Improved Visual-Spatial Reasoning via R1-Zero-Like Training

Anonymous ICCV submission

Paper ID 7

Abstract

Increasing attention has been placed on improving the reasoning capacities of multi-modal large language models (MLLMs). As the cornerstone for AI agents that function in the physical realm, video-based visual-spatial intelligence (VSI) emerges as one of the most pivotal reasoning capabilities of MLLMs. This work conducts a first, in-depth study on improving the visual-spatial reasoning of MLLMs via R1-Zero-like training. Technically, we first identify that the visual-spatial reasoning capacities of small- to medium-sized Qwen2-VL models cannot be activated via Chain of Thought (CoT) prompts. *We then incorporate Group Relative Policy Optimization (GRPO) training for improved visual-spatial reasoning, using the carefully curated **VSI-100k** dataset, following DeepSeek-R1-Zero. During the investigation, we identify the necessity to* keep the KL penalty (even with a small value) in GRPO. *With just 120 GPU hours, our **vsGRPO-2B** model, fine-tuned from Qwen2-VL-2B, can outperform the base model by 12.1% and surpass GPT-4o. Moreover, our **vsGRPO-7B** model, fine-tuned from Qwen2-VL-7B, achieves performance comparable to that of the best open-source model LLaVA-NeXT-Video-72B. Additionally, we compare vsGRPO with supervised fine-tuning and direct preference optimization baselines in terms of both spatial and general abilities. Our observations indicate that GRPO training achieves significant performance superiority while effectively balancing spatial and general capabilities.*

1. Introduction

Recently, Large Language Models (LLMs) [1–3, 11, 32] have emerged as a significant advancement in AI. These models take text as input and generate corresponding responses, demonstrating strong generalization capabilities in various language tasks. To extend their capabilities to the multimodal domain, researchers have developed Multimodal Large Language Models (MLLMs) [9, 23, 30, 34, 40, 48] based on pre-trained LLMs. They not only take text but also images and videos as inputs and generate textual responses, serving as the foundation for various applications, including multi-modal understanding [23, 34], visual language agents [15, 36], autonomous driving [26, 37], etc.

The exhaustive understanding of multi-modal observations hinges on advanced reasoning capability, which has spurred growing interest in investigating reasoning mechanisms within MLLMs. This trend mirrors concurrent advancements in vanilla LLMs [22, 35, 38, 44]. As the foundation for AI agents [5, 42] operating in the physical world, the video-based visual-spatial reasoning stands out as one of the most crucial capacities of MLLMs, enabling applications in embodied AI [20, 21] and autonomous driving. However, VSI-bench [41] demonstrates that, although an MLLM exhibits strong video understanding and linguistic reasoning capabilities, its spatial reasoning abilities are still developing, and even the state-of-the-art models lag significantly behind human performance. This highlights the necessity of enhancing the spatial reasoning capabilities of MLLMs.

This work conducts a systematic study on improving the visual-spatial reasoning capacities of MLLMs based on R1-Zero-like training. Focusing on the Qwen2-VL models [34], we first perform an initial study to evaluate the performance of Qwen2-VL regarding whether simple reasoning-oriented prompts can activate the visual-spatial reasoning capacities. We are concerned about Chain of Thought (CoT) strategies [38] due to their previous success. However, we find that vanilla non-CoT prompts perform the best for small- to medium-sized Qwen2-VL on VSI-bench, compared to various CoT ones. This exposes the issue that such models cannot trade inference FLOPs for improved visual-spatial reasoning.

We naturally choose to improve MLLMs on visual-spatial reasoning data. Although some prior works have proposed visual-spatial datasets, they primarily focus on the image domain [6, 7, 23]. As a result, there is an urgent need for a large-scale video-based visual-spatial dataset to facilitate advancements in training MLLMs that can understand and reason about dynamic visual content. Considering this, we construct a video-based question answering dataset of more than 100k samples, **VSI-100k**, following the protocol of VSI-bench. Specifically, we leverage ScanNet [10] to get high-fidelity video scans accompanied by meticulous object-level 3D annotations, based on which (question, answer) pairs regarding spatial information can be easily crafted.

Given the dataset, we follow the journey of DeepSeek-R1-Zero [14], which has demonstrated that simple rule-based reinforcement learning can activate LLMs to autonomously develop complex reasoning capabilities. We decide to extend Group Relative Policy Optimization (GRPO) to improve the visual-spatial reasoning capacities of Qwen2-VL. Specifically, following common practice [8, 14, 47], we define the rule-based reward function based on the alignment between the model prediction and the ground-truth answer to perform GRPO [29]. We also include a format reward when trying to activate the CoT reasoning behavior.

Experiments using GRPO on VSI-100k turns the pretrained Qwen2-VL-2B model into the performant **vsGRPO-2B** within just 120 GPU hours. We observe that vsGRPO-2B outperforms the base model by 12.1% and even surpasses GPT-4o. The same pipeline also transforms the Qwen2-VL-7B model into **vsGRPO-7B**, achieving performance similar to that of the best open-source model with 72B parameters. During GRPO training, we have identified the necessity to *keep the KL penalty (even with a small value) in the training of GRPO* and observed phenomena such as reward hacking. We also compare GRPO with supervised fine-tuning (SFT) and direct preference optimization (DPO) [27] in terms of both spatial and general abilities, and confirm the superiority of GRPO in improving the visual-spatial reasoning capacities of Qwen2-VL while balancing between these two aspects.

In summary, our key contributions are listed as follows:

- We propose VSI-100k, the first dataset of video-based visual-spatial question-answer pairs and will release it to contribute to the spatial understanding and reasoning of MLLMs.
- We apply GRPO to enhance visual-spatial understanding in Qwen-VL, with vsGRPO-2B outperforming GPT-4o and vsGRPO-7B achieving competitive results with the top open-source 72B model.
- We identify the necessity of KL penalty and the reward hacking phenomenon in GRPO training, providing a thorough comparison with other methods like SFT and DPO.

2. Related Works

2.1. Multimodal Large Language Models

In recent years, large language models (LLMs) such as GPT-4 [2], LLaMA [32], and Qwen [1] have demonstrated impressive text generation capabilities. They leverage vast amounts of data in training to produce coherent and contextually relevant text. Building on the architecture of LLMs, multimodal large language models (MLLMs) have emerged, enabling the processing of various input modalities, primarily images and videos. This capability allows MLLMs to excel in vision tasks, bridging the gap between textual and visual information. Pioneering models such as the LLaVA series [19, 23], the Qwen-VL series [4, 34], and the InternVL series [9, 48] have driven this advancement, demonstrating significant improvements in tasks like image captioning, cross-modal retrieval, visual language agents [15, 36] and autonomous driving [26, 37].

2.2. Multimodal Large Language Model Reasoning

Inspired by recent advancements in LLM reasoning [22, 33, 38], numerous studies have aimed to enhance the reasoning capabilities of MLLMs [25, 43, 46]. For instance, some previous works [40] have developed SFT datasets that incorporate step-level reasoning, yet these datasets often lack sufficient human involvement, which limits their effectiveness in addressing complex reasoning tasks. Additionally, some concurrent studies [12, 45] utilize group relative policy optimization from DeepSeek-R1 [14] to further bolster general reasoning abilities. Distinct from these works, our study focuses on the spatial understanding and reasoning capabilities of MLLMs and provides a thorough analysis of various training methods.

2.3. Spatial Understanding and Reasoning

With the emergence of AI agents [5, 42], spatial understanding and reasoning have gained even more significance in multi-modal large language models (MLLM), proving to be valuable across various fields. In response, VSI-bench [41] is proposed to measure these abilities, focusing on aspects such as spatial relations and object size. For training, some prior works have proposed related datasets. For example, VSR [23] offers a dataset focused solely on spatial relations, while SpatialVLM [7] presents an internet-scale 3D spatial reasoning dataset in metric space, though it is not publicly available. More recently, SAT [28] introduces a simulated spatial aptitude training dataset, synthesized using a photo-realistic physics engine. However, they face challenges due to their reliance on image-based data, issues of validity, or synthetic nature. A concurrent work, VisualThinker-R1-Zero [47], shares a similar idea but is based on the SAT dataset and overlooks video-based visual-spatial abilities. Our work is grounded in constructing data from real-world dynamic scenarios and leverage it to improve video-based visual-spatial ability.

3. Can Visual-spatial Reasoning Capacities Be Activated by Prompting?

We initiate by evaluating Qwen2-VL [34] on the VSI-bench [41] with various prompting strategies.

Table 1. Quantitative comparisons of different prompting strategies on Qwen2-VL-2B and Qwen2-VL-7B on VSI-bench.

Backbone	Methods	Avg	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan.	Appr. Order
Qwen2-VL-2B	Think-mode	22.9	18.4	4.3	31.5	17.3	28.3	22.9	26.2	16.8
	Observe-mode	21.8	16.8	1.7	32.7	22.7	28.8	27.6	26.2	18.1
	Vanilla-mode	23.3	21.4	3.4	32.3	31.1	26.7	27.7	24.7	18.9
Qwen2-VL-7B	Think-mode	31.3	44.8	26.1	25.3	23.4	34.7	30.9	32.9	31.5
	Observe-mode	32.0	29.9	19.0	39.6	32.0	34.6	40.0	36.0	24.4
	Vanilla-mode	32.2	39.4	25.0	25.8	43.2	32.6	30.9	27.8	32.6

Concretely, the VSI-bench includes two types of question-answer problems:

- Numerical Answer (NA), including tasks such as object count, absolute distance measurement, object size evaluation, and room size assessment;
- Multiple-Choice Answer (MCA), including tasks related to relative distance, relative direction, route planning, and appearance order.

To evaluate the reasoning capacities of Qwen2-VL on this dataset, we consider two CoT prompting strategies: the widely adopted think-mode, where the model first thinks and then replies to the question, and the observe-mode, where the model first observes the input video and then replies. The latter follows a human-like pattern and has been explored in related works [39]. We also include a non-CoT vanilla-mode, the default mode in the original evaluation, for comparison. Here is a summarization of them:

- Think-mode: `Let's think step by step and then answer the question using a single word or phrase.`
- Observe-mode: `Please observe the video first and then answer the question using a single word or phrase.`
- Vanilla-mode: `Please answer the question using a single word or phrase.`

CoT prompting is ineffective for small- to medium-sized Qwen2-VL on VSI-bench. As shown in Table 1, despite longer responses, think-mode and observe-mode underperform the simple vanilla-mode. Namely, small- to medium-sized Qwen2-VL cannot trade inference FLOPs for improved visual-spatial reasoning.

We visualize some output examples given by Qwen2-VL-2B in Figure 1. We see that the model can actually understand the instructions for activating thinking, but the final answer is still wrong, the same as that of the vanilla prompting. From the exposed chain of thoughts, we realize that the error may arise from the failure to perceive the sofa in the video.

4. R1-Zero-like Training for Visual-spatial Reasoning

Given the above observations, we realize it is necessary to fine-tune the Qwen2-VL models for improved visual-spatial reasoning. Typically, we opt to focus on Group Relative Policy Optimization (GRPO) [29] given its success in building DeepSeek-R1-Zero.

4.1. Training Data Construction

We first create a video-based question-answering dataset named **VSI-100k** for visual-spatial reasoning. It consists of more than 100k samples and follows the VSI-bench protocol. Specifically, we utilize ScanNet [10] to obtain high-fidelity video scans that come with detailed object-level 3D annotations. These annotations provide rich spatial context, enabling us to generate (question, answer) pairs that are directly related to spatial information.

Specifically, we construct questions regarding seven topics, including object count, relative direction, object size, absolute distance, room size and relative distance. We leave the other two topics in VSI-bench, route planning and appearance order, held out. This corresponds to two reasons: 1) (question, answer) pairs of this topic cannot be simply constructed given rarely the static 3D information, which implies that expensive annotation or complex pipeline can be required; 2) with this, we can test the task generalization ability of the trained models. For the NA type problems, we implement a question template similar to that used in [41]. For the MCA one, we simplify the question format by removing the options. This adjustment enhances the model's capacity to recognize entity correspondence instead of simply matching symbols. Some examples are provided in Figure 2.

We provide a brief introduction to the dataset construction methods for various tasks:

- Object Count: directly leverage the object labels included in the annotations document, yielding a total of 6.4k samples;
- Relative Distance: fix one targeted object and compute

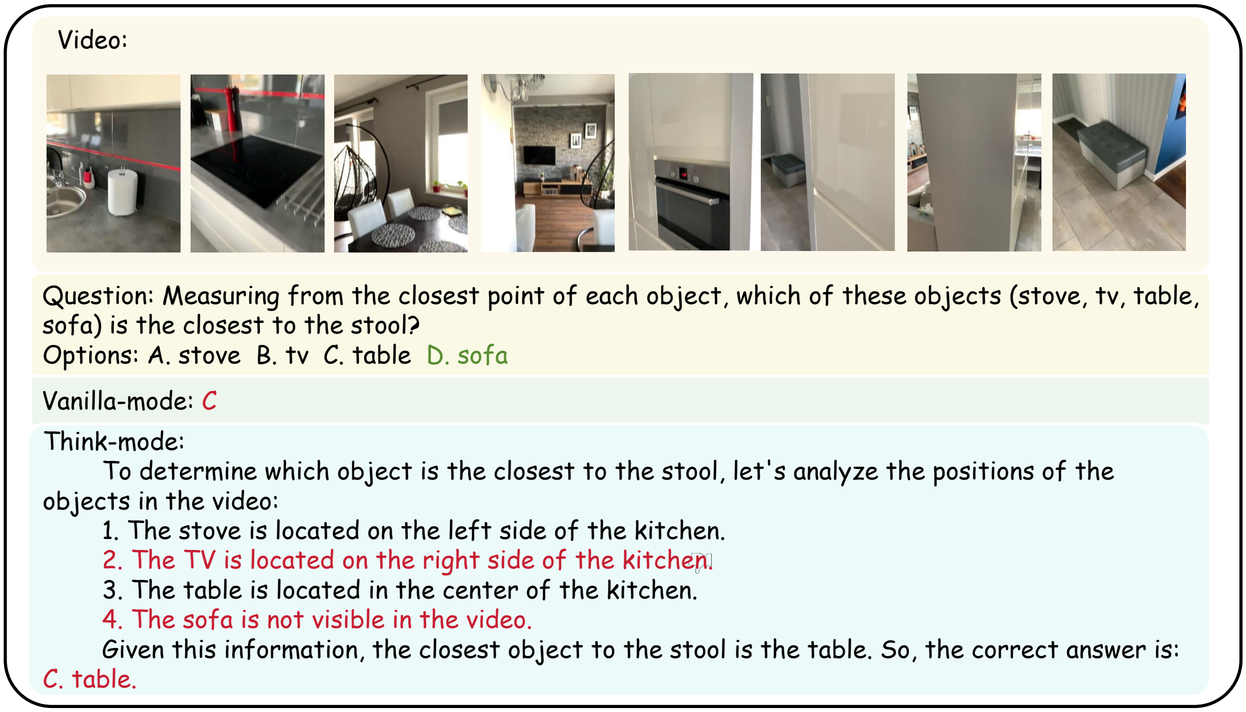

Figure 1. Comparison between the vanilla-mode and think-mode predictions. We use green to label the correct answers and red to indicate incorrect generations. We observe that while using think-mode allows the model to leverage more tokens during generation and initiate the reasoning process in the answer, it still produces the same incorrect response as the vanilla mode. This indicates that Qwen2-VL cannot trade inference FLOPs for improved visual-spatial reasoning.

the absolute distance between it and four other objects to estimate relative distance, yielding 13k samples;

- Object Size: leverage the 3D bounding box to compute the longest dimension of the object, yielding a total of 13k samples;
- Absolute Distance: first remove objects that appear multiple times to ensure specification preciseness, and then calculate the distance between geometric centers of various 3D point-cloud objects, obtaining 75k samples;
- Room size: use the alpha shape algorithm[1] to the total 1.5k scenes, resulting in 1.5k samples.
- Relative Direction: select one object as the front and determine the relative direction of two objects based on their geometric centers of point clouds, getting a total of 8k samples.

4.2. GRPO

GRPO is a type of reinforcement learning (RL) that eliminates the critic model to reduce training costs. Specifically, a group of generated output set $\{o_1, o_2, \cdots, o_G\}$ is sampled for each question q from policy model $\pi_{\theta_{old}}$. Then GRPO optimizes the model π_θ using the following objective:

$$
\begin{aligned}
J_{\mathrm{GRPO}}(\theta) &= \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(O|q)} \\
&\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\frac{\pi_\theta(o_i\mid q)}{\pi_{\theta_{\mathrm{old}}}(o_i\mid q)}\,A_i, \\
&\mathrm{clip}\Big(\frac{\pi_\theta(o_i\mid q)}{\pi_{\theta_{\mathrm{old}}}(o_i\mid q)}, 1-\varepsilon,\, 1+\varepsilon\Big)A_i\Big) \\
&-\beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big)\Bigg],
\end{aligned}
\tag{1}
$$

The reward r guides the direction of the training process and is crucial. We adhere to [8, 24] of using format rewards and accuracy rewards, but with necessary modifications.

Format Reward. Although the CoT prompts are useless for the small-sized Qwen2-VL-2B in inference time, we still wonder if training with them is beneficial for GRPO. As a result, following recent progress in the community, we consider three training prompts for GRPO:

- Think-mode: `Please think step by step and enclose your thinking process in <think> </think> tags and then provide the short answer with one or two words or a number in <answer> </answer>.`

[1] https://en.wikipedia.org/wiki/Alpha_shape

ICCV #7

Figure 2. Illustrations of the VSI-100k. We focus on six subtasks: object counting, relative direction, object size, absolute distance, room size, and relative distance. For each question type, we present one example, accompanied by a video.

- Observe-mode: `Please observe carefully and analyze what you see that helps you to solve the question in the video and enclose it in <observe> </observe> tag, and then provide the short answer with one or two words or a number in <answer> </answer>.`
- Vanilla-mode: `Please provide the short answer with one or two words or a number.`

The format reward quantifies how the responses follow the specified format. It returns a score of 1 or 0. Note that such a reward is omitted for the vanilla-mode.

Accuracy Reward. In the case of non-NA tasks, we employ a character matching method to assess accuracy, awarding a score of 1 for a match and 0 for a mismatch. For NA tasks, we develop a function that computes the absolute difference between the true value and the predicted one and divides the result by the minimum of the two values.

Experimental Settings. Unless specified otherwise, we use Qwen2-VL-2B/7B as the base models due to resource constraints. For training, we employ LoRA [16] training with a learning rate of 10^{-5} for Qwen2-VL-2B with only 120 A800 GPU hours and 5×10^{-6} for Qwen2-VL-7B. To ensure robust performance, we conduct 14 rollouts for each question, allowing for a thorough exploration of possible responses. We set the default sampling temperature to 1 to maintain a balance between creativity and coherence in the generated outputs. The KL divergence coefficient β is configured to 0.0001 as the default setting.

4.3. Results and Analyses

4.3.1. Main Results

Let vsGRPO-T, vsGRPO-O, and vsGRPO-V denote the GRPO training on VSI-100k with prompts of think-mode, observe-mode, and vanilla-mode respectively. We evaluate them with the corresponding test prompts by default. Given the studies in the previous section, we also test the trained models with vanilla-mode prompts.

As shown in Table 2, for models based on Qwen-VL-2B, all GRPO fine-tuned models improve over the baseline. Besides, for the models trained with CoT prompting strategies, their CoT test performance outperforms vanilla one. This indicates that GRPO training can effectively enhance the model's long reasoning capabilities. We present an example in Figure 3. We notice that, despite receiving explicit instructions to utilize CoT reasoning and appropriate tags, Qwen2-VL-2B struggles to adhere to them and tends to provide short answers. With think-mode prompting trained using GRPO, the model can correctly apply the tags and initiate a more structured reasoning process. This also proves the effectiveness and stability of GRPO in enhancing the model's reasoning capabilities.

Table 2. Quantitative results on VSI-bench. vsGRPO-T, vsGRPO-O, and vsGRPO-V refer to GRPO training on VSI-100k with prompts of think-mode, observe-mode, and vanilla-mode respectively. V, T, and O in the Eval. Mode column refer to using vanilla-mode, think-mode, and observe-mode prompts for evaluation, respectively. We also present the best performance of open-source models under the specific model size, like LLaVA-NeXT-Video [19] (LNV for short) and InternVL2 [9] (IVL2 for short), and close-source ones like GPT-4o [18] and Gemini-1.5 Pro [31].

Methods	Eval. Mode	Avg	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Open-source										
Qwen2-VL-2B	V	23.3	21.4	3.4	32.3	31.1	26.7	27.7	24.7	18.9
+ SFT	V	29.6	29.6	23.5	47.4	33.5	26.9	28.3	28.8	18.6
+ DPO	V	23.9	21.7	3.7	34.8	32.4	27.1	28.5	24.2	18.6
+ vsGRPO-T	V	26.1	24.7	10.7	37.4	36.2	27.3	29.5	25.7	17.9
+ vsGRPO-O	V	28.0	26.2	16.4	44.8	38.2	27.0	29.3	24.2	18.2
+ vsGRPO-T	T	29.6	35.0	28.2	34.7	25.2	28.0	38.5	28.5	18.7
+ vsGRPO-O	O	31.2	34.6	22.5	44.8	33.7	29.4	41.8	26.8	15.8
+ vsGRPO-V	V	35.4	53.6	29.0	52.7	43.4	28.1	30.9	26.8	18.9
Qwen2-VL-7B	V	32.2	39.4	25.0	25.8	43.2	32.6	30.9	27.8	32.6
+ SFT	V	38.1	44.7	27.6	46.1	50.4	34.0	35.7	33.0	33.4
+ DPO	V	32.6	39.1	25.2	26.5	44.2	32.6	30.9	29.3	33.3
+ vsGRPO-V	V	40.7	**59.9**	**29.6**	50.8	48.3	35.4	35.6	34.0	31.5
IVL2-2B	V	27.4	21.8	24.9	22.0	35.0	33.8	44.2	30.5	7.1
LNV-7B	V	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
IVL2-40B	V	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LNV-72B	V	**40.9**	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Close-source										
GPT-4o	V	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Pro	V	**48.8**	49.6	28.8	**58.6**	**49.4**	**46.0**	**48.1**	**42.0**	**68.0**

Notably, directly applying the vanilla-mode prompting strategy yields the best performance improvements, particularly for NA questions, and even outperforms GPT-4o. We refer to this model as **vsGRPO-2B** by default. This underscores the conclusion that CoT prompting is ineffective for the small-sized Qwen2-VL-2B on the VSI-bench.

In terms of Qwen2-VL-7B, we only tried vsGRPO-V considering the above results. We observe that vsGRPO-V performs the best on two subtasks—object count and absolute distance. Moreover, the test performance on the Route Planning is also improved, similar to the 2B case. This is possibly because the Route Planning can be divided into sub-tasks that include relative direction, indicating inter-task generalization. With only 7B model size, we note that our model shows performance comparable to that of the leading open-source model, LLaVA-NeXT-Video-72B [19].

4.3.2. Importance of KL Penalty

The KL penalty term plays a crucial role in regulating the divergence between the online policy and the frozen reference one. It avoids the model straying too far from the initial point. While some works [24, 45] advocate for removing the KL penalty to enhance performance, we have observed that doing so can easily lead to training collapse, as illustrated in Figure 4. In contrast, introducing a positive β (even very small, such as 0.0001) can effectively address this issue. This may be attributed to the specific nature of VSI reasoning problems.

4.3.3. Reward Hacking

During training, we observed that the model occasionally discovers strategies to achieve high rewards that do not align with our original intentions. For example, when training in think-mode, some rollouts share the structure: <think> There are [number] [object] in the room</think> <answer>[number]</answer> in the object count task. We also encountered some extreme samples in the rollouts, such as <think> </think> <answer>xx</answer>. While this format is technically correct, it represents a failure to demonstrate a

Figure 3. A comparison between Qwen2-VL-2B and its trained variant, vsGRPO-T-2B. We select the route planning question type, given that the training dataset does not include this category.

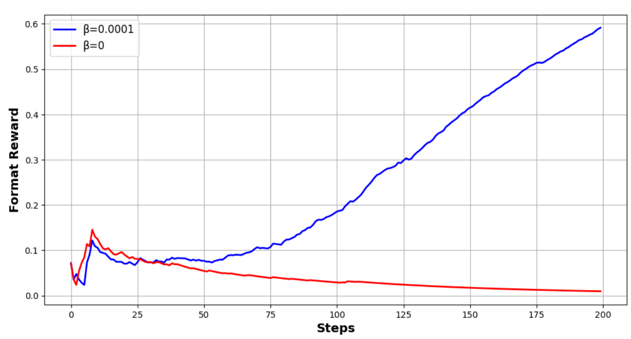

Figure 4. The format reward curve of $\beta = 0$ and $\beta = 0.0001$ during training.

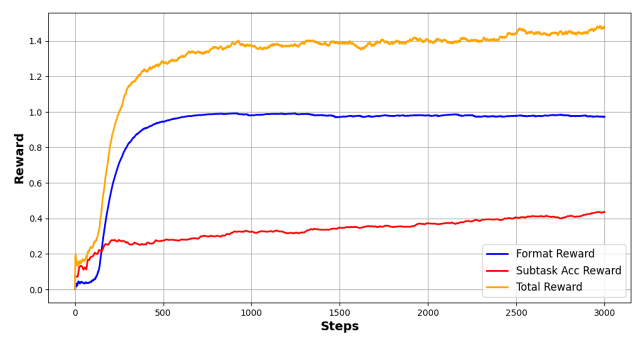

Figure 5. The curve of the format reward (blue), the accuracy reward curve of one subtask of VSI-100k (red), and the total reward curve (yellow) during GRPO training.

genuine thinking process. This phenomenon is reminiscent of observations made in VisualThinker-R1-Zero [47]. In response to this issue, we decided to incorporate a length reward function as a mitigation strategy. However, we soon realized that some new generations merely added extra <think></think> and <answer></answer> tags to exploit the length reward. This approach also does not contribute to a meaningful thinking process. So, more reasonable reward functions should be explored.

4.3.4. Dynamics of Various Rewards

As shown in Figure 5, during the GRPO training, the format reward converges to 1 quickly, while the accuracy reward increases more slowly. This phenomenon suggests that the format is more easily learned than spatial knowledge. We also notice that there appears to be an upper bound for the accuracy reward, indicating that although GRPO can enhance spatial reasoning abilities, there may be inherent limitations to the extent of improvement achievable. This suggests the need for further exploration into alternative methods or additional training paradigms that could push these boundaries and lead to even greater advancements in spatial reasoning tasks.

4.3.5. Comparison to Other Training Approaches

We also compare our approach with commonly used fine-tuning algorithms, SFT and DPO [27], in Table 2. For SFT, we directly use the constructed VSI-100k for tuning. For

Table 3. General understanding and reasoning ability with different training methods. We use VideoMMMU (16 frames) as the video reasoning benchmark and VideoMME as the video general benchmark.

Metrics	VideoMMMU		VideoMME	
	Qwen2-VL-2B	Qwen2-VL-7B	Qwen2-VL-2B	Qwen2-VL-7B
Base	32.3	43.1	49.6	50.9
SFT	30.7	42.7	49.3	41.8
DPO	32.2	43.7	48.8	50.9
vsGRPO-V	31.7	43.0	49.1	45.6

DPO, the correct answer is modified to a wrong one to serve as the less-preferred answer. As shown, the two approaches both improve over the base model on the VSI-bench, but still lag behind vsGRPO-V. Besides, the improvement of DPO is minor, which is perhaps because of the sub-optimal preference pair construction.

In addition to spatial understanding and reasoning abilities, we also consider general understanding and reasoning abilities when training with different methods using visual-spatial data. To evaluate this, we leverage VideoMMMU [17] and VideoMME [13] benchmarks following the setting in [12], shown in Table 3. Combining with the results in Table 2, we find that, although DPO training does not result in a significant decline in general understanding and reasoning abilities, it fails to improve spatial ability. This suggests that using DPO training to inject new knowledge into the models may not be the most effective approach for enhancing spatial reasoning skills. In contrast, when applying SFT and GRPO, while there is a notable improvement in spatial ability, there is also a corresponding drop in general understanding and reasoning performance. This trade-off reveals a critical challenge in balancing the enhancement of specific skills with the maintenance of overall cognitive capabilities. Besides, GRPO demonstrates a smaller overall decline in general understanding and reasoning compared to SFT. This finding highlights the advantages of GRPO, particularly when training with visual-spatial data, suggesting that it may better preserve broader general abilities while still achieving targeted improvements in specific abilities like spatial understanding and reasoning.

5. Conclusion

In this work, we center on the video-based visual-spatial intelligence of MLLMs. Using Qwen2-VL as the base model, we identify that the visual-spatial reasoning capacities of Qwen2-VL-2B/7B cannot be activated via CoT prompts. We construct **VSI-100k** to combat data scarcity and adapt GRPO training. Extensive experiments demonstrate that `vsGRPO-2B` and `vsGRPO-7B` outperform models of the same size in spatial understanding and reasoning ability, and also indicate that the GRPO approach achieves greater improvements in visual-spatial ability while experiencing a smaller drop in general ability when trained with domain-specific data in comparison to SFT and DPO.

6. Limitation & Future Work

During the application of GRPO to spatial reasoning, we made necessary modifications to the reward functions derived from the LLM model, DeepSeek-R1, resulting in impressive outcomes. It is natural to introduce additional reward functions specifically tailored for visual and temporal aspects of video-based visual-spatial reasoning, which we did not explore in this study. In the future, we aim to investigate these additional reward functions to further enhance the model's performance in visual-spatial tasks, ultimately leading to more robust reasoning capabilities.

References

[1] Qwen2 technical report. 2024. 1, 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 1

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[5] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. 1, 2

[6] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 1

[7] 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 Com-*

puter Vision and Pattern Recognition, pages 14455–14465, 2024. 1, 2

[8] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02. 2, 4

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024. 1, 2, 6

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 3

[11] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 1

[12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 2, 8

[13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 8

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[15] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024. 1, 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[17] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 8

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6

[20] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 1

[21] Ziming Li, Huadong Zhang, Chao Peng, and Roshan Peiris. Exploring large language model-driven agents for environment-aware spatial interactions and conversations in virtual reality role-play scenarios. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 1–11. IEEE, 2025. 1

[22] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2

[23] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1, 2

[24] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 4, 6

[25] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 2

[26] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 1, 2

[27] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2, 7

[28] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 2

[29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3

[30] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 1

[31] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[33] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Y Wu, and Zhifang Sui. Math-shepherd: A label-free step-by-step verifier for llms in mathematical reasoning. *arXiv preprint arXiv:2312.08935*, 2023. 2

[34] 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. 1, 2

[35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*. 1

[36] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 1, 2

[37] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 1, 2

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1, 2

[39] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024. 3

[40] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 1, 2

[41] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. 1, 2, 3

[42] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. *arXiv preprint arXiv:2502.13130*, 2025. 1, 2

[43] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 2

[44] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023. 1

[45] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 2, 6

[46] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 2

[47] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025. 2, 7

[48] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, 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. 1, 2