and InternVL3.5-8B with the same two-stage framework. As shown in Tab. 3, all three models improve from Baseline to Stage 2, by 7.80, 4.23, and 4.20 percentage points, respectively, with all gains statistically significant.

Table 4: Transfer to Thai, Bengali, and Swahili, which are unseen during training.

Model	Method	Thai	Bengali	Swahili	Unseen Avg	Δ Unseen (pts)
Qwen2.5VL-3B	Baseline	**65.7**	**61.1**	**45.0**	**57.3**	–
	Stage 1	63.6	41.9	40.0	48.5	-8.8
	Stage 2	63.6	40.4	41.0	48.3	-9.0
Qwen2.5VL-7B	Baseline	75.3	49.5	40.0	54.9	–
	Stage 1	78.3	**61.6**	41.0	60.3	+5.4
	Stage 2	**81.3**	60.6	**43.5**	**61.8**	+6.9
InternVL3.5-8B	Baseline	77.8	**80.3**	41.0	66.4	–
	Stage 1	78.3	78.8	39.5	65.5	-0.8
	Stage 2	**78.8**	78.3	**46.5**	**67.9**	+1.5

4.3.4 Transfer to Unseen Languages

We further evaluate transfer to three lower-resource languages, Thai, Bengali, and Swahili, unseen during RL training. As shown in Tab. 4, the larger Qwen2.5-VL-7B and InternVL3.5-8B models improve average accuracy on these unseen languages after our staged GRPO training, from 0.549 to 0.618 and from 0.664 to 0.679, respectively, compared to the Baseline (Zero-shot Task Instruction). The smaller Qwen2.5-VL-3B fails to improve, with average accuracy decreasing from 0.573 to 0.483.

4.3.5 Reward Ablation for Staged GRPO

Table 5: Ablation of reward design for GRPO on Qwen2.5VL 3B across 12 languages. Fmt: format reward. Acc-Bin-G/Acc-Bin-Q: binary accuracy reward from GPT-4o-mini / Qwen2.5VL 7B. Acc-Cont: continuous accuracy reward from MetaCLIP. S: All 4 step-wise rewards. oe: Object Exclusion step reward. gi: Grammar Inference step reward.

Reward	Avg	EN	FR	DE	KO	ES	TR	ZH	JA	AR	HI	RU	PT
Baseline	71.7	**88.0**	80.4	79.8	72.2	78.0	60.4	81.0	75.5	55.0	34.5	73.0	82.0
Global reward variants													
Fmt	66.8	76.4	72.2	74.1	57.1	64.8	56.4	–	–	–	–	–	–
Fmt+Acc-Bin-G	66.8	76.4	73.2	72.6	57.1	64.8	56.4	–	–	–	–	–	–
Fmt+Acc-Cont	67.3	77.4	70.7	78.2	70.8	60.6	62.3	–	–	–	–	–	–
Fmt+Acc-Bin-G+Acc-Cont	77.6	84.9	**86.9**	**85.9**	71.2	85.0	67.5	78.2	84.0	69.7	**52.3**	**84.8**	80.9
Fmt+Acc-Bin-Q+Acc-Cont	75.4	85.4	86.4	84.3	68.7	83.0	58.9	**81.7**	81.0	66.2	47.2	79.8	81.4
GRPO with step-wise rewards (not staged, unlike ours)													
Fmt+Acc-Bin-Q+Acc-Cont+S	67.7	82.4	78.9	74.8	62.1	76.5	56.4	70.1	71.5	51.5	43.2	65.7	78.9
Fmt+Acc-Cont+oe+gi	67.8	78.9	78.9	80.3	63.1	69.5	55.3	74.1	74.0	50.5	41.7	70.2	77.4
Ours (staged)	**78.7**	84.9	**86.9**	85.4	**74.8**	**87.5**	**69.0**	79.7	**85.0**	**72.7**	49.8	83.8	**84.9**

We conduct reward ablations on Qwen2.5-VL-3B for efficiency. Tab. 5 shows that on the six languages evaluated for the format-only, format+binary accuracy, and format+continuous similarity variants, the format reward alone reduces accuracy from 0.765 to 0.668, and adding binary accuracy or continuous similarity alone yields little improvement. Combining both global rewards with the format reward raises accuracy to 0.776 with GPT-4o-mini and 0.754 with a frozen Qwen2.5-VL-7B judge.

Introducing step-wise rewards from the beginning yields only 0.677–0.678 accuracy. In contrast, staged training reaches 0.787, improving over the strongest global-reward variant (0.776). The ablation supports our two-stage design: global rewards first establish task-level competence, and step-wise rewards then provide additional supervision for intermediate reasoning.

Table 6: Results on SpatiaLQA with Qwen3-VL-8B. p-values are one-sided.

Method	Content F1	p vs. Base	p vs. S1
Baseline: Zero-shot Task Instruction	40.20	–	–
Ours: Stage 1	42.65	0.0006	–
Ours: Stage 2	**44.70**	< 0.0001	0.0006

5 Model-Proposed Reasoning Steps for Spatial Reasoning

5.1 SpatiaLQA and Model-Proposed Step Design

We next evaluate whether the general framework can be instantiated without deriving the reasoning decomposition from human reasoning studies. We consider SpatiaLQA (Xie et al., 2026), a spatial-reasoning benchmark in which the model generates an action plan by reasoning about objects, spatial relations, and dependencies between actions. We partition the dataset by scene into 7,636 training examples and 1,969 test examples.

Given the task definition and 10 representative examples, we prompt GPT-5.5 (OpenAI, 2026) to propose task-specific reasoning steps and corresponding evaluation criteria. We make only minor wording refinements without changing the underlying reasoning structure. The resulting decomposition consists of Target-object Identification, Blocking-object Identification, Spatial-relation Identification, Dependency Reasoning, and Plan-generation Reasoning.

We train Qwen3-VL-8B using the two-stage framework in Sec. 3. In Stage 1, the official metric F1-scores provide continuous rewards, with no separate binary final-answer reward, as the task does not define a binary correctness metric for the generated plan. In Stage 2, all five reasoning steps receive criterion-based rewards from a frozen Qwen3-VL-8B-Instruct judge. Detailed reward formulations and step-evaluation criteria are provided in Appendix N.

5.2 Evaluation

Following the SpatiaLQA evaluation protocol, we report average Content F1. For each test example, GPT-4o-mini determines whether predicted and reference actions are semantically equivalent, and the Hungarian algorithm performs one-to-one matching between predicted and reference actions. We compute action precision and recall from these matches, take their harmonic mean as Content F1 for each example, and average across the test set. We use one-sided paired randomization tests with 100,000 sign-flip samples to evaluate improvements in Content F1. As shown in Tab. 6, average Content F1 increases from 0.4020 at baseline to 0.4265 after Stage 1 and 0.4470 after Stage 2. Stage 2 contributes a further 2.06 percentage points over Stage 1 ($p = 0.0006$), bringing the total improvement over the baseline to 4.50 points ($p < 0.0001$). These results show that the framework can be adapted to a distinct reasoning task using a model-proposed decomposition and evaluation criteria, with step-wise rewards providing additional improvement.

6 Conclusion

We study how MLLMs can benefit from structured multimodal reasoning in photorealistic settings where intermediate steps are difficult to verify exactly. We propose a staged GRPO training framework that first learns the required output and reasoning structure and global task accuracy, then introduces step-wise supervision for intermediate reasoning. We evaluate the framework on MOSAIC, our multilingual visual word inference testbed, with a human-inspired reasoning decomposition. Experiments across 12 languages show improvements over prompting, supervised fine-tuning, and vanilla GRPO baselines, with the largest gains in lower-resource languages. These improvements extend across Qwen and InternVL model families, with the larger models also improving average accuracy on three languages unseen during RL training. We further examine whether the reasoning decomposition must come from human strategies: on SpatiaLQA, model-proposed reasoning steps and evaluation criteria enable the same framework to improve F1-score by 4.50 percentage points over the baseline. Limitations are discussed in Appendix P.

AI USE STATEMENT

As described in the paper, language and vision-language models are components of our methodology: they assist with dataset construction and translation, provide training-time reward signals and evaluation judgments, and propose task-specific reasoning steps and evaluation rubrics for the SpatiaLQA setting.

Separately, we used Codex to assist with manuscript language polishing, adapting training code across different base models, and preparing and submitting training and evaluation jobs to GPU servers. The authors take responsibility for the final manuscript, experimental methodology, code, and reported results.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Ju-Seung Byun, Jiyun Chun, Jihyung Kil, and Andrew Perrault. Ares: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse ai feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4410–4430, 2024.

Honghao Chen, Xingzhou Lou, Xiaokun Feng, Kaiqi Huang, and Xinlong Wang. Unveiling chain of step reasoning for vision-language models with fine-grained rewards. *Advances in Neural Information Processing Systems*, 38:114703–114727, 2025.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16259–16273, 2024.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen-tau Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd conference on computational natural language learning*, pp. 402–412, 2018.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

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

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025a.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Ruilin Luo, Zhuofan Zheng, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin Zeng, and Yujiu Yang. Unlocking multimodal mathematical reasoning via process reward model. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://iclr.cc/virtual/2024/poster/19094. Poster.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Introducing GPT-5.5, 2026. URL https://openai.com/index/introducing-gpt-5-5/.

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 146–162. Springer, 2022. doi: 10.1007/978-3-031-20074-8_9.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://neurips.cc/virtual/2025/loc/san-diego/poster/118345.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton,

Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 715–729, 2022.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 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)*, pp. 9426–9439, 2024.

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. VisualPRM400K: An effective dataset for training multimodal process reward models. In *International Conference on Learning Representations*, 2026.

Yufei Wang, Adriana Kovashka, Loretta Fernández, Marc N. Coutanche, and Seth Wiener. Towards understanding ambiguity resolution in multimodal inference of meaning. In *2025 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–8, 2025. doi: 10.1109/ICDL63968.2025.11204410.

Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Zhu Hao, Rui Tang, Rong Wei, Mingli Song, Yuanyu Wan, and Jie Song. SpatiaLQA: A benchmark for evaluating spatial logical reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2645–2657, 2026.

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

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=a3g2l4yEys.

A Appendix Overview

This appendix provides additional details, extended results, and analyses to support the main paper. It is organized as follows:

- **External RL-tuned checkpoint** (Sec. B): Inference-only evaluation of Visual-RFT on MOSAIC, included as a reference point rather than a controlled baseline.
- **GQA evaluation** (Sec. C): Sanity check for multimodal reasoning evaluation comparing the base model and our trained checkpoint on GQA.
- **Dataset construction details** (Sec. D): Target noun selection, filtering, translation procedure, and seen/unseen language construction.
- **Human evaluation details** (Sec. E): Annotation protocol, scoring rubrics, agreement analysis, and step-wise evaluation results.
- **Prompts** (Sec. F, Sec. G): Full prompts used for baseline models, structured reasoning, revision, and GPT-based evaluation.
- **Implementation details** (Sec. H): Description of answer extraction, training setup, and evaluation pipeline.
- **Additional evaluation analysis** (Sec. I): Comparison between GPT-based and MetaCLIP-based evaluation methods and their limitations.
- **Qualitative examples** (Sec. J): Detailed examples illustrating differences in reasoning behavior between baseline and trained models.
- **MOSAIC Error Analysis** (Sec. K): Manual error analysis.
- **Human annotation rubric** (Sec. L): Full rubric used for evaluating reasoning steps.
- **Training details** (Sec. M): Computational resources needed and per-model sampling and gradient accumulation settings across training stages.
- **Spatial reasoning implementation details** (Sec. N): Reward design, evaluation criteria, and scoring for the model-proposed reasoning steps.
- **Limitations** (Sec. P): Task applicability, semantic judges, and training computation.
- **Prompt-based baseline results** (Sec. O): Structured prompting, few-shot demonstrations, and self-revision.

B Inference-only Evaluation of an External RL-tuned Checkpoint

We additionally evaluate Visual-RFT (Liu et al., 2025b), a released RL-tuned multimodal checkpoint, on MOSAIC. This comparison is not a controlled baseline: Visual-RFT is trained for a different objective and uses Qwen2-VL-7B as its base model, while our staged GRPO model uses Qwen2.5-VL-7B. To make the comparison easier to interpret, we also evaluate the corresponding Qwen2-VL-7B base model.

As shown in Tab. 7, Visual-RFT does not improve over its corresponding Qwen2-VL-7B base model on MOSAIC. In contrast, in our controlled setting, staged GRPO improves over the Qwen2.5-VL-7B base model. We include this result only as a reference point for evaluating an existing multimodal RL-tuned checkpoint, rather than as a direct comparison between methods.

C Sanity Check: GQA Evaluation

To check whether our reward-based training preserves performance on another multimodal reasoning task, we evaluate the base model and our trained checkpoint on GQA (Hudson & Manning, 2019). We use two prompt settings. First, we query both the base model and our trained model with the standard GQA-style prompt, which asks for a short answer without requiring intermediate reasoning. Second, we query our trained model with our step-wise reasoning prompt, which requires the model to produce the four reasoning steps before the final answer.

As shown in Tab. 8, our trained model performs similarly to the base model under the GQA-style prompt, with accuracy changing from 0.608 to 0.609. When we use our step-wise reasoning prompt on GQA, accuracy is slightly lower at 0.581. This decrease may reflect prompt-task mismatch, since our reasoning steps are designed specifically for MOSAIC rather than the broader question types in GQA.

Table 7: Inference-only evaluation of an external RL-tuned checkpoint compared with our controlled Qwen2.5-VL-7B setting. Visual-RFT is not trained on MOSAIC and uses a different base model, so it is included only as a reference point rather than a controlled baseline.

Setting	Method	Avg
External RL-tuned checkpoint		
Qwen2-VL-7B	Zero-shot Task Instruction	0.748
Qwen2-VL-7B	Visual-RFT (Liu et al., 2025b)	0.725
Our controlled setting		
Qwen2.5-VL-7B	Zero-shot Task Instruction	0.811
Qwen2.5-VL-7B	**Staged GRPO (Ours)**	**0.853**

Table 8: GQA experiments on Qwen2.5-VL-7B. We evaluate on the `lmms-lab/GQA` testdev-balanced split using normalized exact-match accuracy. The GQA prompt follows the single-word/phrase answer format; the reasoning prompt uses our four-step reasoning format with `<REASONING>` and `<SOLUTION>` tags.

Method	Prompt	Accuracy	Correct / Total
Base Model	GQA	**0.608**	**7646 / 12578**
Our Method	GQA	0.609	7660 / 12578
Our Method	Reasoning	0.581	7304 / 12578

D Dataset Configuration Details

We start from the XM3600(Thapliyal et al., 2022) dataset with five languages: German, French, Korean, Spanish, and Turkish, which are human-generated. For each image, we first extract all nouns from the captions using SpaCy. For a noun n in language L, we query GPT-4o-mini to determine two conditions: (1) whether n is an object noun that refers to a physical entity, and (2) whether the corresponding object appears in the noun list of the remaining four languages. If both conditions are satisfied, we select n as the target noun. Otherwise, we continue to the next noun in the noun list of language L.

For each selected target noun, we next query GPT-4o-mini again to verify that it is not part of a compound noun phrase (e.g., "pizza" in "pizza box"). If no valid noun is found, we skip the image.

We then add English as a baseline language. Instead of using captions from XM3600, we query GPT-4o-mini to generate an English caption that explicitly includes the selected target word. To further expand the language range, we translate the English captions into additional languages using GPT-4o-mini: Chinese (Simplified), Japanese, Arabic, Hindi, Russian, and Portuguese. We choose to translate from English rather than directly use XM3600 captions for the newly added languages in order to ensure that the target word is consistently preserved in the sentence, and to limit the complexity of finding instances where all languages mention a noun; prior work shows that speakers of different languages will not always mention the same content for the same image (Kádár et al., 2018). We intentionally select languages with diverse writing systems in order to evaluate reasoning performance under varied character systems and writing direction (e.g., Arabic is from "right to left" not "left to right").

E Human Evaluation Details

To understand how current MLLMs perform on the structured reasoning steps, we query Qwen2.5-VL (Bai et al., 2025b) to generate responses following the predefined four-step reasoning structure, followed by the final answer. Unlike the intermediate reasoning steps in mathematical problems, the quality of a reasoning step in our MOSAIC task is not straightforward to define. Therefore, we first define what constitutes correct and undesired behavior for each reasoning step. Table 9 summarizes the corresponding model instructions and evaluation criteria. We formulate step-specific scoring

Table 9: Step-wise scoring criteria. Specific questions are provided in Sec. L.

Step	Instruction	Expected	Undesired
Sentence Analysis	Explain meaning in English; list mentioned object nouns; keep `[blank]`.	Correct translation; no extra inference.	Object hallucination; filling hidden target word; removing `[blank]`.
Image Analysis	Describe main objects and scene.	Accurate and image-grounded description.	Unsupported inference; hallucination of non-visible objects.
Object Exclusion	Remove mentioned and irrelevant objects; list remaining candidates.	Target preserved; mentioned & irrelevant objects excluded; consistent.	Mentioned not excluded; target removed; contradiction with itself.
Grammar Inference	Infer properties via grammar cues (preposition, gender, number).	Correct rule use supported by sentence and image.	Hallucinated rules; incorrect rule use.

questions to distinguish expected from undesired reasoning behavior. To illustrate how the rubric is applied, we provide an example reasoning trace and corresponding scores in Fig. 3. On the 20 human-annotated examples, the automatic step judge agrees with the human majority score on 75.3% of rubric decisions, with agreement of 73.3%, 100%, 75.8%, and 65.0% for Sentence Analysis, Image Analysis, Object Exclusion, and Grammar Inference, respectively.

Most questions are scored on a 0–2 scale, where 2 indicates correct reasoning and 0 indicates incorrect reasoning. An intermediate score of 1 is assigned when the model demonstrates partially correct reasoning, such as minor hallucination or incomplete exclusion. For grammar inference questions, annotators may select *N/A* when the relevant grammatical component (e.g., gender, number, or preposition) does not appear in the sentence.

The annotation was conducted through a Google Form with the predefined scoring rubric (following Fig. 3, with some additional details). Six annotators (colleagues at the authors' organization who were not involved in the project) were divided into two groups of three. Each annotator independently evaluated 10 examples; annotators within a group evaluated the same examples, and the two groups evaluated different examples. This yielded 20 distinct examples, each evaluated by three annotators on 15 scoring items, for a total of $6 \times 10 \times 15 = 900$ labels.

Table 10 summarizes the results, reporting both agreement among annotators, and score given to the corresponding step (averaged over the 20 images and the three annotations per image). It shows four groups of rows, one for each of the first four steps in our reasoning task. Within each group, we show the individual rubrics (e.g., "Description" for Image Analysis). We then show two numbers. Agr is the agreement among the three annotators assigned to each example, showing how many chose the mode among the answers (3 being best). Acc shows the average score our annotators gave the model's output (2 being best). We observe relatively high agreement on most rubric criteria, with lower agreement on some Grammar Inference criteria. Agreement for the last step scoring (not shown in table) equals 2.1. Model performance (Acc) is reasonable for the first two categories (Sentence Analysis and Image Analysis) but declines for some steps within Object Exclusion and especially Grammar Inference. For example, Mentioned Objects are not excluded well. Preposition, Gender and Number inference are not used well. These results show relatively high annotator agreement on several rubric criteria, while model scores are lower for parts of Object Exclusion and Grammar Inference. This motivates targeting these reasoning components with additional training signals.

F Baseline Prompts

Here we present the prompts used in the baseline experiments.

Table 10: **Agr**eement shows how many annotators among the 3 provide the mode (most common answer). Higher is better, max is 3, score is 0 if all responses are different. **Acc**uracy shows model performance (human scores given for model output); max is 2.

Step	Criterion	Agr	Acc
Sentence Analysis	Hallucin.	2.8	1.9
	Translation	2.6	1.5
	Extra	2.7	1.5
	Avg	2.7	1.6
Image Analysis	Descrip.	2.9	1.9
	Extra	2.7	1.6
	Avg	2.8	1.8
Object Exclusion	Right	2.7	1.4
	Mentioned	2.6	1.1
	Irrelevant	2.7	1.6
	Consistency	2.6	1.7
	Avg	2.7	1.5
Grammar Inference	Preposition	2.1	1
	Gender	2.5	0
	Number	1.9	0.9
	Hallucin.	2.5	1.6
	Avg	2.3	0.9

F.1 Task Description

See Tab. 11

Table 11: Prompt describing the task.

You are an expert in the language **{language}**. You help identify objects in images based on sentences with missing words.
Guess the meaning of the word in `[blank]` based on the image and the sentence describing the image:
{sentence}
The word is an object in the image. If there are multiple blanks, they refer to the same object.
Reply with the word wrapped in `<SOLUTION></SOLUTION>` tags. Do not include any explanations.
Example: `<SOLUTION>apple</SOLUTION>`

F.2 Reasoning Steps Instruction

See Tab. 12

F.3 Self-revising Instruction

See Tab. 13

Table 12: Prompt providing instructions for the reasoning steps.

You are an expert in the language **{language}**.
Guess the meaning of the word in `[blank]` based on the image and the sentence describing the image:
{sentence}
The word is an object in the image. If there are multiple blanks, they refer to the same object.
Explain your reasoning process in English.
Did you look at other objects in the image and perform exclusion? If you excluded objects that are not the target word, state the remaining objects that could still be the target word.
Did you utilize grammar rules to infer the meaning?
You may keep the word in its original language.
Explain your reasoning in four steps: **Sentence Analysis**, **Image Analysis**, **Object Exclusion**, and **Grammar and Context**, in that order, and then give the final answer.
In **Sentence Analysis**, translate the sentence into English and list the objects in the sentence.
In **Image Analysis**, describe the image in detail.
In **Object Exclusion**, analyze the image and the sentence to identify the objects that are not the target word. Try to exclude objects that are already mentioned in the sentence and objects that are too small or near the border of the image. Also list the remaining objects after exclusion.
In **Grammar and Context**, analyze the sentence and the image to identify the grammar rules that can be used to infer properties of the target word.
First provide your reasoning between `<REASONING>` and `</REASONING>`, and then provide your final answer between `<SOLUTION>` and `</SOLUTION>`. The final answer must be a single word.

Table 13: Prompt instructions for revising the contradictory responses.

Task Description Prompt Here
Here are your reasoning steps so far:
Sentence Analysis: {sentence_analysis}
Image Analysis: {image_analysis}
Current **{step_name}** step: **{step_text}**
There are contradictions within the **{step_name}** step. Please rewrite **ONLY** the **{step_name}** step so that it is internally consistent, logically coherent, and consistent with the Sentence Analysis, Image Analysis, and the image.
Output **ONLY** the revised **{step_name}** step, starting with the heading **"{step_name}:"**.
Do not repeat the other steps and do not provide the final answer.

G GPT Evaluation Prompt

The prompt used for evaluating the answer correctness through GPT-4o-mini is provided below in Tab. 14. The "Grammar and Context" heading in the original prompt corresponds to the Grammar Inference component used throughout the paper.

Table 14: Prompt utilized in final answer evaluation.

You are a helpful linguist and visual reasoning assistant.
Look at the provided image carefully. I will give you two words.
Your task is to decide whether these two words can refer to the **same object or entity visible in the image**.
Guidelines:

- Focus primarily on what is shown in the image, not on dictionary definitions alone.
- Accept **singular/plural** or minor form differences (e.g., *egg* vs *eggs*).
- Accept **synonyms or translations** that describe the same thing (e.g., *chicken* vs *poule*).
- Accept **subset/superset relations** if both could validly describe the same visible entity (e.g., *person* vs *tourist*, *bird* vs *penguin*).
- Reject if they clearly refer to **different objects or species** (e.g., *cat* vs *dog*), **different parts** (e.g., *tree* vs *leaf*), or **different object categories** (e.g., *food* vs *plate*).

The words can be in language_list.
Respond with only **"yes"** or **"no"**.
Are **{correct_answer}** and **{model_answer}** referring to the same object in this image?

H Final Answer Extraction

To evaluate model predictions, we extract the predicted word from the generated response. The model is instructed to output its reasoning followed by the final answer enclosed between the tokens `<SOLUTION>` and `</SOLUTION>`.

We first attempt to extract the answer using regular-expression pattern matching:

```
pattern = r"<SOLUTION>(.*?)</SOLUTION>"
```

If the pattern is successfully matched, the substring between the tags is taken as the predicted answer. We additionally check that the extracted span contains only a single word, as required by the task definition.

If the tags are missing, or the pattern fails to match, or the extracted span contains multiple words, we instead query GPT-4o-mini to extract the predicted word from the model's response. The prompt used for this fallback extraction is provided below in Tab. 15.

Table 15: Prompt utilized in final answer extraction.

You are a helpful linguist and visual reasoning assistant.
Look at the provided image carefully.
A model was asked to guess a word that fills `[blank]` in this **{language}** sentence:
{sentence}
The model's full response was:
{full_response}
Your task:
Identify what single word the model was trying to guess as the answer.

Table 16: Threshold sensitivity analysis for MetaCLIP-based final answer evaluation based on the human annotations. FP: False positive. FN: False negative.

Threshold	Agreement (%)	FP	FN
0.70	60.0	5	1
0.75	66.7	3	2
0.80	66.7	1	4
0.85	73.3	0	4
0.90	66.7	0	5

I Alternative Evaluation Method

I.1 GPT-based Evaluation - Direct Correctness Prompt

We initially prompted GPT to directly answer whether the prediction was correct with a binary yes/no output and provided explicit conditions for correctness. However, through manual inspection, we observed a large number of false negatives. For example, when the ground-truth word was "tree" and the model predicted "trunk", GPT frequently labeled the answer as incorrect. We hypothesize that this is due to GPT interpreting the semantic meaning of the words too literally. From a lexical perspective, "tree" and "trunk" do not represent the same concept and are not in a hypernym-hyponym relation. However, in the context of the image, both words may refer to the same visible entity. When the model predicts "trunk" while the ground-truth label is "tree", the prediction may still correctly identify the relevant object in the image. For the purpose of this task, such predictions should reasonably be considered correct. Thus, we adopt the reference prompt shown in Tab. 14 for judging answer correctness.

I.2 MetaCLIP-based Evaluation

In addition to GPT-based evaluation, we explored an alternative approach for assessing the correctness of the predicted answer using semantic similarity. Specifically, we compute the cosine similarity between the text embeddings of the target word and the predicted word using MetaCLIP. Since the cosine similarity score is continuous while our final evaluation is binary, a threshold is required to determine whether a prediction should be considered correct. For example, a threshold of 0.70 is indicating that when the similarity between the model answer and target word is greater or equal to 0.70, we consider this answer as correct. In practice, we observed that model performance is highly sensitive to the choice of this threshold, making the evaluation unstable and less reliable. As shown in Tab. 16, the highest agreement with human annotations reaches 73.3%, which is lower than the 80% agreement obtained from the GPT-based accuracy evaluation. We further observe that slightly lower thresholds lead to more false positives, while slightly higher thresholds lead to more false negatives, with the optimal trade-off strongly depending on the specific data.

In addition, the reasoning process is performed in English, which occasionally leads the model to output the final answer in English even when the caption is in another language. In such cases, the cosine similarity between the predicted word and the target word (in different languages) is often significantly lower than when both words are in the same language, resulting in many false negatives.

J Qualitative Examples

As shown in Fig. 4, the baseline model produces an incorrect prediction. For the GRPO model trained with global rewards, several issues appear in the reasoning steps.

In the Sentence Analysis step, the model directly fills in the missing word instead of preserving the blank as instructed. The Image Analysis step correctly describes the scene, but the Object Exclusion step does not follow the intended exclusion procedure and is further influenced by the incorrect filling in of target word introduced in the Sentence Analysis step. In the Grammar Inference step, the model does not utilize the grammatical cues provided in the sentence (e.g. reason that "of" in the sentence is indicating the trunk is one part of the target word) and instead predicts "river" without a clear justification, leading to the incorrect final answer. We also note that the choice of final answer "river"

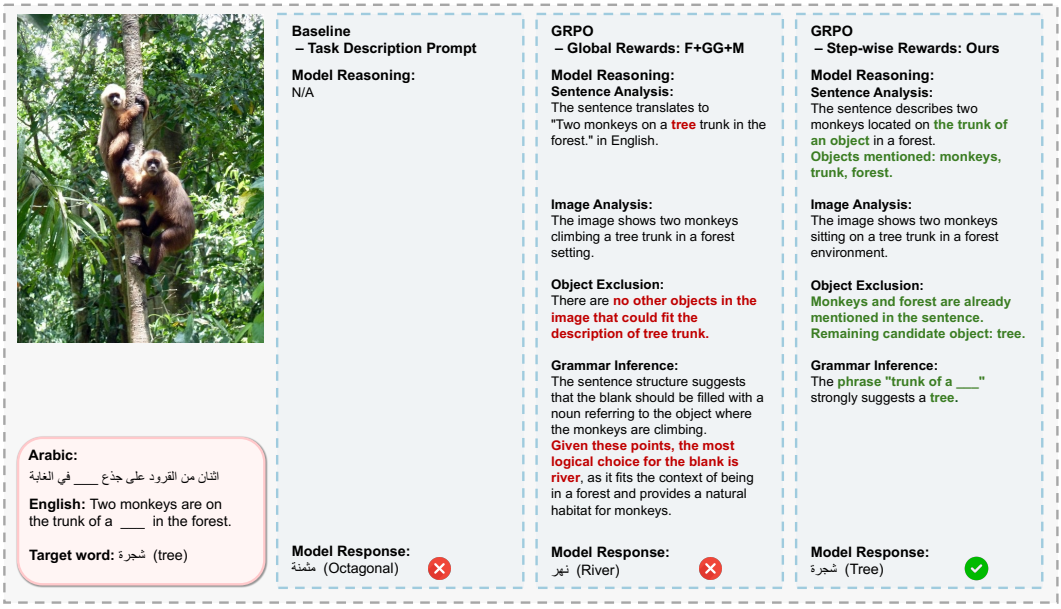


Figure 4: Qualitative example of an Arabic sentence.

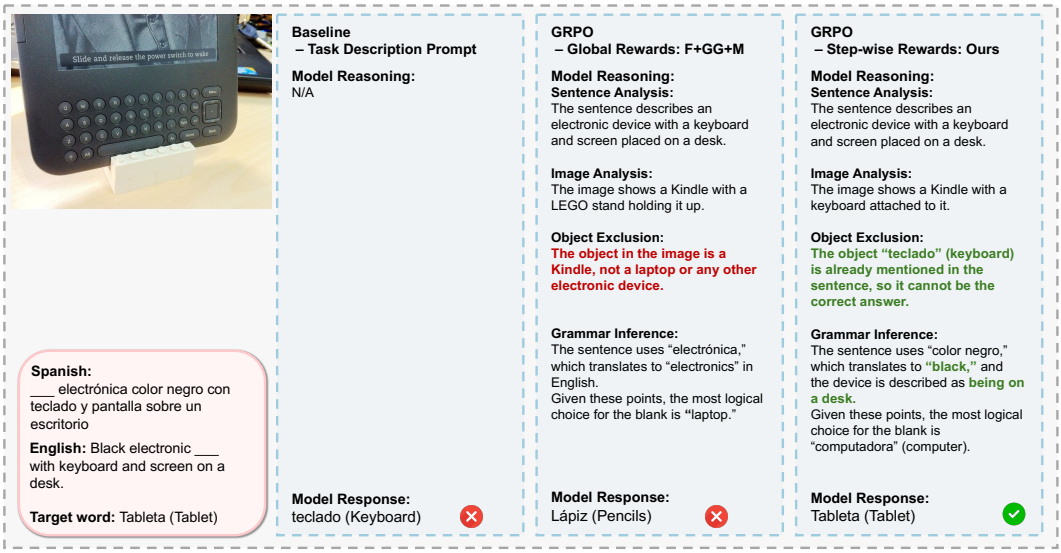


Figure 5: Qualitative example of a Spanish sentence.

shows that the model did not utilize the information it concluded in Image Analysis as there is no river both in the image and in the Image Analysis reasoning step. This example illustrates a failure to maintain consistency across the reasoning steps.

In contrast, our model trained with step-wise rewards produces more structured reasoning. The Sentence Analysis step correctly preserves the blank and lists the objects mentioned in the sentence, and the Image Analysis step accurately describes the scene. The Object Exclusion step properly removes objects already mentioned in the sentence and identifies the remaining candidate object. Although the Grammar Inference step does not explicitly apply the predefined rules for the word "of", it mentions its existence and the reasoning remains consistent with the image and sentence context, resulting in the correct prediction.

Error category	Higher-resource seen ($n = 16$)	Lower-resource seen ($n = 28$)	Unseen ($n = 32$)
Visual misidentification	2 (12.5%)	1 (3.6%)	1 (3.1%)
Sentence/context misunderstanding	6 (37.5%)	4 (14.3%)	5 (15.6%)
Exclusion/candidate-selection failure	5 (31.3%)	2 (7.1%)	3 (9.4%)
Reasoning–answer inconsistency	3 (18.8%)	18 (64.3%)	23 (71.9%)
Language/output error	0 (0.0%)	3 (10.7%)	0 (0.0%)

Table 17: Distribution of primary error categories among retained MOSAIC examples.

We show another qualitative example in Fig. 5, where the baseline model simply repeats a word already present in the sentence.

The GRPO model trained with global rewards performs well in the Sentence Analysis and Image Analysis steps, but the Object Exclusion step does not follow the intended exclusion procedure and provides little useful information for identifying the target word. In the Grammar Inference step, the model again fails to apply explicit grammar rules, which in this case is tied to the preposition "on". Although it infers that the object should be an electronic device and even reaches a reasonable candidate ("laptop"), the final answer being put between the solution tags in the response is inconsistent with this reasoning and is incorrect.

In contrast, in this example, our model produces more accurate structured reasoning. The Sentence Analysis step correctly interprets the sentence, and the Image Analysis step is largely accurate despite minor inaccuracies. The Object Exclusion step successfully removes objects already mentioned in the sentence. The Grammar Inference step applies the positional cue provided by the preposition "on" to infer that the target object is on top of the desk as well as identifying an adjective word "black" that indicates the color of the target object, resulting in the correct prediction.

K MOSAIC Error Analysis

We conduct a qualitative error analysis of MOSAIC failures across three language groups from the reasoning traces generated by our framework's final checkpoint. From examples with negative final-answer judgments, we randomly sample 50 examples from higher-resource seen languages (English, French, German, Spanish, Portuguese, Russian, Chinese, Japanese, and Korean), 50 from lower-resource seen languages (Hindi, Turkish, and Arabic), and 50 from languages unseen during RL training (Thai, Bengali, and Swahili).

For the sampled examples, we try to assign one primary error category:

- **Visual misidentification**: the model incorrectly identifies a visible object or its visual category.
- **Sentence/context misunderstanding**: the reasoning misinterprets the role of the blank, a grammatical constraint, or a relevant spatial relation.
- **Exclusion/candidate-selection failure**: the model retains candidates that should be ruled out, including objects already mentioned in the sentence.
- **Reasoning–answer inconsistency**: the reasoning identifies a supported target, but the final answer does not express that target.
- **Language/output error**: the completion contains a word-form or output defect.

After excluding examples that we cannot categorize cleanly, 16 higher-resource seen, 28 lower-resource seen, and 32 unseen-language examples remain for categorical analysis. Table 17 reports the distribution of primary error categories within these retained examples.

Within the retained higher-resource seen examples, sentence/context misunderstanding (37.5%) and exclusion/candidate-selection failure (31.3%) are the most frequent categories. Within the retained lower-resource seen and unseen examples, reasoning–answer inconsistency accounts for 64.3% and 71.9% of classified errors, respectively. Exclusion/candidate-selection failures are relatively uncommon among the retained lower-resource seen and unseen examples, accounting for 7.1% and 9.4% of classified errors, respectively. In reasoning–answer inconsistency cases, the reasoning trace identifies a supported target, but the generated final answer does not reflect that conclusion. Visual

misidentification is also relatively infrequent in all three retained samples. These patterns indicate that reasoning–answer inconsistency is a prominent failure mode in the sampled lower-resource seen and unseen examples.

L Reasoning Step Rubric Questions in Google Form

Table 18: Human annotation rubric used in the reasoning step evaluation form.

Step	Rubric questions and scoring criteria
Sentence Analysis	**Hallucination of Object.** 2: only mentions objects/concepts clearly stated in the sentence. 1: mentions an object/concept not clearly stated in the sentence but visible in the image. 0: mentions an object/scene not present in the image, or explicitly names the hidden target word. **Translation Correctness.** 2: translation is correct. 0: translation contains an error. **Extra Inference.** 2: no extra inference beyond translation and object listing. 0: extra information is provided.
Image Analysis	**Image Description Accuracy.** 2: accurately describes the visible image, including main objects and scene. 0: inaccurate description or object misidentification. **Extra Inference.** 2: no extra inference beyond describing the image. 0: extra information is provided.
Object Exclusion	The model should exclude objects that cannot be the target word, including objects already mentioned in the sentence and small/peripheral irrelevant objects, while preserving plausible candidates. **Preserved Right Answer.** 2: target word is preserved. 0: target word is not preserved. **Excluded Mentioned Objects.** 2: objects mentioned in the sentence are excluded. 0: they are not excluded. **Excluded Irrelevant Objects.** 2: irrelevant small/peripheral objects are excluded. 0: they are not excluded or are excluded incorrectly. **Consistency.** 2: no inconsistency. 0: inconsistent response.
Grammar Inference	The model should use grammar rules in the original language, not only the English translation. **Preposition Rules.** 2: correctly applied. 0: incorrectly applied or omitted when needed. N/A: not present. **Gender Inference.** 2: correctly applied. 0: incorrectly applied or omitted when needed. N/A: not present. **Number Inference.** 2: correctly applied. 0: incorrectly applied or omitted when needed. N/A: not present. **Grammar Hallucination.** 2: no hallucinated grammar rules. 0: hallucinated grammar rules.
Final Answer	**Translation Issue Check.** A: reasoning correct and final answer correct. B: reasoning correct but final answer incorrect. C: reasoning incorrect but final answer correct. D: reasoning incorrect and final answer incorrect. **Answer Form Correctness.** Correct; Incorrect singular/plural; Incorrect masculine/feminine; Incorrect other.

We include the rubric questions used in the Google Form for human annotations. The form asked annotators to score each intermediate reasoning step independently using the criteria below. The rubric defines evaluation criteria for all four reasoning steps. For training, the Object Exclusion and Grammar Inference criteria are used to compute the step-wise rewards, while all four sets of criteria are used for reasoning-trace evaluation.

Table 19: Sampling and gradient accumulation settings across models and training stages. We adjust the number of sampled responses and gradient accumulation to maintain a consistent effective batch size across models.

Model	Stage 1		Stage 2	
	Responses	Grad Acc	Responses	Grad Acc
Qwen2.5-VL 3B	4	8	4	8
Qwen2.5-VL 7B	2	16	16	2

Table 20: Training configurations for additional model and task settings.

	InternVL3.5-8B		SpatiaLQA (Qwen3-VL-8B)	
Setting	**Stage 1**	**Stage 2**	**Stage 1**	**Stage 2**
Updates to evaluated model	2,880	1,560	1,080	2,000
Learning rate	10^{-5}	10^{-5}	5×10^{-6}	2×10^{-6}
LoRA rank / alpha	32 / 32	32 / 32	16 / 32	16 / 32
Responses per input	8	8	4	4
Gradient accumulation	4	4	4	4
Maximum prompt tokens	2,048	2,048	6,144	6,144
Maximum completion tokens	768	768	1,024	1,024
Optimizer	8-bit AdamW		AdamW	
Training reward judge	Frozen Qwen3-VL-8B-Instruct			

Annotation scale. Unless otherwise specified, questions are scored using discrete ratings: 2 indicates correct behavior, 0 indicates incorrect behavior, and 1 is used only for minor hallucination in Sentence Analysis. For grammar questions, annotators could also select N/A when the relevant grammatical rule was not present.

M Training Details

MOSAIC rewards. For Qwen2.5-VL-7B, Stage 1 uses

$$r^{(1)} = r_{\text{fmt}} + 2r_{\text{acc-bin}} + 2r_{\text{acc-cont}}, \tag{5}$$

where $r_{\text{acc-bin}} \in \{0, 1\}$ denotes judge-assessed final-answer correctness and $r_{\text{acc-cont}}$ is the MetaCLIP cosine similarity.

Stage 2 drops both final-answer rewards and uses

$$r^{(2)} = r_{\text{fmt}} + 2.8r_{\text{oe}} + 2.8r_{\text{gi}}, \tag{6}$$

where r_{oe} and r_{gi} are the Object Exclusion and Grammar Inference rubric scores, each divided by its applicable maximum. The coefficient 2.8 combines a normalization scale of 2 and a step weight of 1.4.

We use parameter-efficient fine-tuning with LoRA and DrGRPO (Liu et al., 2025a). We sample multiple responses per input and use gradient accumulation during optimization. GPT-4o-mini is used for final word accuracy evaluation, separately from the local Qwen judges used for InternVL training.

Hyperparameters for different models. The details are shown in Table 19.

Experiment training settings. Tab. 20 summarizes the training configurations for InternVL3.5-8B on MOSAIC and Qwen3-VL-8B on SpatiaLQA. Both settings use LoRA-based fine-tuning with cosine learning-rate decay, 10% warmup, weight decay of 0.1, gradient clipping at 0.1, sampling temperature 0.7, and no LoRA dropout. Stage 2 is initialized from the corresponding Stage 1 checkpoint. A separate frozen Qwen3-VL-8B-Instruct model provides training-time reward judgments, while

GPT-4o-mini is used for final evaluation. Additional model- and stage-specific hyperparameters are reported in Tab. 20.

Continued-Stage-1 control. To control for additional training, we continue Stage 1 for the same number of additional updates while retaining the Stage 1 reward composition. Stage 2 outperforms this control on both seen and unseen languages, as shown in Tab. 21.

Table 21: Stage 2 versus continued Stage 1 with matched additional training updates.

Evaluation	Paired gain (pp)	One-sided p
Seen languages	**+1.13**	**0.0245**
Unseen languages	**+3.19**	**0.0435**

InternVL rewards. Stage 1 sums a tag-format reward (maximum 2), a four-heading format reward (maximum 2), binary answer accuracy scaled by 2, and MetaCLIP cosine similarity scaled by 1.5. A reasoning-length penalty increases linearly from 0 at 180 words to 1 at 300 words and is capped at 1. Stage 2 retains the tag-format reward and replaces the other components with Object Exclusion and Grammar/Context rewards. Each step score is normalized by its applicable rubric maximum and multiplied by $2 \times 1.4 = 2.8$.

N Spatial Reasoning Implementation Details

Model-proposed reasoning decomposition. We provide GPT-5.5 (OpenAI, 2026) with the SpatiaLQA task definition and 10 representative examples and ask it to propose intermediate reasoning steps together with corresponding evaluation criteria. We make only minor wording refinements without changing the underlying reasoning structure. GPT-5.5 is used only for this design stage and does not participate in training-time reward computation.

Evaluation Metric Specifics. SpatiaLQA evaluates action plans using Content F1 and Precondition F1, which measure action-content correctness and dependency agreement, respectively. Precondition F1 is the harmonic mean of precision and recall over prerequisite edges matched to the reference plan after action matching. It therefore depends on both action matching and agreement with the reference dependency structure, which may penalize alternative valid plans with different action decompositions. Thus, we focus on Content F1 in the main text and report Precondition F1 here. Average Precondition F1 improves from 0.0359 for the baseline to 0.0548 after Stage 1 and 0.0647 after Stage 2.

Stage 1 rewards. We evaluate the predicted action-plan JSON using a frozen Qwen3-VL-8B-Instruct semantic matcher and Hungarian one-to-one matching. Content F1 measures matched action content; Precondition F1 measures matched prerequisite edges after predicted action IDs are mapped to reference IDs. For SpatiaLQA, Content F1 and Precondition F1 instantiate the graded task-level reward in place of the MetaCLIP similarity used for MOSAIC. No separate binary whole-plan reward is used. The Stage 1 reward is

$$r^{(1)} = 0.5 f_{\text{tag}} + f_{\text{heading}} + f_{\text{schema}} + 4F_{\text{content}} + 2F_{\text{precondition}}, \tag{7}$$

where the three format scores are normalized to $[0, 1]$ and check the reasoning/answer tags, the five ordered reasoning headings, and the validity of the action-plan JSON, respectively.

Stage 2 rewards. Stage 2 uses

$$r^{(2)} = 0.25 f_{\text{tag}} + 0.5 f_{\text{heading}} + \sum_{k=1}^{5} s_k, \tag{8}$$

where s_k is the sum of binary criteria for reasoning step k. Stage 2 uses no final-answer accuracy or answer-schema reward. Training settings are listed in Tab. 20.

Step evaluation. We summarize the step-evaluation rubric for SpatiaLQA in Tab. 22. The model generates five reasoning sections followed by an action-plan JSON specifying atomic actions and their prerequisites. A separate frozen Qwen3-VL-8B-Instruct model serves as the judge for both the Stage 1 global rewards and the Stage 2 step-wise rewards. During Stage 2, the judge evaluates each reasoning section using the image, task instruction, previous steps, and reference action plan. The predicted final-answer JSON is not provided to the step judge. The reference action plan is provided only to the frozen judge for assessing the quality of the intermediate reasoning step; it is not included in the policy model input. The policy model generates both its reasoning and final action plan from the task input alone and receives the resulting step score only as a training reward.

Scoring. Each criterion receives 1 if satisfied and 0 otherwise. Criteria are scored independently and summed within each step, yielding scores of 0–3 for the first three steps and 0–4 for the last two. Missing, empty, unrelated, or uninterpretable sections receive 0 for all criteria.

Table 22: Step-evaluation criteria for SpatiaLQA.

Step	Scoring criteria
Target-object Identification	**Target correctness.** Identifies the intended object, destination, or target state. **Role distinction.** Does not confuse the target with a blocker, support, tool, cover, or incidental object. **Grounding.** Does not invent targets unsupported by the image, task, or reference plan.
Blocking-object Identification	**Required objects.** Identifies blockers, covers, supports, or access objects that must be handled. **Relevance.** Excludes visible objects unrelated to completing the task. **No hallucination.** Does not invent blockers unsupported by the image, task, or reference plan.
Spatial-relation Identification	**Relevant relations.** States spatial relations needed for the plan. **Relation correctness.** Uses specific relations such as on, inside, under, or blocking correctly. **Consistency.** Avoids relations that contradict the image, task, or reference plan.
Dependency Reasoning	**Prerequisites.** Identifies necessary preceding actions, or explicitly states that none are needed. **Causal ordering.** Explains why actions must occur in a particular order. **Necessary dependencies.** Avoids irrelevant prerequisites. **Valid dependencies.** Avoids circular, forward-looking, or physically or temporally impossible dependencies.
Plan-generation Reasoning	**Coverage.** Includes the actions needed to complete the task. **Atomicity.** Keeps distinct actions separate. **Relevant actions.** Avoids unnecessary or hallucinated actions. **Task completion.** The plan would satisfy the task if executed in order.

Computational cost. Table 23 reports performance together with training cost and average inference-time generation length on SpatiaLQA. Stage 1 improves Content F1 from 40.20 to 42.65 with 6.27 H200 GPU-hours of training. After Stage 2, Content F1 reaches 44.70 and the cumulative training cost increases to 25.81 H200 GPU-hours. The additional computational cost of Stage 2 is primarily a one-time training cost, since the frozen reward judge is not used after training. At inference, Stage 2 introduces little additional generation-length overhead relative to Stage 1, with the average response length increasing only from 272.2 to 275.4 tokens. For comparison, the baseline generates 65.1 tokens on average.

Method	Content F1 (%)	Avg. gen. tokens	Cumulative train cost (H200 GPU-hours)
Baseline	40.20	65.1	0.00
Stage 1	42.65	272.2	6.27
Stage 1 + Stage 2	44.70	275.4	25.81

Table 23: Performance, average generated response length, and cumulative training cost on SpatiaLQA. The reward judge is used only during training and is not required at inference.

Table 24: Prompt-based baselines (Qwen2.5-VL-7B), on 12 languages ordered by zero-shot accuracy.

Method	Avg	EN	FR	ZH	PT	DE	ES	JA	RU	KO	AR	TR	HI
Zero-shot Task Instruction	**.811**	**.925**	**.900**	**.865**	**.875**	**.894**	**.870**	**.885**	**.865**	**.758**	.670	.685	.540
Zero-shot Step-wise Instruction	.722	.789	.804	.772	.809	.833	.735	.700	.778	.647	.611	.594	.593
1-shot Step-wise Instruction	.774	.834	.869	.822	.814	.884	.825	.750	.843	.707	.652	**.701**	.583
3-shot Step-wise Instruction	.785	.834	.824	.833	.859	.849	.830	.800	.828	.727	**.702**	.695	**.638**
Self-revision Baseline	.687	.769	.789	.701	.734	.818	.750	.660	.677	.616	.576	.584	.568

O PROMPT-BASED BASELINE RESULTS

We first test whether prompting alone can effectively guide the ambiguity resolution reasoning. The vanilla zero-shot task-instruction baseline directly asks the model to infer the missing word without requesting intermediate reasoning. We tested several prompt-based variants which give more guidance on the task. However, as shown in Tab. 24, they all perform worse than the vanilla baseline. Zero-shot step-wise prompting drops average accuracy from 0.811 to 0.722. Handcrafted 1-shot and 3-shot demonstrations partially recover this drop, reaching 0.774 and 0.785, respectively, but still fall below the vanilla task-instruction baseline overall. Although they provide some gains for lower-resource languages, these improvements remain limited, highlighting the difficulty of the proposed reasoning task.

We also evaluate a training-free self-revision baseline, where the model checks consistency across its own reasoning steps using proposition-level contradiction detection (Mündler et al., 2024), revising as needed. This baseline performs worst, reaching 0.687 average accuracy. Manual inspection shows the model often misses contradictions, such as excluding an object and later listing it as a candidate. Overall, explicit reasoning prompts do not reliably improve performance on this task, motivating our staged GRPO.

P LIMITATIONS

In this paper, we study step-wise GRPO for multimodal reasoning beyond mathematically verifiable settings. We evaluate the framework on our multilingual MOSAIC testbed and SpatiaLQA, but further work is needed to understand which types of reasoning structure are most suitable for step-wise semantic rewards and how broadly the approach transfers across tasks. A second limitation is the use of automatic judges for intermediate reasoning steps. These judges cannot guarantee correct evaluation of every reasoning trace, particularly when the criteria are semantically defined rather than uniquely verifiable. The improvements in downstream task performance suggest that these imperfect scores nevertheless provide useful learning signals, but they should not be interpreted as evidence that every generated reasoning step is correct. Finally, step-wise reward evaluation introduces additional training computation because intermediate steps must be scored during Stage 2. Future work could explore more efficient judge models, selective evaluation of only the most informative reasoning steps, or other ways to reduce this overhead.