LENGUARD-GPC: LENGTH GUARDING WITH GUIDED-PROMPT CONSISTENCY FOR SPATIAL REASONING REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view spatial reasoning requires vision-language models to compare visual evidence across images, align object correspondences, and infer spatial relations over long visual contexts, a setting where chain-of-thought reasoning tends to grow verbose without becoming more accurate. Reinforcement learning with verifiable rewards is a natural fit for this task, but standard GRPO reward relies on sparse outcome-level feedback and gives no signal about where a reasoning trajectory goes wrong, nor any control over its length. We propose LENGUARD-GPC, a dense reward framework that addresses both problems together. For each sampled trajectory, it compares the token-wise predictive distributions under a standard prompt and a guided prompt, and uses the resulting token-sum KL divergence as a dense reward signal. Since this KL penalty accumulates over tokens and would otherwise reward shorter responses regardless of their quality, we introduce a staged length bonus that keeps reasoning length within a controlled range without simply encouraging brevity.

1 INTRODUCTION

Vision-language models (VLMs) have recently made remarkable progress in bridging visual perception and language-based reasoning. By scaling multimodal pretraining and instruction tuning, models such as LLaVA (Liu et al., 2023b; 2024; Li et al., 2024b), Flamingo (Alayrac et al., 2022), Gemini (Team et al., 2023), and Qwen-VL (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025a) have achieved strong performance on a wide range of vision-language tasks such as visual question answering, image captioning and visual grounding (Kan et al., 2025; Sarto et al., 2025; Wang et al., 2025b; Wu et al., 2025d; Tao et al., 2026a; Gou et al., 2025; Tao et al., 2025), where a single image usually contains enough evidence to support an answer. Multi-view spatial reasoning is a different kind of problem: given several images of the same scene taken from different viewpoints, a model has to relate objects across images before it can answer a question about their spatial relations.

Consider two images of the same room, one from the front and one from the side. A model must recognize that a sofa near the wall in one image and a sofa near the door in the other are the same object before it can answer a question such as which object is closer to the window, a step that standard vision-language benchmarks never require and that even strong VLMs tend to fail at. When prompted to reason step by step over several such images, models typically describe what they see in each image in turn, producing traces that grow longer without ever explicitly resolving which object in one image corresponds to which in another, the step the question actually depends on. Longer traces can also compound errors: once several claims accumulate across views, an early misreading tends to persist uncorrected, and irrelevant detail crowds out the one relation that matters. Reasoning quality here is therefore not a function of length; what matters is whether the trace resolves cross-view correspondence at all.

Reinforcement learning with verifiable rewards offers a promising direction for improving reasoning behavior, since spatial reasoning tasks often admit objective answer-level supervision. Methods such as GRPO can optimize models using outcome rewards without requiring dense human annotations. However, directly applying reward-only optimization to multi-view spatial reasoning remains insufficient. A final correctness reward provides limited information about where the reasoning process goes

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.

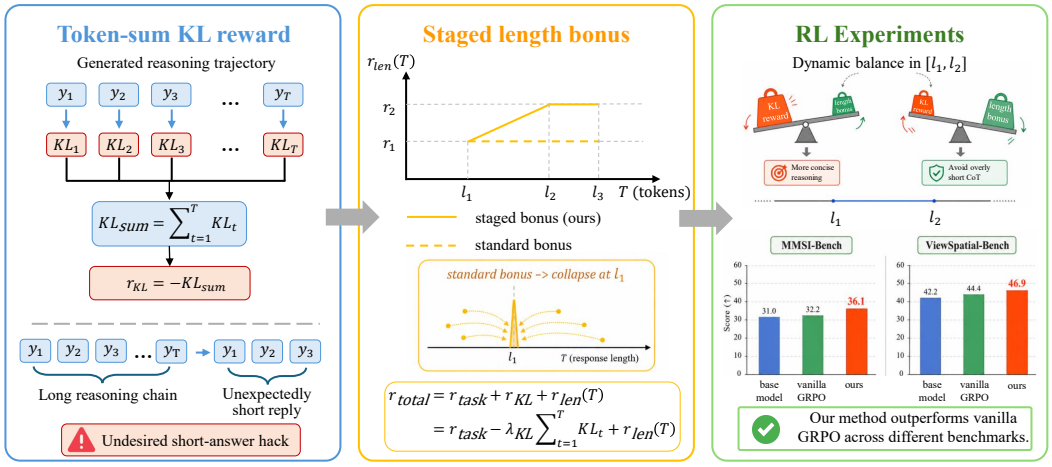

Figure 1: Overview of LENGUARD-GPC for multi-view spatial reasoning. Within the staged interval $[l_1, l_2]$, LENGUARD-GPC adaptively balances the negative KL-based consistency reward and the length bonus. When the mean per-token KL is high, the consistency reward dominates and encourages more concise reasoning. When the mean per-token KL is low, the length bonus becomes relatively stronger, preventing the model from collapsing into overly short and content-poor chains of thought.

wrong, especially when the failure is caused by subtle local errors such as mismatched viewpoints, incorrect object correspondence, or invalid spatial relation inference. Moreover, outcome-only reward optimization may encourage undesirable reasoning patterns, including unnecessarily long rationales, repetitive visual descriptions, or superficial answer guessing. Therefore, an effective RL objective for multi-view spatial reasoning should provide denser process-level guidance while also controlling the length and structure of the reasoning trajectory.

To address these challenges, we propose LENGUARD-GPC, a length-guarded guided-prompt consistency framework for GRPO-based spatial reasoning. The guided-prompt design is inspired by Self On-Policy Distillation (Hübotter et al., 2026), which compares model behavior under different prompting conditions. The central idea is to use a guided prompt as a reference condition for evaluating the model's sampled trajectory. Given a generated response, we compute the token-wise predictive distributions under two conditions: the original prompt and a guided prompt that provides additional reasoning guidance. We then measure the KL divergence between these two distributions along the trajectory. This guided-prompt consistency score reflects the consistency between the model's current reasoning behavior and the behavior encouraged by the guided prompt. Rather than introducing an auxiliary supervised objective, LENGUARD-GPC directly incorporates the trajectory-level KL score into the reward function used by GRPO. In this way, the model receives denser feedback than a final correctness reward while preserving the simplicity and flexibility of reinforcement learning optimization.

However, using token-sum KL as a reward component introduces a length-related challenge. Since KL scores accumulate over tokens, longer reasoning chains naturally produce larger cumulative values. This may create an incentive for the model to generate overly short responses, even when a moderate amount of reasoning is necessary. A simple flat length bonus can partially compensate for this bias, but it may cause response lengths to concentrate around the earliest rewarded boundary. To avoid both short-answer collapse and boundary-seeking behavior, LENGUARD-GPC introduces a staged length bonus controlled by length boundaries (l_1, l_2, l_3) and bonus levels (r_1, r_2). The bonus is not applied before l_1, increases from r_1 to r_2 within the interval $[l_1, l_2]$, and remains constant until l_3. This schedule creates a dynamic balance between KL-guided compactness and length-based sufficiency: when the mean per-token KL is high, the consistency reward encourages more concise reasoning; when the mean per-token KL is low, the length bonus prevents the model from collapsing into overly short and content-poor chains of thought.

Although LENGUARD-GPC is developed for multi-view spatial reasoning, its principle can extend to broader long-context multimodal reasoning tasks, such as video understanding, multi-image

comparison, document-grounded visual reasoning, and embodied perception. These tasks also require models to reason over dense multimodal inputs, where sparse outcome rewards may be insufficient and unconstrained chain-of-thought generation can become verbose or poorly grounded. By requiring only a guided prompt and token-level distributional comparison, LENGUARD-GPC offers a general reward-based framework for improving reasoning behavior in multimodal reinforcement learning. Our contributions are summarized as follows:

1. We identify a key challenge in GRPO-based multi-view spatial reasoning: sparse outcome rewards provide limited guidance for local cross-view reasoning errors.

2. We propose LENGUARD-GPC, which leverages guided-prompt token-wise KL consistency scores as a GRPO-compatible reward signal, enabling dense process-level feedback without an additional supervised loss.

3. We introduce a staged length-guard mechanism that mitigates both short-answer collapse and boundary-seeking behavior, and demonstrate LENGUARD-GPC as a general RL framework for utilizing guided-prompt consistency rewards in long-context multimodal reasoning.

2 RELATED WORK

2.1 REINFORCEMENT LEARNING FOR MULTIMODAL LARGE LANGUAGE MODEL REASONING

Reinforcement learning (RL) and preference optimization have been increasingly used to improve the reasoning quality and behavioral alignment of multimodal large language models (MLLMs) (Sun et al., 2024; Yu et al., 2024a;b; Xie et al., 2024; Huang et al., 2025). Recent reasoning-oriented RL methods optimize models with verifiable rewards or preference signals to encourage stronger multimodal deliberation. For example, Insight-V (Dong et al., 2025) learns from self-generated reasoning trajectories selected by a multi-agent framework, while R1-VL (Zhang et al., 2025a) adopts step-wise GRPO with dense rule-based rewards. GRPO-style training has also been extended to video reasoning, including Video-R1 (Feng et al., 2025), Video-RTS (Wang et al., 2025d), and Video-STR (Wang et al., 2025c). In parallel, several works encourage models to "observe first" through explicit context descriptions or observation stages, such as HumanOmniV2 (Yang et al., 2025a), Visionary-R1 (Xia et al., 2025), ViewFusion (Tao et al., 2026b) and Observe-R1 (Guo et al., 2025). Existing RL-based MLLM reasoning methods largely overlook reasoning length. In long-context multimodal tasks, overly long chains can become redundant, while overly short responses may be insufficient. We address this with a guided-prompt consistency reward and staged length guard to balance compactness and reasoning sufficiency.

2.2 SPATIAL REASONING WITH MLLMS

Spatial reasoning has emerged as a key frontier for MLLMs, aiming to move beyond object recognition toward understanding relative position, orientation, viewpoint transformation, and occlusion-aware relations in 3D scenes. A growing body of work seeks to strengthen spatial reasoning in MLLMs by improving grounding and spatial representations, for example via spatially aware instruction tuning and curated supervision (Liu et al., 2023a; Tong et al., 2024; Chen et al., 2024; Yu et al., 2025; Gholami et al., 2025; Wu et al., 2025b; Zhao et al., 2025; Batra et al., 2025; Wang et al., 2025a; Fan et al., 2025; Li et al., 2024a). More recently, Visual Spatial Tuning (Yang et al., 2025b) trains vision-language models with large-scale spatial perception and reasoning data, producing notably stronger spatial reasoning performance and improved generalization across spatial benchmarks. To systematically evaluate these advances under multi-view inputs, several benchmarks have been proposed (Zhang et al., 2025b; Lee et al., 2025). MMSI-Bench (Yang et al., 2025e) focuses on multi-view, multi-image spatial intelligence and includes problems that require aligning evidence across views rather than solving from a single snapshot. ViewSpatial (Li et al., 2025a) further stresses viewpoint-dependent spatial localization and cross-view reference frames, revealing substantial generalization gaps when camera viewpoints shift.

3 LENGUARD-GPC

3.1 REASONING-LENGTH CHALLENGE IN MULTI-VIEW SPATIAL REASONING

Chain-of-thought prompting is widely used to elicit intermediate reasoning in language and multimodal models (Wei et al., 2022; Zhang et al., 2023). However, in multi-view spatial reasoning, long reasoning traces often become verbose and inefficient, repeatedly describing irrelevant details or comparing similar regions without advancing the target inference, which distracts from key cross-view evidence. Recent studies suggest that excessive chain-of-thought length can hurt both accuracy and computational efficiency, while concise reasoning may achieve comparable or better performance (Nayab et al., 2024; Xu et al., 2025; Hassid et al., 2025). Thus, a key challenge is to regulate reasoning length and encourage concise, sufficient, and visually grounded reasoning.

Simply encouraging shorter reasoning is insufficient. Although long-to-short methods can improve efficiency, overly aggressive reduction may reduce reasoning transparency or cause behavioral inconsistency (Yang et al., 2025c). Likewise, rewards favoring the shortest valid output may lead to answer-only behavior, format collapse, or other reward-hacking strategies rather than genuinely improved reasoning. These observations motivate our RL formulation. We introduce a guided-prompt consistency reward for dense trajectory-level feedback and combine it with a staged length guard to avoid both overly long and overly short reasoning. Inspired by Self On-policy Distillation (Hübotter et al., 2026), our method compares behavior under different prompting conditions, but uses this comparison as an RL reward rather than a supervised distillation objective.

3.2 GUIDED-PROMPT CONSISTENCY REWARD

Given a multi-view reasoning instance, we denote the visual inputs as $\mathcal{I} = \{I_1, \ldots, I_m\}$ and the question as q. The standard prompt is written as $x = (\mathcal{I}, q)$, under which the policy model π_θ samples a reasoning trajectory $y = (y_1, \ldots, y_T)$. To provide denser feedback during training, we additionally construct a guided prompt $x^{\mathrm{gp}} = (\mathcal{I}, q, h)$, where h provides task-specific guidance for evaluating the generated trajectory. In our implementation, h is a one-sentence guidance generated by GPT-5.3-Chat for each training instance. The guided prompt is used only for reward computation during training and is not required at inference time.

For the same sampled trajectory y, we compare the model's token-level predictive distributions under the standard prompt and the guided prompt. Both distributions are computed along the same generated trajectory using teacher forcing:

$$p_t^{\mathrm{std}} = \pi_\theta(\cdot \mid x, y_{<t}), \quad p_t^{\mathrm{gp}} = \pi_\theta(\cdot \mid x^{\mathrm{gp}}, y_{<t}), \quad D_t = D_{\mathrm{KL}}\big(p_t^{\mathrm{std}} \,\|\, p_t^{\mathrm{gp}}\big)\,. \tag{1}$$

Here, D_t measures how much the model's prediction under the standard prompt deviates from its prediction under the guided-prompt condition at the same reasoning state.

We aggregate the token-level KL scores over the full trajectory and define the guided-prompt consistency reward as the negative accumulated divergence:

$$S_{\mathrm{KL}}(y) = \sum_{t=1}^{T} D_t, \quad r_{\mathrm{KL}}(y) = -S_{\mathrm{KL}}(y). \tag{2}$$

In this way, trajectories that remain closer to the guided-prompt behavior receive a smaller penalty, while trajectories that deviate substantially from the guided condition receive a lower reward. Importantly, this KL score is used as an RL reward rather than an auxiliary supervised objective. Thus, the model is not forced to imitate a fixed reasoning path, but is instead encouraged to generate trajectories that are more consistent with the guided-prompt evaluation signal.

3.3 STAGED LENGTH GUARD

Although the guided-prompt consistency reward provides dense trajectory-level feedback, it also introduces a length-related bias. Since the KL score is accumulated over tokens, longer reasoning trajectories tend to receive larger KL penalties even when their reasoning process is valid. If used alone, this penalty may encourage the policy to shorten its responses excessively, resulting in insufficient or answer-only reasoning. A straightforward solution is to add a flat length bonus

after a minimum length threshold, but this can create another degenerate behavior: the model may concentrate its response length around the earliest rewarded boundary, producing just-enough but underdeveloped reasoning.

To address this issue, we introduce a staged length guard. Instead of assigning a constant length bonus, we divide the reasoning length into stages. No length bonus is assigned before the lower boundary l_1, which discourages overly short responses. Within the interval $[l_1, l_2]$, the bonus gradually increases from r_1 to r_2, preventing the policy from collapsing all responses around l_1. After l_2, the bonus is kept constant, so the model does not receive additional incentives for producing unnecessarily long reasoning chains. Formally, the staged length bonus is defined as

$$r_{\text{len}}(T) = \begin{cases} 0, & T < l_1 \text{ or } T > l_3, \\ r_1 + (r_2 - r_1)\dfrac{T - l_1}{l_2 - l_1}, & l_1 \le T \le l_2, \\ r_2, & l_2 < T \le l_3. \end{cases} \tag{3}$$

where T denotes the number of generated tokens, (l_1, l_2, l_3) are the length boundaries, and (r_1, r_2) are the lower and upper bonus levels. In practice, l_3 corresponds to the maximum effective reasoning length considered by the length guard.

The staged design balances compactness and sufficiency during RL optimization. When the mean per-token KL is high, the KL reward discourages unnecessary or misaligned reasoning steps. When the mean per-token KL is low, the length guard prevents the model from collapsing into overly short responses. As a result, the policy is encouraged to produce reasoning trajectories that are not merely short, but concise, sufficient, and better aligned with the guided-prompt evaluation signal.

3.4 GRPO Training with Guided-Prompt Consistency Reward

3.4.1 Preliminary: Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) optimizes a policy by comparing multiple responses sampled from the same input. Given a multi-view reasoning prompt x, the policy π_θ samples G trajectories $\{y_i\}_{i=1}^{G}$, where each $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ contains the reasoning process and final answer. GRPO estimates trajectory advantages through group-level reward normalization without an external value model.

For each sampled trajectory, the task reward is a weighted sum of accuracy and format rewards:

$$r_{\text{task}}(y_i) = \lambda_{\text{acc}} r_{\text{acc}}(y_i) + \lambda_{\text{format}} r_{\text{format}}(y_i). \tag{4}$$

Here, $r_{\text{acc}}(y_i)$ evaluates answer correctness, while $r_{\text{format}}(y_i)$ encourages valid output formatting.

The final trajectory reward combines the task reward, guided-prompt consistency reward, and staged length bonus:

$$r_i = r(y_i) = \lambda_{\text{task}} r_{\text{task}}(y_i) + \lambda_{\text{KL}} r_{\text{KL}}(y_i) + \lambda_{\text{len}} r_{\text{len}}(T_i). \tag{5}$$

Here, $r_{\text{KL}}(y_i)$ measures token-wise deviation between the standard and guided prompts, while $r_{\text{len}}(T_i)$ regulates reasoning length.

GRPO then normalizes the trajectory rewards within each group to obtain relative advantages:

$$A_i = \frac{r_i - \text{mean}\left(\{r_j\}_{j=1}^{G}\right)}{\text{std}\left(\{r_j\}_{j=1}^{G}\right) + \epsilon}. \tag{6}$$

where ϵ is a small constant for numerical stability. The standard GRPO objective then encourages above-average trajectories and suppresses below-average ones.

During training, responses are sampled only from the standard prompt x, while the guided prompt x^{gp} is used solely to compute the guided-prompt consistency reward. Thus, our method introduces no additional inference-time overhead.

3.4.2 Training Procedure

Algorithm 1 summarizes the training process. For each instance, we construct both the standard prompt and the guided prompt. The policy samples a group of trajectories from the standard prompt,

and each trajectory is evaluated using the accuracy reward, format reward, guided-prompt consistency reward, and staged length bonus. The combined rewards are normalized within the group and used to update the policy with GRPO.

Overall, the proposed training procedure treats guided-prompt consistency as an RL reward rather than a target sequence or an auxiliary loss. By combining task correctness, format control, guided-prompt consistency, and staged length regulation, the model is optimized to produce reasoning trajectories that are correct, compact, and sufficiently grounded.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training Data. Our training set is derived from VST-500K (Yang et al., 2025b). We sample approximately 13,000 instances from its multi-view spatial reasoning data, where each example contains multiple images captured from different viewpoints and requires the model to reason over cross-view spatial relations. Unlike single-image VQA, these samples require cross-view comparison, object correspondence, and spatial inference from partially overlapping observations. This makes the dataset suitable for studying GRPO-based multi-view reasoning and evaluating guided-prompt consistency under long, visually dense contexts.

Algorithm 1 GRPO Training with Guided-Prompt Consistency Reward

Require: Training instance $(\mathcal{I}, q, a)$, policy π_θ, group size G
1: Construct the standard prompt $x = (\mathcal{I}, q)$ and the guided prompt $x^{\mathrm{gp}} = (\mathcal{I}, q, h)$.
2: Sample G trajectories $\{y_i\}_{i=1}^{G}$ from $\pi_\theta(\cdot \mid x)$.
3: **for** each trajectory y_i **do**
4: Compute reward components r_{acc}, r_{format}, r_{KL}, and r_{len}.
5: Combine them into the total reward $r_i = r(y_i)$.
6: **end for**
7: Normalize $\{r_i\}_{i=1}^{G}$ within the group to obtain advantages $\{A_i\}_{i=1}^{G}$.
8: Update π_θ with the standard GRPO objective.

Implementation Details. We use Qwen3-VL-8B-Instruct as the base model and train the policy with GRPO. The learning rate is set to 1×10^{-6}. The task reward is composed of an accuracy reward and a format reward:

$$r_{\mathrm{task}}(y) = \lambda_{\mathrm{acc}} r_{\mathrm{acc}}(y) + \lambda_{\mathrm{format}} r_{\mathrm{format}}(y). \tag{7}$$

Here, $r_{\mathrm{acc}}(y)$ measures whether the final answer is correct, and $r_{\mathrm{format}}(y)$ measures whether the response follows the required output format. In implementation, we use a format reward of 0.2 to encourage valid response formatting.

The guided-prompt consistency reward is computed from the accumulated token-level KL score:

$$S_{\mathrm{KL}}(y) = \sum_{t=1}^{T} D_t, \tag{8}$$

and is incorporated into the final reward with coefficient $\lambda_{\mathrm{KL}} = 5 \times 10^{-3}$. Equivalently, this corresponds to a KL-score coefficient of -5×10^{-3} in implementation.

For the staged length guard, we set the length boundaries as $l_1 = 160$, $l_2 = 360$, and $l_3 = 768$. The length bonus starts from $r_1 = 0.25$ at $l_1 = 160$ tokens and increases by 2.5×10^{-4} for each additional token until it reaches $r_2 = 0.30$ at $l_2 = 360$ tokens. From l_2 to l_3, the length bonus remains fixed at 0.30. No length bonus is granted beyond $l_3 = 768$ tokens, so the staged guard does not encourage excessively long reasoning trajectories.

The final reward combines task performance, guided-prompt consistency, and staged length regulation:

$$r(y) = \lambda_{\mathrm{task}} r_{\mathrm{task}}(y) - \lambda_{\mathrm{KL}} S_{\mathrm{KL}}(y) + \lambda_{\mathrm{len}} r_{\mathrm{len}}(T). \tag{9}$$

These settings are designed to balance the negative guided-prompt consistency reward with sufficient reasoning length, preventing short-answer collapse while avoiding unnecessary verbosity.

Table 1: Overall accuracy (%) on six evaluation benchmarks. Benchmarks marked with $*$ follow the same multi-view spatial reasoning domain as our training data, while benchmarks marked with $\diamond$ are used as out-of-domain evaluations. MindCube is additionally marked with ‡ because we only use it for evaluation and do not use its provided training set.

Model	Size	MMSI*	ViewSpatial*	MindCube*,‡	3DSR$^\diamond$	BLINK$^\diamond$	CV$^\diamond$	Avg.
Random Choice	–	25.0	26.3	33.0	45.8	38.1	42.5	35.1
Closed-source models								
Gemini-2.5-Pro	–	38.0	46.0	57.6	59.3	73.5	85.9	60.1
GPT-5	–	41.8	45.5	56.3	60.3	68.0	84.6	59.4
Gemini-3-Pro-Preview	–	45.2	50.3	70.8	68.9	76.0	92.0	67.2
Open-source General Models								
InternVL3-2B (Zhu et al., 2025)	2B	26.5	32.5	37.5	47.7	50.8	76.5	45.3
InternVL3-8B (Zhu et al., 2025)	8B	28.0	38.6	41.5	44.3	53.5	81.0	47.8
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	3B	28.6	31.9	37.6	43.5	48.7	71.8	43.7
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	7B	26.8	36.8	36.0	47.5	55.9	75.4	46.4
Qwen3-VL-2B-Instruct (Bai et al., 2025a)	2B	28.9	36.9	34.5	47.5	53.2	78.4	46.6
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	4B	30.1	42.5	37.0	52.5	62.6	84.7	51.6
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	8B	31.1	42.2	29.4	53.9	66.7	85.1	51.4
Spatial Intelligence Models								
SpatialLadder-3B (Li et al., 2025b)	3B	27.4	39.8	43.4	42.8	43.0	73.7	45.0
SpaceR-7B (Ouyang et al., 2025)	7B	27.4	35.8	37.9	40.5	49.6	74.8	44.3
ViLaSR-7B (Wu et al., 2025c)	7B	30.2	35.7	35.1	46.6	51.4	76.7	46.0
Cambrian-S-3B (Yang et al., 2025d)	3B	27.0	39.0	32.5	41.4	37.7	75.2	42.1
Cambrian-S-7B (Yang et al., 2025d)	7B	27.1	40.9	39.6	45.0	37.9	76.9	44.6
VST-3B-RL (Yang et al., 2025b)	3B	32.0	45.0	36.4	56.5	57.2	84.2	51.9
VST-7B-RL (Yang et al., 2025b)	7B	34.8	42.4	39.1	60.1	62.6	86.5	54.3
Vanilla GRPO	8B	32.2	44.4	44.7	59.4	65.7	85.2	55.3
Ours								
LENGUARD-GPC	8B	**36.1**	**46.9**	**48.4**	60.1	66.8	85.5	**57.3**

Evaluation Setting. We evaluate the trained models on six benchmarks covering both in-domain and out-of-domain reasoning scenarios. Specifically, MMSI-Bench (Yang et al., 2025e), MindCube (Yin et al., 2025), and ViewSpatial-Bench (Li et al., 2025a) are multi-image spatial reasoning benchmarks, where each instance requires the model to compare visual evidence across multiple views and infer spatial relations. These benchmarks are most aligned with our training data, which also consists of multi-view spatial reasoning samples. Notably, MindCube is used only for evaluation, and we do not use its provided training set.

To further assess the generalization ability of our method and verify that the training procedure does not overfit to the in-domain data, we additionally evaluate on three out-of-domain benchmarks: 3DSR (Ma et al., 2025) and CV-Bench (Tong et al., 2024), which focus on single-image spatial reasoning, and BLINK (Fu et al., 2024), which involves multi-image reasoning but is not primarily centered on spatial relation inference. This evaluation protocol allows us to examine whether the proposed guided-prompt consistency reward improves performance on in-domain multi-view spatial reasoning tasks while maintaining robustness on related spatial and multi-image reasoning tasks beyond the training distribution.

4.2 QUANTITATIVE RESULTS

Table 1 reports the overall accuracy on six evaluation benchmarks. Compared with the base Qwen3-VL-8B-Instruct model, vanilla GRPO improves the average accuracy from 51.4% to 55.3%, showing that RL training can enhance multi-view spatial reasoning performance. Our LENGUARD-GPC further improves the average accuracy to 57.3%, yielding a gain of $+2.0$ points over vanilla GRPO and $+5.9$ points over the base model. The improvement is most pronounced on benchmarks that follow the same multi-view spatial reasoning domain as our training data. On MMSI, LENGUARD-GPC improves over vanilla GRPO from 32.2% to 36.1%, corresponding to a $+3.9$ point gain. On MindCube, which is used only for evaluation without using its provided training set, the accuracy increases from 44.7% to 48.4%. On ViewSpatial, LENGUARD-GPC further improves the accuracy

Table 2: Fine-grained accuracy (%) breakdown on MMSI-Bench. We compare LENGUARD-GPC with baselines that are included in the main results table and also have fine-grained MMSI results available.

Models	Positional Relationship						Attribute		Motion		MSR	Avg.
	Cam.–Cam.	Obj.–Obj.	Reg.–Reg.	Cam.–Obj.	Obj.–Reg.	Cam.–Reg.	Meas.	Appr.	Cam.	Obj.	–	
RandomChoice	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Closed-source models												
Gemini-2.5-Pro	38.7	34.0	40.7	44.2	38.8	41.0	62.5	30.3	39.2	25.0	33.3	38.0
GPT-5	41.9	33.0	35.8	49.8	42.4	68.7	54.7	37.4	28.3	40.8	36.4	41.8
Open-source General Models												
InternVL3-2B (Zhu et al., 2025)	31.2	22.3	28.4	30.2	28.2	28.9	25.0	22.7	16.2	28.9	26.8	26.5
InternVL3-8B (Zhu et al., 2025)	22.6	22.3	34.6	31.4	42.4	33.7	25.0	19.7	20.3	34.2	24.8	28.0
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	36.6	30.9	28.4	26.7	28.2	31.3	31.2	16.7	16.2	35.5	28.8	28.6
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	28.0	26.6	19.8	32.6	38.8	28.9	23.4	21.2	20.3	30.3	24.8	26.8
Qwen3-VL-2B-Instruct (Bai et al., 2025a)	26.9	29.8	30.9	38.4	35.3	33.7	23.4	28.8	29.7	28.9	21.2	28.9
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	28.0	37.2	32.1	31.4	35.3	38.5	37.5	15.2	27.0	28.9	29.8	31.1
Spatial Intelligence Models												
SpatialLadder-3B (Li et al., 2025b)	36.6	29.8	29.6	32.6	30.6	24.1	18.8	31.8	23.0	23.7	23.2	27.4
Spatial-MLLM-4B (Wu et al., 2025a)	24.7	21.3	28.4	30.2	29.4	28.9	18.8	34.9	10.8	23.7	29.8	26.1
SpaceR-7B (Ouyang et al., 2025)	25.8	31.9	29.6	25.6	31.8	22.9	26.6	28.8	16.2	34.2	27.3	27.4
ViLaSR-7B (Wu et al., 2025c)	29.0	35.1	28.4	39.5	40.0	44.6	31.2	16.7	17.6	31.6	23.2	30.2
Cambrian-S-3B (Yang et al., 2025d)	25.8	28.7	24.7	48.8	24.7	33.7	29.7	22.7	20.3	28.9	18.7	27.0
Cambrian-S-7B (Yang et al., 2025d)	24.7	26.6	24.7	47.7	22.4	31.3	32.8	24.2	12.2	30.3	24.2	27.1
Vanilla GRPO	47.3	29.8	28.4	34.9	32.9	43.4	43.8	27.3	32.4	27.6	21.2	32.2
Ours												
LENGUARD-GPC	**48.4**	**35.1**	**29.6**	**48.8**	**40.0**	**55.4**	**50.0**	19.7	**33.8**	23.7	**24.8**	**36.1**

from 44.4% to 46.9%. These results suggest that the proposed reward design generalizes beyond the exact training source while remaining effective within the multi-view spatial reasoning domain. On out-of-domain benchmarks, LENGUARD-GPC also maintains consistent gains over vanilla GRPO. It improves 3DSR from 59.4% to 60.1%, BLINK from 65.7% to 66.8%, and CV-Bench from 85.2% to 85.5%. Although these gains are smaller than those on in-domain benchmarks, they indicate that the guided-prompt consistency reward and staged length guard do not overfit to the training distribution or sacrifice general reasoning ability.

Compared with existing open-source baselines, LENGUARD-GPC achieves the best average performance among the listed open-source models. These results demonstrate that explicitly regulating reasoning behavior with guided-prompt consistency and staged length control is effective for improving multi-view spatial reasoning while preserving robustness on out-of-domain evaluations.

Table 2 further provides a fine-grained breakdown on MMSI-Bench. Compared with vanilla GRPO, LENGUARD-GPC improves the overall accuracy from 32.2% to 36.1% and achieves gains in most subcategories. The largest improvements appear in cross-view positional reasoning, including Cam.–Obj. ($+13.9$ points), Cam.–Reg. ($+12.0$ points), Obj.–Reg. ($+7.1$ points), and Obj.–Obj. ($+5.3$ points), suggesting that our reward design mainly strengthens spatial relation reasoning across views. In addition, LENGUARD-GPC improves Attribute Meas., Motion Cam., and MSR, while the drops on Attribute Appr. and Motion Obj. indicate that appearance-based attributes and object-motion reasoning remain challenging.

Although LENGUARD-GPC evaluates each sampled trajectory under both the standard and guided prompts, this does not require two full autoregressive generations. The trajectory is sampled only once from the standard prompt, while the guided-prompt distribution is computed over the fixed trajectory using teacher forcing, allowing token-level predictions to be evaluated in parallel. Moreover, the remaining GRPO operations, including advantage computation, backpropagation, and parameter updates, are shared with the vanilla training pipeline. Consequently, the additional guided-prompt evaluation increases the overall training time by only approximately $1.6\times$, rather than doubling it. The shorter trajectories induced by LENGUARD-GPC further offset part of this additional cost.

4.3 ABLATION STUDY

We conduct ablation studies on MMSI-Bench to evaluate the contributions of guided-prompt consistency and the staged length reward. As shown in Table 3, vanilla GRPO improves the base Qwen3-VL-8B-Instruct model from 31.1% to 32.2%. Adding guided-prompt consistency without the staged length reward further improves the accuracy to 33.5%, showing that the proposed con-

Table 3: Ablation study on MMSI-Bench (accuracy, %), comparing vanilla GRPO, the variant without the staged length reward, and the full method on fine-grained subcategories and overall performance.

Models	Positional Relationship						Attribute		Motion		MSR	Avg.
	Cam.–Cam.	Obj.–Obj.	Reg.–Reg.	Cam.–Obj.	Obj.–Reg.	Cam.–Reg.	Meas.	Appr.	Cam.	Obj.	–	
Qwen3-VL-8B-Instruct	28.0	37.2	32.1	31.4	35.3	38.5	37.5	15.2	27.0	28.9	29.8	31.1
Vanilla GRPO	47.3	29.8	28.4	34.9	32.9	43.4	43.8	27.3	32.4	27.6	21.2	32.2
w/o Staged Length Reward	45.2	34.0	23.5	46.5	32.9	47.0	45.3	25.8	27.0	26.3	24.8	33.5
Full Method	48.4	35.1	29.6	48.8	40.0	55.4	50.0	19.7	33.8	23.7	24.8	36.1

sistency reward provides additional supervision beyond the standard GRPO objective. The full LENGUARD-GPC achieves 36.1%, outperforming the variant without the staged length reward by 2.6 points. Improvements are observed across most fine-grained categories, especially cross-view positional reasoning. These results indicate that the staged length reward effectively complements guided-prompt consistency by preventing overly short reasoning while preserving the benefits of compact reasoning.

4.4 TOKEN EFFICIENCY ANALYSIS

In addition to accuracy, we further evaluate whether the proposed method can produce more compact reasoning trajectories. Figure 2 compares the average generated token cost between vanilla GRPO and LENGUARD-GPC on MMSI and ViewSpatial. Compared with vanilla GRPO, LENGUARD-GPC reduces the average token cost from 307.9 to 261.9 on MMSI, corresponding to a 14.9% reduction. The reduction is more substantial on ViewSpatial, where the average token cost decreases from 196.5 to 121.6, corresponding to a 38.1% reduction.

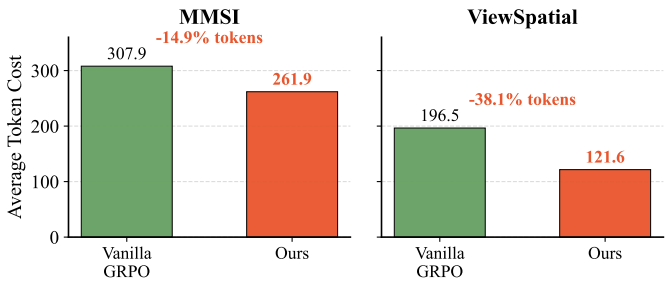

Figure 2: Average generated token cost on MMSI and ViewSpatial. Compared with vanilla GRPO, LENGUARD-GPC produces shorter reasoning trajectories while maintaining stronger accuracy, indicating more compact and adaptive chain-of-thought reasoning.

This result suggests that LENGUARD-GPC does not simply encourage uniformly long or short responses, but instead makes the reasoning length more adaptive. ViewSpatial contains both single-image and multi-image questions with different levels of reasoning difficulty. For easier single-image cases, the model can avoid unnecessary chain-of-thought expansion, while for harder multi-view cases, it can still retain sufficient reasoning when needed. Therefore, the proposed guided-prompt consistency reward and staged length guard compress redundant reasoning steps and improve the flexibility of reasoning length, leading to more efficient multimodal spatial reasoning.

5 CONCLUSION

In this paper, we propose LENGUARD-GPC, a GRPO-based framework for improving multi-view spatial reasoning through guided-prompt consistency and staged length regulation. By comparing model behavior under standard and guided prompts, our method provides dense trajectory-level feedback without inference-time overhead, while the staged length guard encourages sufficient but concise reasoning. Experiments on six benchmarks show that LENGUARD-GPC consistently outperforms the base Qwen3-VL-8B-Instruct model and vanilla GRPO, with clear gains on in-domain tasks. Fine-grained results further show stronger cross-view positional reasoning. Overall, LENGUARD-GPC enables more accurate, compact, and robust multimodal spatial reasoning.

AI USE STATEMENT

Large language models (LLMs) were used solely for data synthesis, specifically to generate auxiliary guidance for training instances. They were not involved in the conception of research ideas, experimental design, data analysis, or interpretation of results. All scientific content, methods, and conclusions were developed independently by the authors.

ETHICS STATEMENT

This work does not involve human subjects, sensitive data, animal experiments, or any other aspect that raises ethical concerns. No potential risks of misuse or negative societal impact have been identified.

REPRODUCIBILITY STATEMENT

We are committed to ensuring reproducibility of our results. All code, along with instructions for data preprocessing, model configuration, and evaluation, will be released upon publication to enable full replication of the experiments and results reported in this paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. *arXiv preprint arXiv:2511.07403*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

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

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kxnoqaisCT.

Zirun Guo, Minjie Hong, and Tao Jin. Observe-r1: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning. *arXiv preprint arXiv:2505.12432*, 2025.

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don't overthink it. preferring shorter thinking chains for improved llm reasoning. *arXiv preprint arXiv:2505.17813*, 2025.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Zhehan Kan, Yanlin Liu, Kun Yin, Xinghua Jiang, Xin Li, Haoyu Cao, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, et al. Taco: Think-answer consistency for optimized long-chain reasoning and efficient data learning via reinforcement learning in lvlms. *arXiv preprint arXiv:2505.20777*, 2025.

Kanghee Lee, Injae Lee, Minseok Kwak, Kwonyoung Ryu, Jungi Hong, and Jaesik Park. Spatialmosaic: A multiview vlm dataset for partial visibility. *arXiv preprint arXiv:2512.23365*, 2025.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1786–1807, 2024a.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025b.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2024.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Xingjian Tao, Yiwei Wang, Yujun Cai, Zhicheng Yang, and Jing Tang. Understanding gui agent localization biases through logit sharpness. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23361–23374, 2025.

Xingjian Tao, Yiwei Wang, Yujun Cai, Yihong Luo, Kai Han, and Jing Tang. Mitigating coordinate prediction bias from positional encoding failures. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20635–20650, 2026a.

Xingjian Tao, Yiwei Wang, Yujun Cai, Yifan Song, and Jing Tang. Viewfusion: Structured spatial thinking chains for multi-view reasoning. *arXiv preprint arXiv:2603.06024*, 2026b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Feiyang Wang, Nan Luo, and Wangyu Wu. Visioncube: 3d-aware vision-language model for multi-step spatial reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3270–3279, 2025a.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b.

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

Wentao Wang, Heqing Zou, Tianze Luo, Rui Huang, Yutian Zhao, Zhuochen Wang, Hansheng Zhang, Chengwei Qin, Yan Wang, Lin Zhao, et al. Video-str: Reinforcing mllms in video spatio-temporal reasoning with relation graph. *arXiv preprint arXiv:2510.10976*, 2025c.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28114–28128, 2025d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025a.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. *arXiv preprint arXiv:2505.17012*, 2025b.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025c.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for mllms. *arXiv preprint arXiv:2506.17901*, 2025d.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025a.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.

Shu Yang, Junchao Wu, Xuansheng Wu, Derek Wong, Ninhao Liu, and Di Wang. Is long-to-short a free lunch? investigating inconsistency and reasoning efficiency in lrms. *arXiv preprint arXiv:2506.19492*, 2025c.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025d.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025e.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025.

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

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv e-prints*, pp. arXiv–2405, 2024b.

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

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Wanglu Wanglu. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Prompt Templates

For each training instance, the original question is inserted into a fixed reasoning template. The standard prompt used for trajectory generation is:

```
Question: {question}

Please reason step-by-step according to the following
instructions:
Provide your detailed reasoning for the raw question
within the <thinking> </thinking> tags. Use the multiple
perspectives to discover objects or details not visible in
a single image. Ensure your logic synthesizes the visual
evidence into a coherent conclusion.
Final Output: Provide only the single option letter (e.g.,
A, B, C, D) within the <answer> </answer> tags.
```

Accordingly, the expected output has the form

$$\texttt{<thinking>} \cdots \texttt{</thinking>} \;\; \texttt{<answer>} \cdots \texttt{</answer>}. \tag{10}$$

For multi-image examples, if explicit image placeholders are not already present in the prompt, the data loader prepends one image placeholder for each input image. The same visual inputs are used under the standard and guided prompting conditions.

B Implementation Details

This appendix provides additional implementation details of LenGuard-GPC. Unless otherwise specified, all settings follow the released training configuration.

B.1 Training Configuration

The training set is loaded from a local JSONL file and shuffled with random seed 1. Validation is disabled during training. The actor uses full-shard FSDP together with gradient checkpointing. In addition to the proposed guided-prompt consistency reward, the underlying GRPO implementation retains its standard reference-policy KL regularization, using the low-variance KL estimator with coefficient 4×10^{-2}.

Table 4: Additional training hyperparameters used in our implementation.

Hyperparameter	**Value**
Training epochs	1
Optimizer	AdamW
Weight decay	1×10^{-2}
LR warmup ratio	0
Rollout batch size	16
Actor global batch size	8
Update micro-batch size / device	1
Experience micro-batch size / device	2
Maximum gradient norm	1.0
Random seed	1
FSDP full sharding	Enabled
Gradient checkpointing	Enabled
Reference KL penalty	low-var KL
Reference KL coefficient	4×10^{-2}

B.2 Image Preprocessing

The training entry point specifies the minimum and maximum visual sizes through image-token limits. The minimum and maximum numbers of image tokens are set to 4 and 512, respectively, with a spatial unit of 28×28 pixels. The resulting allowed image-area range is

$$4 \times 28^2 \leq HW \leq 512 \times 28^2, \tag{11}$$

i.e.,

$$3{,}136 \leq HW \leq 401{,}408. \tag{12}$$

Images exceeding the maximum area or falling below the minimum area are resized while preserving their aspect ratio. Non-RGB images are converted to RGB before model input.