SpaceR: Reinforcing MLLMs in Video Spatial Reasoning

Kun Ouyang[1] **Yuanxin Liu**[1] **Haoning Wu**[2] **Yi Liu**[1]
Hao Zhou[3] **Jie Zhou**[3] **Fandong Meng**[3] **Xu Sun**[1*]
[1] National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2] Nanyang Technological University [3] WeChat AI, Tencent Inc., China
kunouyang10@gmail.com xusun@pku.edu.cn

Abstract

Video spatial reasoning, which involves inferring the underlying spatial structure from observed video frames, poses a significant challenge for existing Multimodal Large Language Models (MLLMs). This limitation stems primarily from 1) the absence of high-quality datasets for this task, and 2) the lack of effective training strategies to develop spatial reasoning capabilities. Motivated by the success of Reinforcement Learning with Verifiable Reward (RLVR) in unlocking LLM reasoning abilities, this work aims to improve MLLMs in video spatial reasoning through the RLVR paradigm. To this end, we introduce the **SpaceR** framework. First, we present **SpaceR-151k**, a dataset with 91k questions spanning diverse spatial reasoning scenarios with verifiable answers, and 60k samples for maintaining general multimodal understanding. Second, we propose **Spatially-Guided RLVR (SG-RLVR)**, a novel reinforcement learning approach that extends Group Relative Policy Optimization (GRPO) with a novel map imagination mechanism, which encourages the model to infer spatial layouts in the thinking process, thereby facilitating more effective spatial reasoning. Extensive experiments demonstrate that SpaceR achieves state-of-the-art performance on spatial reasoning benchmarks (e.g., VSI-Bench, STI-Bench, and SPAR-Bench), while maintaining competitive results on video understanding benchmarks (e.g., Video-MME, TempCompass, and LongVideoBench). Remarkably, SpaceR surpasses the advanced GPT-4o by 11.6% accuracy on VSI-Bench and is on par with the leading proprietary model Gemini-2.0-Flash, highlighting the effectiveness of our SpaceR-151k dataset and SG-RLVR in reinforcing spatial reasoning ability of MLLMs. Code, model, and dataset are available at https://github.com/OuyangKun10/SpaceR.

1 Introduction

Video spatial reasoning [34] requires reconstructing 3D spatial layouts from sequences of observed frames. This task demands a higher-level cognitive ability than conventional video understanding tasks, such as video captioning [31], video question answering [1], and temporal grounding [10], which typically require only recall of video content. Although recent advancements in Multimodal Large Language Models (MLLMs) have significantly improved performance on conventional video understanding [4, 2, 12], these models still struggle with video spatial reasoning [17, 34]. This limitation stems mainly from two factors: 1) the absence of a high-quality dataset specifically

*Corresponding Author(s)

Preprint. Under review.

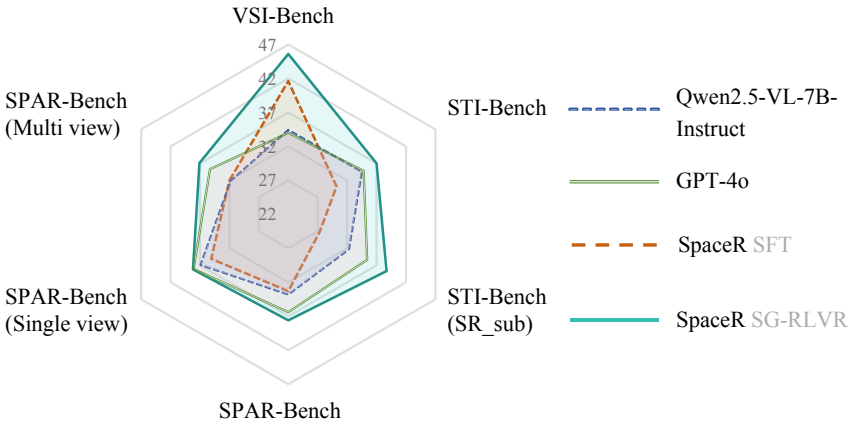


Figure 1: Performance comparison on spatial reasoning benchmarks and their corresponding subsets for Qwen2.5-VL-7B-Instruct [2], GPT-4o [12], our SpaceR SFT and SpaceR SG-RLVR.

designed for spatial reasoning, and 2) the reliance of most existing MLLMs on supervised fine-tuning (SFT) during post-training, which is insufficient for fostering deep reasoning capabilities.

In contrast to SFT, recent studies have demonstrated that Reinforcement Learning with Verifiable Rewards (RLVR) is more effective in enhancing the reasoning capabilities of both LLMs and MLLMs within pure-text [11] and multimodal tasks [7, 8, 23]. For example, DeepSeek-R1-Zero [11] utilizes the Group Relative Policy Optimization (GRPO) algorithm to unlock the reasoning capabilities of LLMs. Concurrent work Video-R1 [8] introduce a large-scale multimodal understanding dataset Video-R1-260k and extend GRPO with a novel temporal reward for RLVR, which also achieves promising performance in video understanding benchmarks like MVBench [16]. Inspired by these findings, this work aims to advance MLLMs' spatial reasoning abilities in video through RLVR.

Specifically, we propose the **SpaceR** framework, which encompasses two key innovations: **First**, we introduce the SpaceR-151k dataset, which consists of 151k samples, including 91k spatial reasoning QA pairs (SR-91k) curated based on a 3D reconstruction dataset ScanNet [5], and 60k samples drawn from the general multimodal understanding dataset Video-R1-260k [8]. In particular, SR-91k spans six spatial reasoning tasks (e.g., relative direction, object/room size, and appearance order), filling a critical gap in available resources. **Second**, we extend the GRPO [28] paradigm to enhance spatial reasoning by designing task-specific verifiable rewards for various QA formats (e.g., multiple choice, numerical). Furthermore, we design a novel map imagination mechanism, where models are prompted to generate an optional cognitive map, a structured representation of object positions in space, within specialized tags `<map>···</map>`. And a map reward is employed to evaluate the quality of these inferred spatial layouts, which incentivizes models to think in space deeply.

Extensive experiment demonstrate that our SpaceR delivers consistent and significant performance gains across several challenging spatial reasoning benchmarks, including VSI-Bench [34], STI-Bench [17], and SPAR-Bench [39], while maintaining promising results in representative video understanding benchmarks like Video-MME [9], TempCompass [22], and LongVideoBench [32]. Notably, our model achieves 45.6% accuracy on VSI-Bench [34], outperforming the advanced proprietary model GPT-4o [12] by 11.6% accuracy, which is presented in Figure 1. These empirical results validate both the utility of the SpaceR-151k dataset and the effectiveness of our SG-RLVR in unlocking spatial reasoning capabilities of MLLMs.

Our contributions are threefold.

- We introduce the SpaceR-151k dataset specifically designed for video spatial reasoning. It consists of questions spanning diverse spatial reasoning scenarios with verifable answers, addressing the scarcity of resources in this domain.
- We propose SG-RLVR, a spatially-guided reinforcement learning framework that integrates a novel map imagination mechanism. It encourages the model to explicitly generate spatial layouts to facilitate video spatial reasoning.

- We conduct extensive evaluations across spatial reasoning and video understanding benchmarks, demonstrating that SpaceR achieves state-of-the-art spatial reasoning capabilities and promising generalizability in video understanding, validating the effectiveness of both the SpaceR-151k dataset and our SG-RLVR framework.

2 Related Works

2.1 Video Spatial Reasoning

Video understanding tasks like video captioning [31], temporal grounding [10, 14], and temporal perception [24, 19], primarily focus on recalling or summarizing video content. For instance, video captioning necessitates models to generate relevant textual descriptions of the video based on human prompts. Recent advances in Multimodal Large Language Models (MLLMs) have significantly improved performance on these tasks. Unlike these conventional understanding tasks, video spatial reasoning requires models not only to perceive visual content but also to infer and reconstruct the spatial structure of entire scenes. It is worth noting that video spatial reasoning is crucial for the development of world models [20] and embodied agents [6]. Recent studies [17, 34, 39] have highlighted the persistent shortcomings of MLLMs on spatial reasoning, underscoring the need for further research in this area. A key limitation contributing to this gap is the scarcity of high-quality training data specifically tailored for video spatial reasoning, which we aim to address in this work.

2.2 Reinforcement Learning with Verifiable Reward

Recent works like o1 [13], DeepSeek-R1 [11], Kimi k1.5 [29] have demonstrated significant breakthroughs in enhancing the reasoning capabilities of large language models (LLMs) through Reinforcement Learning (RL). In particular, the Group Relative Policy Optimization (GRPO) [28] algorithm, applied in DeepSeek-R1, has revealed the strong potential of Reinforcement Learning with Verifiable Reward (RLVR) framework in equipping MLLMs with advanced reasoning capacity. Building on this foundation, several subsequent efforts [23, 25] have employed RLVR to boost visual reasoning performance. For example, Visual-RFT [23] improved MLLMs in multimodal detection [26], grounding [36], and classification [27]. LMM-R1 [25] empowers 3B MLLMs with strong reasoning abilities of mathematics through two-stage rule-based RL. Nevertheless, research specifically targeting video spatial reasoning remains underexplored. Motivated by this gap, our work seeks to design an effective reasoning paradigm to enhance MLLMs' capabilities in video spatial reasoning.

3 Dataset Construction

To address the scarcity of high-quality data for video spatial reasoning and maintain general video understanding in the meanwhile, we construct **SpaceR-151k**, a large-scale dataset consisting of two parts: 1) **SR-91k**, a tailored spatial reasoning dataset built upon the 3D indoor scene reconstruction dataset ScanNet [5], and 2) 60k QA instances resampled from the general multimodal understanding dataset Video-R1-260k [8]. The construction process follows three stages: data collection, QA generation, and data filtering. An overview of QA types and examples is provided in Figure 2, while data statistics are presented in Figure 3.

Data Collection. 1) Spatial reasoning. We first parse ScanNet into a unified meta-information format, including object categories, appearance indices, bounding box and other relevant attributes, to facilitate QA generation. RGB frames of ScanNet are resampled at 24 FPS to form video clips. Besides, we construct a 10×10 map for each video to summarize the object distribution of the room, which is exemplified in Figure 2. Each object's coordinate is determined by the center point of its bounding box and projected onto a 2D map, which is represented in a dictionary format for structured spatial reference. 2) General understanding. To preserve general comprehension capabilities, we uniformly sample $60,000$ diverse QA instances from Video-R1-260k. This subset covers multiple QA types including multi-choice, numerical, OCR, free-form, and regression.

QA Generation. Leveraging the parsed meta-information of ScanNet, we automatically generate the question-answering (QA) pairs for spatial reasoning tasks, which are categorized into multi-choice QA (e.g., relative distance, relative direction, and appearance order) and numerical QA (e.g., object/room size, absolute distance, and counting). The QA examples are presented in Figure 2.

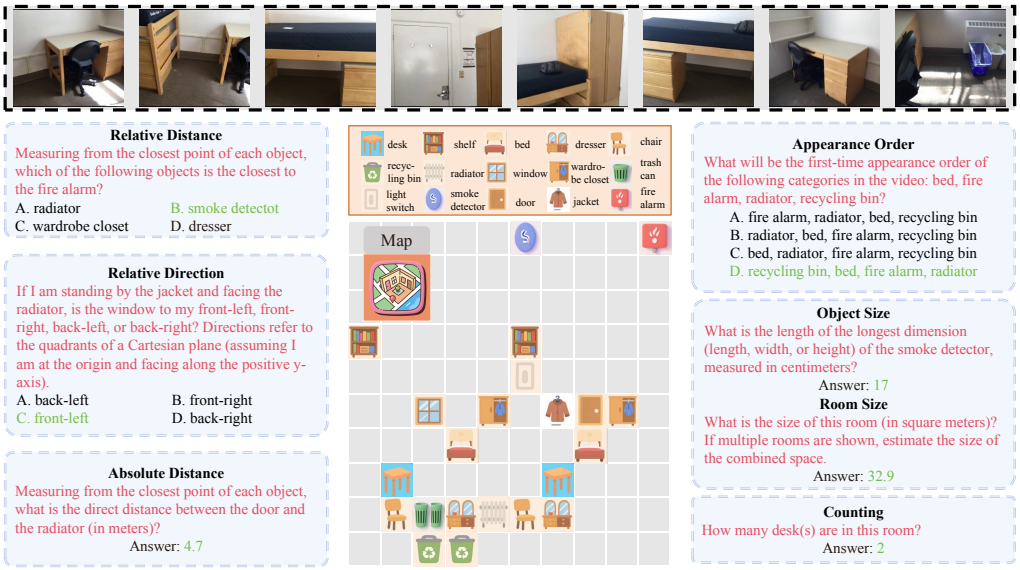


Figure 2: The overview of the Question-Answering examples from SR-91k, including multi-choice QA (e.g., relative distance, relative direction, appearance order) and numerical QA (e.g., object/room size, absolute distance, and counting), as well as the corresponding map for the video.

- **Relative Distance**. For each video, we first identify unique objects, randomly select a target object and four candidate objects to be incorporated into the question template. Finally, the minimum Euclidean distance between each target and candidate is computed to determine the answer.
- **Relative Direction**. Utilizing previous identified unique objects in the video, we randomly select three of them to be integrated in the question template. Relative directions are determined on the basis of their center points of bounding box.
- **Appearance Order**. We record the first frame index where each object appears, and randomly sample four objects from them to generate the questions. The ordering is determined by their first frame indices.
- **Object/Room Size**. Object size is defined as the longest dimension of an object computed from point clouds, and is converted to centimeters. Room size (in square meters) is estimated via the Alpha Shape algorithm[2].
- **Absolute Distance**. We uniformly sample points within the object bounding boxes and estimate the minimum Euclidean distance between two unique objects in the video.
- **Counting**. We obtain the number of each object appearing in the video from the meta-information of ScanNet.

Data Filtering. To ensure the quality of the spatial reasoning QA pairs, we apply a series of filtering steps. First, we limit the number of QA pairs per video to promote scene diversity. For multi-choice QA, we randomly shuffle the positions of correct answers to balance answer distribution and eliminate position bias. In addition, we meticulously adjust the numerical value distribution in the numerical QA to prevent skewed or unrealistic value shifts. After filtering, we retain 91k high-quality QA pairs, forming the SR-91k dataset for spatial reasoning.

Data Statistics. The data statistics of SpaceR-151k are exhibited in Figure 3. This dataset comprises a total of $151,310$ samples, integrating 91k spatial reasoning QA pairs (SR-91k) with 60k instances drawn from the Video-R1-260k. SpaceR-151k features a diverse range of QA types: multi-choice, numerical, OCR, free-form, and regression, whose answers are verifiable. Above all, SpaceR-151k provides rich sources for both spatial reasoning and general understanding, which is the foundation of subsequent training. More details for dataset construction and statistics can be found in Appendix A.

[2] `https://en.wikipedia.org/wiki/Alpha_shape`.

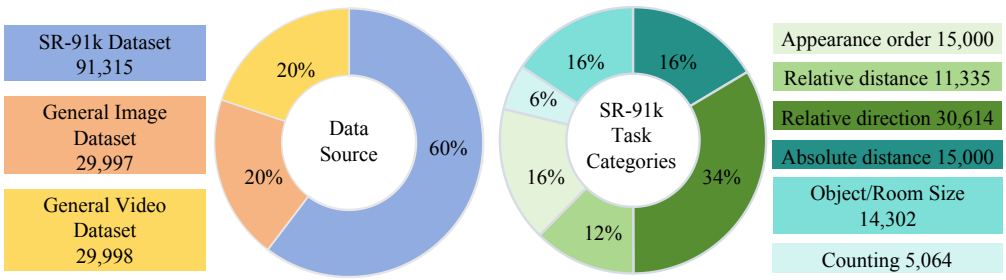


Figure 3: Data statistics of our SpaceR-151k. Left: the distribution of data sources. Right: the task category distribution within SR-91k.

4 Spatially-Guided Reinforcement Learning with Verifiable Reward

To reinforce video spatial reasoning in MLLMs, we propose a reinforcement learning framework named SG-RLVR, which builds on Group Relative Policy Optimization (GRPO) [11] by introducing verifiable reward functions tailored to diverse QA types and a novel map imagination mechanism to guide spatial reasoning.

4.1 Verifiable Reward Function

To supervise model outputs across multiple QA types, including multi-choice, numerical, OCR, free-form, and regression, we design a set of verifiable reward functions that assess either response format or correctness based on task-specific criteria.

Format Reward. To ensure the model responses adhere to a predefined structure, we define a format reward R_{format} based on whether the model wraps its reasoning process and answer within `<think>···</think>` and `<answer>···</answer>` tags, respectively:

$$R_{format}(\hat{y}) = \begin{cases} 1, & \text{if } \hat{y} \text{ matches format,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Multi-choice Reward. For multi-choice QA, the reward R_{mc} is binary, based on exact match with the ground truth:

$$R_{mc}(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y} = y, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\hat{y}$ is the model's response and y is the ground truth.

Numerical Reward. To assess numerical values, we compute relative accuracy across varying confidence thresholds $\theta_i \in \{0.5, 0.55, \ldots, 0.95\}$. The numerical reward R_{num} is defined as:

$$R_{num}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(\frac{|\hat{y} - y|}{y} \leq 1 - \theta_i \right), \tag{3}$$

N is the number of confidence thresholds. Besides, for general multimodal understanding data from Video-R1-260k, we incorporate three additional reward functions: OCR, free-form, and regression rewards [8]. The OCR reward is computed based on the Word Error Rate (WER)[3], which measures the edit distance between the predicted and reference text. The free-form reward is calculated as the average of ROUGE-1, ROUGE-2, and ROUGE-L scores [18] between the model's response and the ground truth. The regression reward is determined by the relative distance between the numerical values of response and ground truth.

4.2 Map-Based Spatial Reasoning

Since previous RLVR frameworks like GRPO [11] lack explicit reward signals for spatial information comprehension when applied to video spatial reasoning, we propose a map imagination mechanism

[3] `https://en.wikipedia.org/wiki/Word_error_rate`

that encourages the model to think in space. Specifically, the model is guided to generate a $M \times M$ map to identify object distributions within the scene, supporting downstream reasoning and lead to more reliable answer. To evaluate the quality of the generated map, we design a novel map reward R_{map} that provides precise quantitative feedback to facilitate spatial reasoning. Particularly, we first calculate the relative accuracy between predicted object and ground truth object by their relative distance $\frac{\sqrt{(x_{p,i}-x_{g,i})^2+(y_{p,i}-y_{g,i})^2}}{\sqrt{M^2+M^2}}$, where M is the size of grid map, and average the relative accuracy across all objects to derive map reward R_{map}. Mathematically,

$$R_{map} = \sum_{i=1}^{k} \left(\frac{n_i}{\sum_{j=1}^{k} n_j} \times \left(1 - \frac{\sqrt{(x_{p,i} - x_{g,i})^2 + (y_{p,i} - y_{g,i})^2}}{\sqrt{M^2 + M^2}} \right) \right), \tag{4}$$

where k is the number of object categories, n_i is the number of i-th object. $(x_{p,i}, y_{p,i})$ and $(x_{g,i}, y_{g,i})$ are the coordinates of the i-th object in the predicted map and ground truth map. To regulate the reasoning process, we introduce a length-based reward R_l that encourages outputs to fall within a defined length range: $[l_{min}, l_{max}]$. This helps strike a balance between promoting sufficient reasoning and avoiding overthinking. R_l is applied only when the model produces a correct answer within the desired length. Formally, the map imagination augmented reward R_m is defined as:

$$R_m = \begin{cases} R_{format} + R_{task} + R_{map} + R_l, & \text{if } R_{task} = 1 \\ R_{format} + R_{task} + R_l, & \text{otherwise,} \end{cases} \tag{5}$$

where $task \in \{mc, num, ocr, free, reg\}$. This reward shaping ensures that when the model answers correctly and can properly understand the space of the indoor scene, it receives additional reward, pushing the optimization toward adopting a more spatial aware reasoning policy. The advantage A_i, representing the relative quality of the i-th response o_i, is computed over the updated rewards within each group of responses $[o_1, o_2, \cdots, o_G]$, where G is the number of output responses. The final optimized policy π_θ is prevented from deviating far from the original model parameters π_{ref} by adding a KL-divergence term $\mathcal{D}_{\text{KL}}(\cdot\|\cdot)$ to the following formulation:

$$A_i = \frac{R_m - \text{mean}(\{R_m\})}{\text{std}(\{R_m\})}. \tag{6}$$

The final policy update follows the clipped surrogate objective of GRPO:

$$\begin{aligned} J(\theta) = \mathbb{E}_{q,\{o_i\}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Bigg(& \min \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) A_i \right) \Bigg) \\ & - \beta \mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Bigg], \end{aligned} \tag{7}$$

where β is a regularization coefficient, preventing excessive deviation from the reference policy during optimization, ϵ is a positive coefficient limits the policy updating degree.

5 Experiment

5.1 Experimental Setups

Implementation Details. 1) In the training stage, we adopt Qwen-2.5-VL-7B-Instruct[4] as the base model. The training process is conducted for a maximum of 2 epochs with a per-device batch size of 1. 8 response candidates are generated for each sample. The maximum completion length is set to $1,024$ tokens. l_{min} and l_{max} are set to 360 and 512, respectively. To balance computational efficiency and model performance, we restrict the number of video frames to 16, with each frame processed at a resolution of $128 \times 28 \times 28$. 2) In the evaluation period, we prompt SpaceR to explicitly perform a step-by-step reasoning process on spatial reasoning benchmarks, while directly generating answers for video understanding benchmarks. For the base model, we prompt it to directly produce answers,

[4] `https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct`.

as it is not trained to perform intermediate reasoning. The generation temperature is uniformly set to 0.01. The maximum number of new tokens is set to $1,024$ when reasoning steps are included, and 128 for direct answer generation. The number of video frames is standardized to 32 during evaluation, and each frame is processed at a resolution of $448 \times 28 \times 28$.

Benchmarks. A diverse set of evaluation benchmarks is employed to comprehensively assess the model's capabilities in both spatial reasoning and video understanding. We conduct extensive evaluation on three spatial reasoning benchmarks (i.e., VSI-Bench [34], STI-Bench [17], and SPAR-Bench [39]) and three video understanding benchmarks (i.e., Video-MME [9], TempCompass [22], and LongVideoBench [32]). The detailed description and usage for each benchmark, as well as the evaluated baselines, are summarized in Appendix B.

	#Params	Frames	Spatial Reasoning						Video Understanding		
			VSI-Bench	STI-Bench		SPAR-Bench			VM	TC	LV
				Overall	SR_sub	Overall	Single-view	Multi-view			
Closed-source Models											
GPT-4o [12]	-	-	34.0	34.8	35.4	36.4	38.1	35.3	71.9	73.8	66.7
Gemini 1.5 Pro	-	-	48.8	-	-	-	-	-	75.0	67.1	64.0
Gemini 2.0 Flash	-	-	45.4	38.7	39.8	-	-	-	-	-	-
Gemini 2.5 Pro	-	-	-	40.9	40.5	-	-	-	-	-	-
Open-source Models											
VideoLLaMA3-7B [38]	7B	-	-	26.9	27.2	-	-	-	66.2	68.1	59.8
LLaVA-OneVision-7B [15]	7B	-	32.4	-	-	31.2	33.1	29.9	58.2	-	56.3
MiniCPM-V-2.6 [35]	8B	-	-	26.9	29.6	-	-	-	60.9	-	54.9
Kimi-VL-A3B-Instruct [30]	3B/16B	16	37.4	-	-	-	-	-	62.3	70.3	58.0
InternVL2.5-78B [3]	78B	-	-	28.4	29.8	-	-	-	72.1	-	63.6
Qwen2.5-VL-72B-Instruct [2]	72B	32	35.6	40.8	36.9	36.4	40.6	33.6	61.3	75.3	57.1
Qwen2.5-VL-7B-Instruct [2]	7B	32	34.4	34.5	32.3	33.8	36.9	31.8	56.3	71.1	53.5
SpaceR SFT	7B	32	41.6	30.2	27.3	33.3	35.1	32.0	57.6	69.3	54.3
SpaceR SG-RLVR	7B	32	45.6 (↑ 11.2)	37.0 (↑ 2.5)	38.7 (↑ 6.4)	37.6 (↑ 3.8)	38.2 (↑ 1.3)	37.1 (↑ 5.3)	57.9 (↑ 1.6)	71.4 (↑ 0.3)	54.6 (↑ 1.1)

Table 1: Evaluation results of base model Qwen2.5-VL-7B-Instruct , SpaceR , and other baselines on spatial reasoning benchmarks (VSI-Bench, STI-Bench, and SPAR-Bench), and video understanding benchmarks: **VM** (Video-MME), **TC** (TempCompass), and **LV** (LongVideoBench). SR_sub is a subset containing six spatial reasoning sub-tasks of STI-Bench.

5.2 Main Results

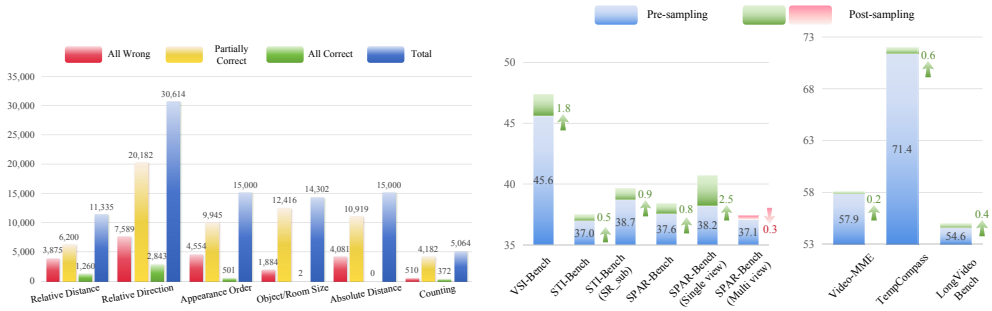
The evaluation results on the six benchmarks are presented in Table 1. And we have the following observations and analyses.

Overall Analysis. Overall, our SpaceR consistently outperforms the base model Qwen2.5-VL-7B-Instruct across all benchmarks. In spatial reasoning benchmarks, SpaceR even surpasses the proprietary GPT-4o model, highlighting its superior spatial reasoning capabilities. Notably, SpaceR achieves a significant improvement in accuracy gains 11.2 on VSI-Bench, a representative benchmark for spatial reasoning, underscoring its enhanced reasoning ability to model complex spatial relationships. Beyond spatial reasoning, SpaceR also generalizes well to video understanding tasks, achieving higher accuracy across all three benchmarks: Video-MME, TempCompass, and LongVideoBench, compared to Qwen2.5-VL-7B-Instruct. This indicates that the spatial reasoning enhancements and general multimodal understanding training samples contribute to broader video comprehension capabilities.

SG-RLVR vs SFT. We further compare the effectiveness of our proposed SG-RLVR and SFT. While SFT yields localized improvements on benchmarks, such as VSI-Bench, Video-MME, and LongVideoBench, it leads to performance degradation on other benchmarks, indicating limited generalizability. In contrast, SG-RLVR consistently improves performance across both spatial reasoning and video understanding benchmarks, highlighting its better generalizability. These results support the claim that "SG-RLVR generalizes, while SFT memorizes," establishing SG-RLVR as a more effective training paradigm for enhancing spatial reasoning in MLLMs.

Impact of Data Sampling on Model Performance. To improve training efficiency and model generalization, we conduct a sample selection strategy on the SR-91k dataset by filtering out samples deemed too easy or too difficult. Using Qwen2.5-VL-7B-Instruct to generate 8 responses per sample, we categorize samples into all correct, partially correct, and all wrong, based on model response consistency, those all correct samples (low learning value) and all wrong samples (potential noise) are excluded. As illustrated in Figure 4(a), the remaining samples span a balanced range of task categories. Retraining SpaceR on this filtered dataset results in consistent performance gains across nearly all benchmarks, as shown in Figure 4(b). These findings suggest that targeted resampling enhances

model training by focusing on samples with high learning utility, ultimately leading to improved generalization in both spatial reasoning and video understanding tasks. Additional analyses for the impacts of thinking, model size, and data scale on model performance are provided in Appendix C.



(a) Sample distribution.

(b) Performance variations between pre-sampling and post-sampling.

Figure 4: Sample distribution and Performance variations between pre-sampling and post-sampling.

	Frames	Spatial Reasoning						Video Understanding		
		VSI-Bench	STI-Bench		SPAR-Bench			VM	TC	LV
			Overall	SR_sub	Overall	Single-view	Multi-view			
w/o-map imagination	32	43.9	34.0	33.3	34.1	34.7	33.8	56.9	**71.9**	52.8
w/o-general data	32	**46.9**	**37.1**	37.2	37.1	35.5	**39.5**	56.3	71.0	53.6
w/o-SR data	32	26.2	34.2	33.9	30.2	30.7	29.5	57.4	70.5	52.9
SpaceR SG-RLVR	32	45.6	37.0	**38.7**	**37.6**	**38.2**	37.1	**57.9**	71.4	**54.6**

Table 2: Ablation results of SpaceR SG-RLVR, where the best results are in boldface.

5.3 Ablation Study

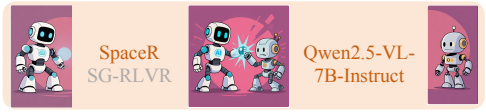
To explore the contributions of individual components in our method, we introduce three variants of our SpaceR: 1) w/o-map imagination, which eliminates the map imagination mechanism in the training and inference stage. 2) w/o-general data, which excludes 60k general multimodal understanding data in the training stage. 3) w/o-SR data, which removes SR-91k data in the training process. The ablation results are presented in Table 2, based on which we have the following findings. a) SpaceR consistently outperforms w/o-map imagination on three spatial reasoning benchmarks, which validates the advantage of map imagination mechanism to guide MLLMs in understanding spatial information. b) SpaceR exceeds w/o-general data on video understanding benchmarks, which prove the effectiveness of the $60,000$ general multimodal understanding training data. c) w/o-SR data shows significantly lower performance on spatial reasoning benchmarks compared to SpaceR, emphasizing the critical importance of SR-91k dataset in enhancing spatial reasoning capabilities.

5.4 Qualitative Analysis

To get an intuitive understanding on the advancement of SpaceR in video spatial reasoning, we present two cases from VSI-Bench in Figure 5. In case (a), SpaceR demonstrates a clear qualitative superiority over Qwen2.5-VL-7B through structured reasoning, explicit spatial mapping, and an accurate conclusion. SpaceR correctly identifies the "table" as the object closest to the "stove", supporting its decision with a cognitive map that simulates the spatial layouts of the scene. In contrast, Qwen2.5-VL-7B relies on shallow, assumption-based heuristics and incorrectly selects the "stool", reflecting a lack of verifiable spatial reasoning. This emphasizes SpaceR's enhanced spatial awareness and reasoning depth, enabled by its map imagination mechanism in the SG-RLVR framework. Similarly, in case (b), SpaceR also beats Qwen2.5-VL-7B on a "chair-counting" task. By reasoning across multiple frames and accounting for partially occluded objects, SpaceR accurately concludes the presence of at least three chairs and reinforces its answer with a cognitive map. In contrast, Qwen2.5-VL-7B underestimates the count, providing a wrong answer of two chairs. Together, these cases prove the SpaceR's improved ability of spatial reasoning.

Measuring from the closest point of each object, which of these objects (stool, tv, table, refrigerator) is the closest to the stove?

A. stool B. tv C. table D. refrigerator



SpaceR SG-RLVR:

<think>Let me think through this step by step. ... The stove is located near the center of the kitchen, and there's a table positioned directly in front of it. The table has a wooden top and white legs, and it appears to be a central feature in the kitchen. The stool is also present in the kitchen but seems to be placed further away from the stove compared to the table. The tv is far from the stove in the right side of the kitchen. And the refrigerator is located in the left side of the kitchen. ... The stool tv and refrigerator , while present, arenot as close to the stove as the table is. ... Let's summarize the map: <map>{'stool': [[7, 4]], 'tv': [[9, 3]], 'table': [[3, 3]], 'stove':[[3,2]], 'refrigerator': [[0, 3]]}</map> The table is the closest to the stove based on the spatial arrangement and the function of the objects in the kitchen.</think><answer>C</answer>

Qwen2.5-VL-7B-Instruct: B

(a) Comparison on object relative distance task.

How many chair(s) are in this room?

Answer: 3

SpaceR SG-RLVR:

<think>Let me think through this step by step. ... In the first frame, there's a chair near the desk, but it's not fully visible. In the second frame, we see more of the room, including another chair that appears to be a standard office chair with wheels. The third frame shows a different angle of the room, and there seems to be a third chair partially obscured behind some boxes. The fourth frame focuses on a corner of the room, and there's a glimpse of what looks like a fourth chair, though it's difficult to confirm due to the angle and the clutter. Considering the visible evidence and the typical layout of an office, it's reasonable to conclude that there are at least three chairs in the room. ... The question asks for the number of chairs, and based on the visible evidence, we can confidently say there are at least three chairs. Now, let's summarize the cognitive map: <map>... 'chair': [[2, 3], [5, 3], [6, 3]] ...</map> </think> <answer>3</answer>

Qwen2.5-VL-7B-Instruct: 2

(b) Comparison on object counting task.

Figure 5: Two samples from VSI-Bench [34], as well as the corresponding responses from SpaceR SG-RLVR and Qwen2.5-VL-7B-Instruct [2].

6 Conclusion

In this work, we introduce SpaceR, a novel framework designed to enhance video spatial reasoning capabilities. To this end, we construct SpaceR-151k, a comprehensive dataset that includes 91k high-quality spatial reasoning QA paris (SR-91k) and 60k samples for general video understanding. Building upon this dataset, we propose Spatially-Guided Reinforcement Learning with Verifiable Reward (SG-RLVR), a novel reinforcement framework, which integrates task-specific reward functions and a map imagination mechanism to guide models in spatial layout inference and foster structured spatial reasoning. Extensive evaluations across three spatial reasoning benchmarks and video understanding benchmarks validate the effectiveness and generalizability of SpaceR. Nevertheless, certain limitations remain. For example, the current framework lacks mechanisms for adaptively controlling the depth of reasoning, which may affect inference efficiency in practice. This is expected to be explored in our future work. We hope that SpaceR serves as a solid foundation for advancing research in video spatial reasoning and inspires further exploration into reasoning-aware training for MLLMs.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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

[3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839.

[6] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488.

[7] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*.

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

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

[10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

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

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

[13] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

[14] Yang Jin, Zehuan Yuan, Yadong Mu, et al. 2022. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–29204.

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

[16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2023. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*.

[17] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. 2025. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*.

[18] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[19] Kun-Yu Lin, Jia-Run Du, Yipeng Gao, Jiaming Zhou, and Wei-Shi Zheng. 2023. Diversifying spatial-temporal perception for video domain generalization. *Advances in Neural Information Processing Systems*, 36:56012–56026.

[20] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*.

[21] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. 2025. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*.

[22] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8731–8772.

[23] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

[24] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

[25] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

[26] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

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

[29] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

[30] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

[31] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504.

[32] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

[33] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

[34] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

[35] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

[36] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

[37] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

[38] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

[39] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. 2025. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*.

A More Details for Data Construction

Relative Distance	Measuring from the closest point of each object, which of these objects ({object a}, {object b}, {object c}, {object d}) is the closest to the {target object}?
Relative Direction	1. If I am standing by the {positioning object} and facing the {orienting object}, is the {querying object} to the left or the right of the {orienting object}? 2. If I am standing by the {positioning object} and facing the {orienting object}, is the {querying object} to my left, right, or back? An object is to my back if I would have to turn at least 135 degrees in order to face it. 3. If I am standing by the {positioning object} and facing the {orienting object}, is the {querying object} to my front-left, front-right, back-left, or back-right? Directions refer to the quadrants of a Cartesian plane (assuming I am at the origin and facing the positive y-axis).
Appearance Order	What will be the first-time appearance order of the following categories in the video: {choice a}, {choice b}, {choice c}, {choice d}?
Object/Room Size	1. What is the length of the longest dimension (length, width, or height) of the {object}, measured in centimeters? 2. What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.
Absolute Distance	Measuring from the closest point of each object, what is the direct distance between the {object 1} and the {object 2} (in meters)?
Counting	How many {object}(s) are in this room?

Figure 6: Question templates for QA pairs of SR-91k.

QA Generation. To format the generated QA pairs, we incorporate the corresponding objects into the specified question templates, which are presented in Figure 6.

Data Filtering. We remove the QA pairs that involve some noisy objects (e.g., "wall", "floor", and "ceiling"). And we also drop the numerical QA pairs, where the objects are too small to identify. Considering VSI-Bench [34] is partially built on ScanNet [5], we exclude the overlapped videos in our SR-91k to ensure fair evaluation.

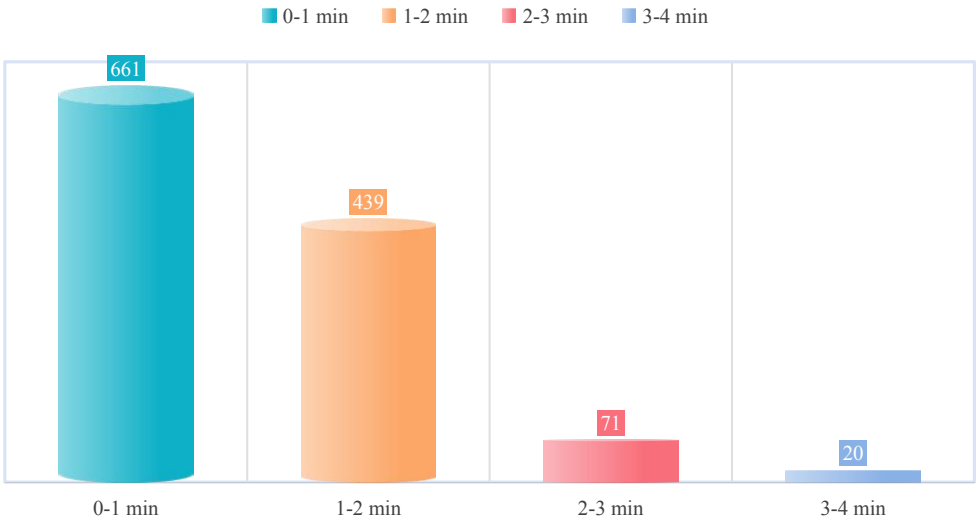


Figure 7: Video duration distribution of SR-91k.

Data Statistics. We visualize the duration distribution of videos from SR-91k in Figure 7.

B More Details for Experiment

B.1 Benchmarks Description

- **VSI-Bench** [34] is a comprehensive benchmark for evaluating the visual-spatial intelligence of Multimodal Large Language Models (MLLMs). It comprises over $5,000$ question-answer pairs across 288 real-world indoor scene videos, covering diverse environments such as homes, offices, and factories, and is specifically designed to assess spatial reasoning capabilities.
- **STI-Bench** [17] evaluates the spatial understanding abilities of MLLMs using real-world videos spanning desktop, indoor, and outdoor scenarios. It includes eight challenging tasks, with the subset SR_sub, which contains more than $2,000$ QA pairs across six sub-tasks (i.e., Dimensional Measurement, Displacement & Path Length,Ego-Centric Orientation, Spatial Relation, Speed & Acceleration, Trajectory Description), being most relevant to our focus on spatial reasoning.
- **SPAR-Bench** [39] is specifically designed to measure the spatial understanding of MLLMs. It contains over $7,000$ QA pairs covering a spectrum of tasks from basic perception to complex spatial reasoning. The benchmark is further divided into single-view and multi-view settings, allowing for comprehensive assessment across varying spatial contexts.
- **Video-MME** [9] serves as a comprehensive benchmark for evaluating general video understanding in MLLMs. It includes 900 videos and $2,700$ high-quality multi-choice questions (three per video), spanning a wide range of scenarios and tasks. We exclude the subtitles of videos in the evaluation.
- **TempCompass** [22] focuses on temporal perception in MLLMs. It consists of 410 videos and $7,540$ questions designed to evaluate models' understanding of temporal dynamics.
- **LongVideoBench** [32] is a benchmark for long-context multimodal video understanding. It features $6,678$ carefully constructed multi-choice questions derived from videos of varying durations, extending up to one hour, and encompasses diverse real-world themes. We utilize the validations set of it and remove the subtitles of videos.

B.2 Baselines Description

- **GPT-4o** [12] is a state-of-the-art MLLM developed by OpenAI, exhibiting strong performance across a variety of vision-language tasks.
- **Gemini 1.5 Pro, Gemini 2.0 Flash, Gemini 2.5 Pro** are advanced MLLMs from Google's Gemini family[5]. These models have shown leading performance across several video understanding benchmarks (e.g., Video-MME [9], and LongVideoBench [32]). Gemini 2.0 Flash and Gemini 2.5 Pro, in particular, exhibit enhanced abilities in complex reasoning tasks.
- **VideoLLaMA3-7B** [38] is an MLLM tailored for both image and video understanding. It adopts Qwen2.5-7B [33] as its language backbone and integrates siglip-so400m-patch14-384 [37] as the vision encoder.
- **LLaVA-OneVision-7B** [15] represents a strong advancement in open-source multimodal language models (LMMs), combining the Qwen2 [33] language backbone with the SigLIP [37] vision encoder. This integration pushes the performance boundaries of open LMMs, particularly in tasks requiring fine-grained visual understanding.
- **MiniCPM-V-2.6** [35] is developed based on SigLIP-400M [37] and Qwen2-7B [33], and introduces enhanced capabilities for multi-image and video understanding. Its architectural improvements and task-specific design make it a competitive model for complex multimodal understanding tasks.
- **InternVL2.5-78B** [3] is a high-performing open-source MLLM that combines InternViT-6B-448px-V2_5 [4] as the vision encoder with Qwen2.5-72B-Instruct [33] as the LLM backbone.

[5] `https://aistudio.google.com`.

- **Kimi-VL-A3B-Instruct** [30] is an efficient open-source MLLM based on a Mixture-of-Experts (MoE) architecture. It incorporates the Moonlight [21] MoE language model and the high-resolution MoonViT [30] vision encoder.
- **Qwen2.5-VL-7B-Insturct, Qwen2.5-VL-72B-Insturct** [2] are part of the Qwen2.5-VL series, which combine the Qwen2.5 [33] language model with a redesigned Vision Transformer (ViT) architecture for enhanced visual grounding and understanding.

B.3 Hardware Usage

Model training is conducted under 8 L20 80 GiB GPUs or 4 A800 80 GiB GPUs. Model is evaluated under 4 L20 80 GiB GPUs.

C More Empirical Results and Analyses

In this section, we supply more comparison results, and more analyses for impact of data scale on model performance.

	#Params	Frames	VSI-Bench	STI-Bench		SPAR-Bench			Avg. Token	VM	TC	LV	Avg. Tokens
				Overall	SR_sub	Overall	Single-view	Multi-view					
Qwen2.5-VL-3B-Instruct [2]	3B	32	-	-	-	-	-	-	-	-	-	-	-
+ non-think	-	-	26.7	36.7	37.5	25.4	25.3	25.5	-	52.4	65.7	49.7	-
+ think	-	-	25.9 (↓ 0.8)	34.3 (↓ 2.4)	37.2 (↓ 0.3)	26.8 (↑ 1.4)	26.6 (↑ 1.3)	26.9 (↑ 1.4)	66.9	52.0 (↓ 0.4)	63.2 (↓ 2.4)	49.0 (↓ 0.7)	0.3
SpaceR-Tiny SFT	3B	32	-	-	-	-	-	-	-	-	-	-	-
+ non-think	-	-	34.8	33.0	36.5	24.8	24.5	24.9	53.4	63.8	50.7	-	-
SpaceR-Tiny SG-RLVR	3B	32	-	-	-	-	-	-	-	-	-	-	-
+ non-think	-	-	40.5	36.6	38.7	30.1	30.7	29.6	-	52.9	66.4	50.1	-
+ think	-	-	41.2 (↑ 0.7)	37.8 (↑ 1.2)	40.1 (↑ 1.4)	30.9 (↑ 0.8)	31.4 (↑ 0.7)	30.6 (↑ 1.0)	274.2	51.6 (↓ 1.3)	65.4 (↓ 1.0)	49.4 (↓ 0.7)	237.1
Qwen2.5-VL-7B-Instruct [2]	7B	32	-	-	-	-	-	-	-	-	-	-	-
+ non-think	-	-	34.4	34.5	32.3	33.8	36.9	31.8	-	56.3	71.1	53.5	-
+ think	-	-	30.2 (↓ 4.2)	33.2 (↓ 1.3)	34.4 (↑ 2.1)	31.6 (↓ 2.2)	31.2 (↓ 5.7)	31.8 (- 0.0)	104.0	54.0 (↓ 2.3)	68.1 (↓ 3.0)	46.6 (↓ 6.9)	68.6
SpaceR SG-RLVR	7B	32	-	-	-	-	-	-	-	-	-	-	-
+ non-think	-	-	45.0	36.7	34.8	36.1	36.2	36.0	-	57.9	71.4	54.6	-
+ think	-	-	45.6 (↑ 0.6)	37.0 (↑ 0.3)	38.7 (↑ 3.9)	37.6 (↑ 1.5)	38.2 (↑ 2.0)	37.1 (↑ 1.1)	345.6	56.4 (↓ 1.5)	70.0 (↓ 1.4)	51.7 (↓ 2.9)	265.7

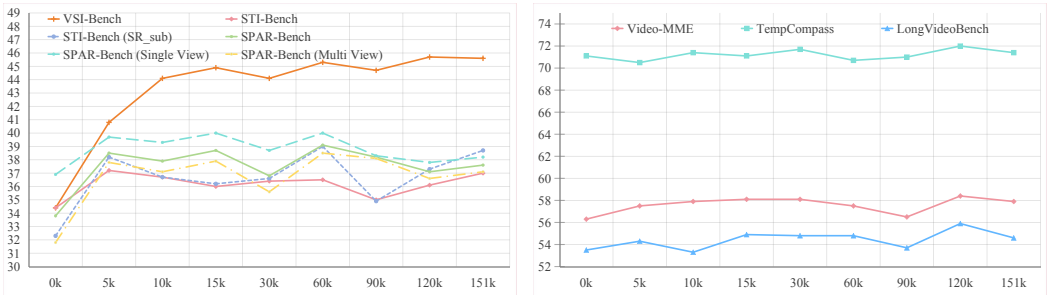
Table 3: We compare non-think and think modes of SpaceR-Tiny, Qwen2.5-VL-3B-Instruct, SpaceR, Qwen2.5-VL-7B-Instruct on spatial reasoning benchmarks (VSI-Bench, STI-Bench, and SPAR-Bench), and video understanding benchmarks: **VM** (Video-MME), **TC** (TempCompass), and **LV** (LongVideoBench). non-think means directly outputting answers, while think refers to explicitly engaging thinking process during inference. **Avg. Tokens** denotes to the tokens number of thinking process in the generated responses.

C.1 Impact of thinking on Model Performance

To assess the effect of explicitly engaging the thinking process during inference, we compare model performance under two modes: *non-think*, which outputs answers directly, and *think*, which includes a structured reasoning process. As shown in Table 3, models not explicitly trained to reason, such as Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, and SpaceR-Tiny SFT, exhibit a significant performance drop across most benchmarks in *think* mode. This degradation is expected, since these models lack sufficient reasoning capability and often generate shorter, less informative reasoning traces, which could be reflected by their lower average token counts. In contrast, our SpaceR SG-RLVR demonstrates consistent gains on spatial reasoning benchmarks when utilizing the *think* mode. This indicates that our SG-RLVR method helps the model develop useful reasoning strategies. However, the same trend does not extend to video understanding tasks, where SpaceR shows a slight performance drop in *think* mode. This suggests that for such tasks, unnecessary or inaccurate reasoning may introduce noise, hindering model predictions. This raises a critical question: *Does Thinking Really Helps?* We hypothesize two contributing factors: 1) While our SG-RLVR framework effectively optimizes for accurate answers, it does not provide direct supervision for the reasoning process itself, leading to suboptimal or inconsistent reasoning traces; 2) For some tasks like video understanding, reasoning may be redundant or even misleading, especially when the question can be answered directly and incorrect reasoning introduces spurious information. These observations highlight a valuable direction for future work: enabling models to decide *when to think* and *how to think accurately*. Developing mechanisms to supervise or adaptively trigger the reasoning process could further enhance performance and interpretability in MLLMs.

C.2 Impact of Model Size on Performance.

To examine the scalability of our SG-RLVR framework across different model sizes, we fine-tune both Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct [2] on the SpaceR-151k dataset using both supervised fine-tuning (SFT) and our proposed SG-RLVR method. As shown in Table 3, the base models exhibit limited spatial reasoning capabilities, with noticeable performance drops in the *think* mode, where they, especially Qwen2.5-VL-3B-Instruct, struggle to produce coherent reasoning and often revert to direct answer outputs. Although SFT leads to modest gains on benchmarks such as VSI-Bench, it fails to enhance generalization and does not substantially improve reasoning ability. In contrast, our SpaceR-Tiny SG-RLVR and SpaceR SG-RLVR consistently outperform the base models and their SFT counterparts across multiple spatial reasoning benchmarks, particularly in the *think* mode. Meanwhile, they show promising generalizability on video understanding benchmarks. These results confirm both the effectiveness and scalability of the SG-RLVR framework in enhancing spatial reasoning across model sizes.



(a) Spatial Reasoning Benchmarks. (b) Video Understanding Benchmarks.

Figure 8: Performance variations with progressively increasing data scale on spatial reasoning and video understanding benchmarks.

C.3 Impact of Data Scale on Model Performance.

To examine the relationship between data scale and model performance, we train SpaceR on six progressively larger subsets of the SpaceR-151k dataset. The results, presented in Figure 8, reveal two key observations. First, SpaceR demonstrates notable performance improvements on spatial reasoning benchmarks such as VSI-Bench and SPAR-Bench (Single View) even with small-scale subsets (e.g., 5k–15k), highlighting the strong data efficiency of our SG-RLVR training paradigm. Second, while performance generally plateaus or slightly fluctuates as the data scale increases to 30k, a substantial jump is observed when training on the full dataset (151k), especially for VSI-Bench, where accuracy surpasses 45%. These findings suggest that our method not only benefits from larger training sets but also exhibits strong generalization from limited supervision, underscoring its effectiveness in both low-resource and full-scale settings.

D Documentation and Licensing

The SpaceR-151k dataset includes annotations in JSONL format, along with associated images and videos. The SpaceR model adopts the same architecture as Qwen2.5-VL-7B-Instruct. Both the dataset and the model are released under the CC BY-NC 4.0 license[6] and are intended for academic research purposes only. Additionally, the ScanNet [5] dataset has been released under MIT license.

[6] `https://creativecommons.org/licenses/by-nc/4.0/`