Bridging the Dynamic Perception Gap: Training-Free Draft Chain-of-Thought for Dynamic Multimodal Spatial Reasoning

Anonymous ACL submission

Abstract

While chains-of-thought (CoT) have advanced complex reasoning in multimodal large language models (MLLMs), existing methods remain confined to text or static visual domains, often faltering in dynamic spatial reasoning tasks. To bridge this gap, we present GRASSLAND, a novel maze navigation benchmark designed to evaluate dynamic spatial reasoning. Our experiments show that augmenting textual reasoning chains with dynamic visual drafts, overlaid on input images, significantly outperforms conventional approaches, offering new insights into spatial reasoning in evolving environments. To generalize this capability, we propose D2R (Dynamic Draft-Augmented Reasoning), a training-free framework that seamlessly integrates textual CoT with corresponding visual drafts into MLLMs. Extensive evaluations demonstrate that D2R consistently enhances performance across diverse tasks, establishing a robust baseline for dynamic spatial reasoning without requiring model fine-tuning.

1 Introduction

Humans often exhibit effective behavioral strategies that inspire multimodal large language models (MLLMs) (Yang et al., 2023a; Li et al., 2024; Wu et al., 2024; Yao et al., 2024) to tackle complex tasks, particularly in the realm of multimodal reasoning. In such tasks, humans commonly create drafts to support step-by-step thinking when processing visual information that integrates text and imagery. This drafting approach is especially beneficial for extracting insights from dynamic images, where chronological, incremental reasoning is highly effective.

Current MLLMs primarily emphasize step-by-step reasoning patterns or simple visualization techniques, exemplified by methods such as ToT (Yao et al., 2023) and ICoT (Gao et al., 2025), but they lack mechanisms for draft creation based on input images. While these frameworks achieve strong results on textual and static visual tasks (Chen et al., 2024a; Lu et al., 2024; Jiang et al., 2025; Hessel et al., 2022), they often suffer from loss of rich visual information and diminished spatial awareness—factors critical for dynamic multimodal spatial reasoning. Since dynamic spatial reasoning plays a pivotal role in many real-world applications, it is important to investigate how well existing models perform in this domain.

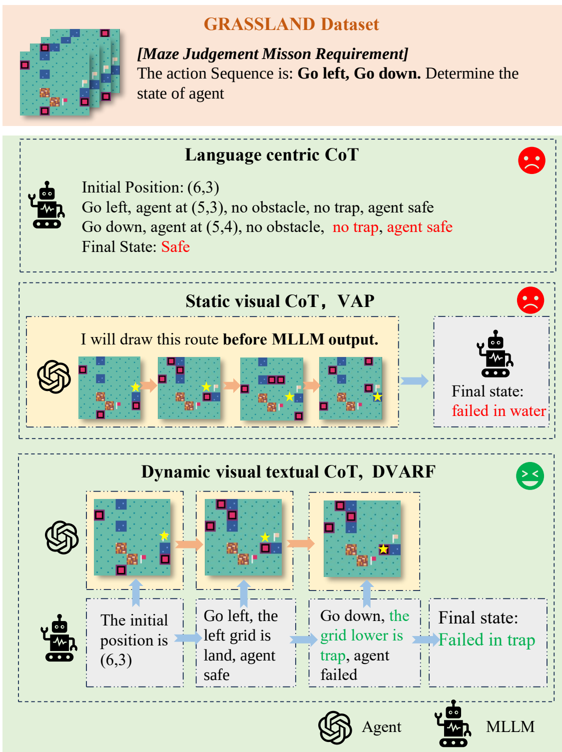

Figure 1: The demonstration of the Draft CoT with D2R. Compared to the spatial information gaps in language-centric CoT, and the incomplete dynamic information in static visual CoT, which only visualizes the input rather than the MLLM's thought process, Draft CoT excels at dynamic spatial reasoning.

To address this, we develop GRASSLAND, a dynamic maze environment modeled as a classi-

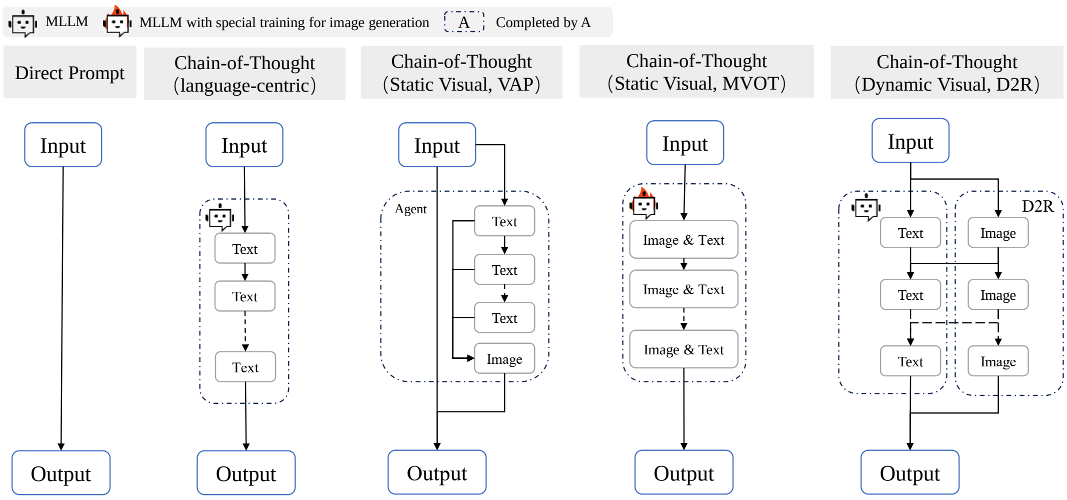

Figure 2: Illustration of the difference between our method and others. Direct prompting and language-centric CoT face significant limitations in dynamic spatial reasoning tasks without images. VAP can only generate static images based on agent prompts, without MLLM involvement for dynamic perception. MVOT requires MLLMs powerful in image generation by training on specialized datasets. In contrast, D2R marks the textual thought in the image as draft and integrates it into the Draft CoT, enhancing the MLLM's dynamic spatial reasoning ability without specific training.

cal pixel grid world with evolving environment grids. We define two dynamic spatial reasoning tasks—Maze Judgment and Maze Navigation—to evaluate models' ability to perform complex visual analysis in changing contexts. As illustrated in Figure 1, our experiments reveal that existing MLLMs and reasoning frameworks struggle with these tasks, often overlooking or misinterpreting spatial context, such as inaccurately judging locations or ignoring special grid features. To overcome these challenges, we propose the Draft Chain-of-Thought (Draft CoT) approach, which integrates textual reasoning with corresponding drafts over dynamic input images. This method significantly outperforms previous approaches, providing fresh insights into dynamic spatial reasoning.

Despite its effectiveness, Draft CoT relies on image generation capabilities not universally available across all MLLMs. To broaden its applicability, we introduce a training-free framework named the **Dynamic Draft Augmented Reasoning Framework** (D2R). As shown in Figure 2, D2R seamlessly integrates both visual and textual inputs, enhancing reasoning by enabling cross-modal information exchange. It first generates a global plan based on the task prompt and tool set, then iteratively performs chronological reasoning by updating textual thoughts as drafts on dynamic images. Finally, D2R signals the MLLM to produce the final output, concluding the iterative process.

Extensive experiments on the two dynamic spatial reasoning tasks demonstrate that D2R surpasses existing text-only and static vision-based reasoning methods. Moreover, tests on multiple MLLMs confirm D2R's ease of transfer, robustness, and broad applicability as a training-free enhancement.

In summary, this paper makes three main contributions:

- **A novel benchmark for dynamic spatial reasoning:** We introduce GRASSLAND, a classical pixel grid world with dynamic environment changes, along with two challenging tasks—Maze Judgment and Maze Navigation—to rigorously evaluate dynamic spatial reasoning capabilities.

- **A new Draft Chain-of-Thought method:** We propose Draft CoT, which combines textual reasoning with corresponding drafts over dynamic input images, significantly improving performance over existing reasoning frameworks on dynamic spatial tasks.

- **A training-free framework for broad applicability:** We develop the Dynamic Draft Augmented Reasoning Framework (D2R) that seamlessly integrates Draft CoT into existing MLLMs without additional training, enabling enhanced dynamic multimodal reasoning across various models.

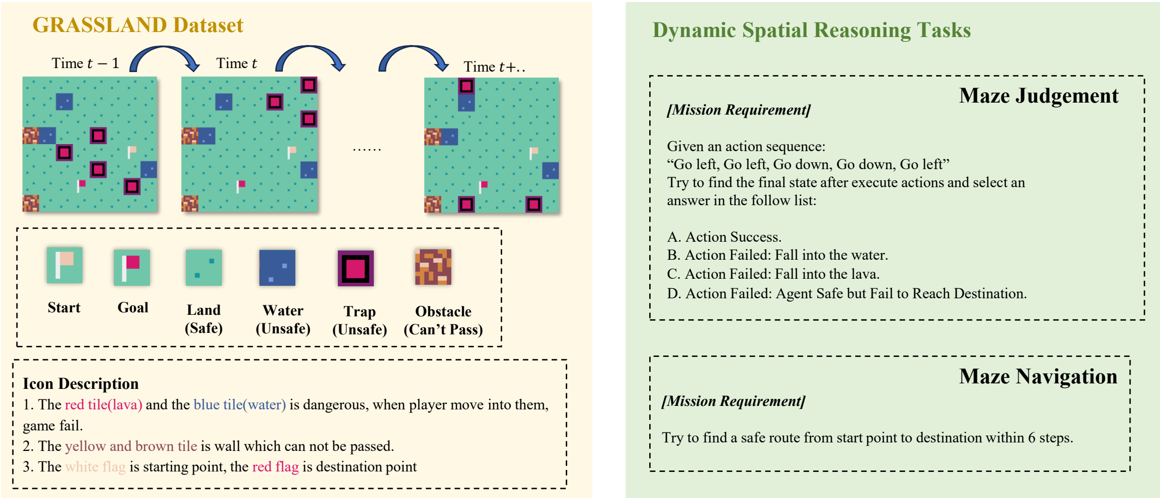

Figure 3: Example of dynamic scenario sequence in GRASSLAND. The left part is the illustration of the dynamic images and grids in GRASSLAND, and the right part is the description of the two tasks.

2 Related Works

2.1 MultiModal Large Language Models

Multimodal Large Language Models (Liu et al., 2025; Xu et al., 2025b; Zhu et al., 2025) have made remarkable progress by integrating various modalities—such as text, images, and video—into a unified framework for understanding and reasoning. In this framework, different modality encoders project inputs into a shared semantic space, which is then processed by a language model to generate responses (Yin et al., 2024). However, most existing MLLMs adopt a unimodal generation strategy: they rely solely on text for auto-regressive response generation, treating non-text modalities merely as auxiliary context during encoding (Liu et al., 2024b). As a result, the rich and dynamic information contained in modalities like images and videos is not fully utilized during the generation process, which significantly limits the model's performance on multimodal reasoning tasks (Liu et al., 2024a). In contrast, OpenAI o3 (OpenAI, 2025) demonstrates the potential of step-by-step generation that jointly conditions on both visual and textual inputs. Unfortunately, current MLLMs are not capable of this generation pattern due to inherent limitations in image processing. In this paper, we propose a Dynamic Draft Augmented Reasoning Framework, which achieves adaptiveness-enhanced reasoning with multiple domain inputs by utilizing external tools to generate a bimodal chain-of-thought.

2.2 MLLMs Reasoning

Multimodal reasoning tasks are designed to evaluate the ability to integrate information from different modalities and perform comprehensive reasoning (Gao et al., 2025; Zheng et al., 2023). The most common method is the language-centric multimodal reasoning pattern, which focuses on extracting information from the visual modality and downscaling it to the linguistic domain for inference (Yang et al., 2023b; Xu et al., 2025a; Mitra et al., 2024). Rather from the language-centric pattern, the collaborative multimodal reasoning introduces the visual domain into the reasoning process, such as VAP (Xiao et al., 2024) and MVoT (Li et al., 2025). However, VAP merely visualizes the input of the model instead of the model's thought process, while MVoT requires the model to generate multimodal output. Both methods overlooks the need to enhance the generalization ability of existing models across multimodal reasoning tasks. In this paper, we propose Dynamic Draft Augmented Reasoning Framework, which enhances the reasoning capabilities of existing MLLMs by realizing bimodal chains of thought through the combination of textual thought and their corresponding drafts in the input images.

3 Dynamic Multimodal Spatial Reasoning

To further evaluate the performance of the existing MLLMs on the dynamic spatial reasoning task, we propose GRASSLAND, a dynamic maze navigation scenario for the dynamic spatial reasoning task.

Task	Maze Judgment			Maze planning		
	easy	normal	hard	easy	normal	hard
Grid Size		7×7			5×5	
Obstacles	0	1	2	1	2	3
Dynamic Trap	2	3	4	1	2	2
Static Trap	0	1	2	0-4	0-4	0-6
Route Length	5.32	6.00	5.67	3.47	3.75	4.34

Table 1: Statistics of the dataset information, covering three levels of complexity in two tasks.

As shown in Figure 3, it simulates a classical pixel grid world W with a start point p_s and destination point p_e. Also, parts of the environment grids contain obstacles('the walls')P_o, dynamic traps ('the lava')P_l, and stationary traps ('the water') P_w. The model is required to determine the next action or state based on the given prompt and scenario.

3.1 Task Formulation

Based on this dataset, we define two scenarios for the dynamic spatial reasoning tasks: Maze Judgment and Maze Navigation. These scenarios require models to analyze time and spatial sequences, locate special objects, make action decisions, and predict states when actions are executed. The details are presented in Table 1.

The Maze Judgment Scenario To assess the ability of MLLMs to perceive dynamic spatial locations, we introduce the maze judgment scenario. In this task, the MLLM must determine the final state based on actions and the map, which are divided into success, failure, and loss. This process is modeled within a discrete state space, S, where each state $s_t \in S$ represents the agent's status at time t. In practice, the model must predict the state in time t defined as s_t, and determine the final state s_{end}, given a world map W and a sequence of actions $R_{\text{action}} = \{r_1, r_2, \ldots, r_T\}$. This process is performed as follows:

$$s_t = f(W, R_{action<t}, S_{<t}) \quad t \in \{1, \ldots, T\}, \tag{1}$$

$$s_{end} = s_T. \tag{2}$$

The Maze Navigation Scenario To examine the ability of MLLM to reason dynamic spatial location, we propose the maze navigation scenario. In this task, the MLLM should reach the destination from the starting point, while avoiding all dangers and doing so as quickly as possible. This route is defined with the current position p_t and next action r_t. In practice, MLLM should lay out a safe route R_{action} that can stay out of danger positions set $P_D = P_l \cup P_w$ (i.e., $\forall t < T, p_t \notin P_D$), and reach the destination p_e within a limited steps L (i.e., $T \leq L$). This process is performed as follows:

$$r_t, p_t = f(W, r_{t-1}, p_{t-1}), \forall t \in \{1, \ldots, T\} \tag{3}$$

$$R_{action} = \{r_t\}_{t=1}^{T} \tag{4}$$

If the agent cannot reach the final destination within a limited steps or fall into the danger set, the agent will be judged as a failure in this case.

3.2 Interesting Findings

Poor abilities of MLLMs To explore the abilities of MLLMs on dynamic spatial reasoning, we measured two tasks on different MLLMs. As shown in Table 2, MLLM exhibits a poor ability to follow the long action sequence and collaborative processing of information across multiple modalities. Among the failed cases, we note that MLLMs often ignore or misjudge the scenario context in their thinking process, such as misjudging the location or ignoring special grids. These findings suggest that current MLLMs lack a robust mechanism for integrating spatial and contextual cues over time.

Limit gains of existing methods To investigate what factors can enhance the dynamic spatial reasoning capabilities of MLLMs, we conduct experiments on the hard judgment task using a variety of methods. As shown in Figure 4, various language-centric Chain-of-Thought approaches yield only marginal performance improvements and in some cases, even underperform compared to the original baseline. On the other hand, incorporating the VAP method with ground-truth positional images fails to improve model effectiveness and instead introduces noise that degrades performance.

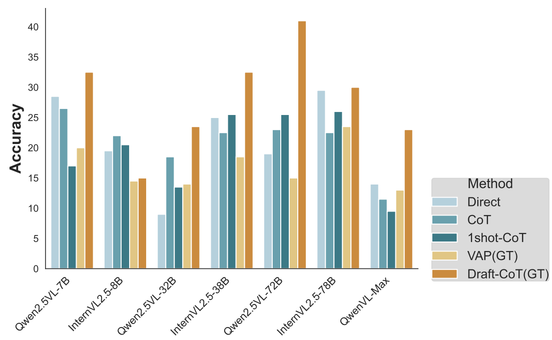

Figure 4: Accuracy with different models and methods in the hard Maze Judgment task. GT denotes that this result is obtained by ground truth in the route.

Model	Maze Judgment			Maze Navigation		
	easy	normal	hard	easy	normal	hard
VideoLLaMA3-7B (Zhang et al., 2025)	18.0	12.5	11.0	1.0	1.5	0.0
Qwen2.5VL-7B (Bai et al., 2025)	22.5	34.0	28.5	1.0	2.0	1.0
InternVL2.5-8B (Chen et al., 2024b)	21.0	18.5	19.5	3.5	1.0	0.5
Qwen2.5VL-32B (Bai et al., 2025)	14.0	9.0	9.0	0.0	0.0	0.0
InternVL2.5-38B (Chen et al., 2024b)	22.5	26.0	25.0	13.5	11.5	3.5
Qwen2.5VL-72B (Bai et al., 2025)	**61.0**	**38.5**	19.0	**31.0**	**21.5**	**6.5**
InternVL2.5-78B (Chen et al., 2024b)	28.5	26.0	**29.5**	15.0	9.0	1.5
QwenVL-Max (Bai et al., 2023)	40.0	21.5	14.0	19.5	10.5	1.5

Table 2: Performance of various models in Maze Judgment task and Maze Navigation task with direct prompt. The best results of each dimension are **bold** and the secondary results are underlined.

These results highlight the limitations of existing approaches and underscore the need for more effective integration of dynamic spatial information during the reasoning process.

Drafts over dynamic images: Bring Surprise Inspired by the previous findings, we introduced visual navigation cues into the dynamic input images and combined them with textual CoT. This method allows the reasoning process to unfold through textual thought with its drafts over dynamic input images, termed as Draft CoT. Specifically, we directly edited the dynamic images by overlaying visual guidelines to depict the path. As shown in Figure 4, this approach significantly improves accuracy across all models, regardless of their underlying reasoning abilities, even outperforming the one-shot CoT setting in average accuracy. Moreover, as shown in Figure 5, the accuracy of all four options improves, rather than just increasing the success rate of a single option, further highlighting the robustness of Draft CoT across all scenarios. These results demonstrate the effectiveness of incorporating corresponding drafts over dynamic input images into the textual CoT process, providing new insights for dynamic spatial reasoning tasks.

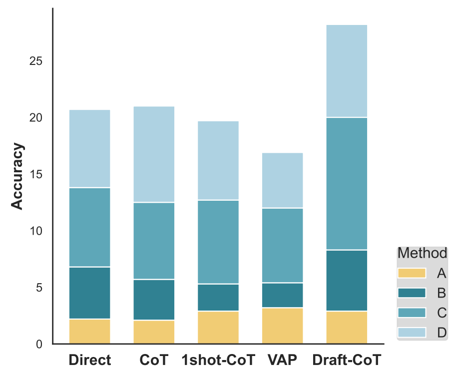

Figure 5: Average accuracy of models for each choice using various methods in the Maze Judgment task.

4 Methodology

Although the Draft CoT can obtain great performance gains, it rely on image generation capabilities not universally available across all the MLLMs. To broaden its applications, we propose the Dynamic Draft Augmented Reasoning Framework (D2R), a training-free framework to generate intermediate thoughts on both textual thoughts and visual drafts. D2R extends the reasoning space from a signal language domain $\mathcal{L}$ to multiple domains $\mathcal{L} \cup \mathcal{V} \cup \mathcal{T}$, where $\mathcal{V}$ represents the visual domain and $\mathcal{T}$ represents the chronology domain. It enables models to reason in dynamic visual information by splitting it into steps and marking drafts over the input images in each step. By combining textual thoughts with corresponding drafts, this novel reasoning paradigm offers a more intuitive and accurate method with enhanced ability to collaborate on details between these two modalities.

4.1 Toolkits for Synthesis and Drafting

Drafting in the visual domain can enhance the ability to reason. However, MLLMs lack the ability to edit dynamic visual information and are weak in long text processing scheduling. Therefore, it is necessary to leverage external toolkits to enhance MLLM's performance. Therefore, we introduce the Dynamic-Information-Extract and Position-Draw tools for visual editing. Additionally, we also introduced an external LLM as a scheduling hub to organize the utilization of those tools.

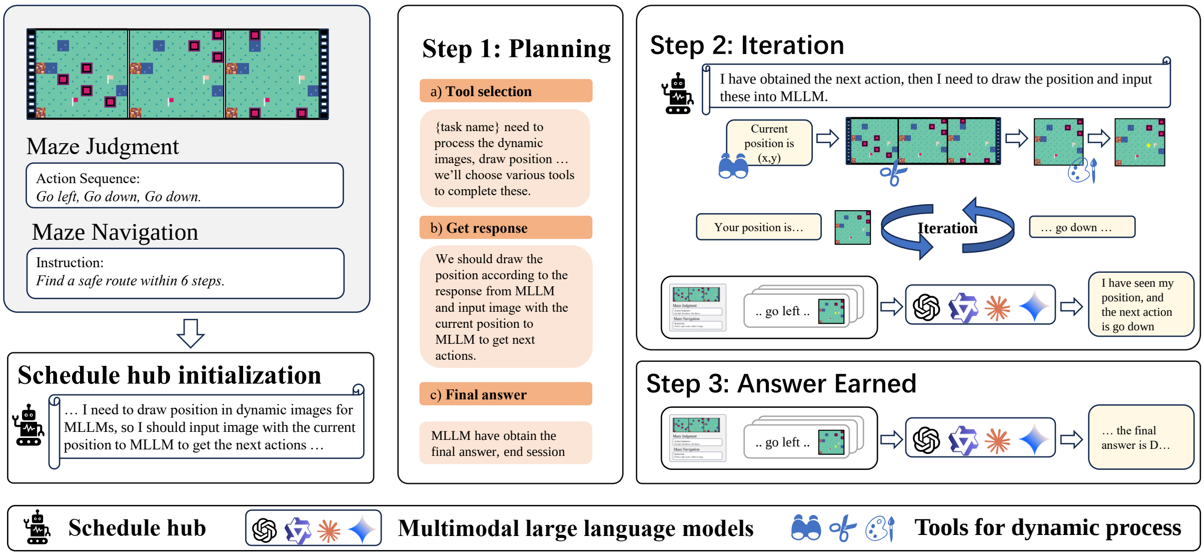

Figure 6: Illustration of D2R reasoning process. After the schedule hub initialization, the process consists of planning, iteration, and answering three parts.

4.2 Procedures of D2R

We analogize D2R's process to an iterative process. Scheduled by the scheduling hub, D2R will autonomously determine the task type and generate a tool invocation plan, and it will maintain a real-time updated draft chain that is continuously supplemented with the most up-to-date information during the iteration process until the answer is generated. The whole process are as follows:

Algorithm 1: Procedures of Dynamic Draft Augmented Reasoning Framework

Input: Text instruction $\mathbb{G}$
Dynamic images $\mathcal{I}$
Output: Final answer $\mathcal{A}$

Initialization:
$\mathcal{D}_p \leftarrow$ Scheduling hub
$\mathbb{E} \leftarrow$ Tool set
$\mathcal{C}_0 \leftarrow \varnothing, n \leftarrow 0$
Step 1: Planning
$\varphi \leftarrow \mathcal{D}_p(\mathbb{G}, \mathbb{E})$
Step 2: Iteration
while *not* $\mathcal{D}_p$ *decides to stop*
do
  $c_n \leftarrow \mathrm{MLLM}(\mathbb{G}, \mathcal{I}, \mathcal{C}_{<n})$
  $e_n \leftarrow \mathcal{D}_p(c_n, \varphi)$
  $C_n \leftarrow e_n(\varphi, \mathcal{I}, c_n)$
  $n \leftarrow n + 1$
Step 3: Final Answer
$\mathcal{A} \leftarrow \mathrm{MLLM}(\mathbb{G}, \mathcal{I}, \mathcal{C}_{\mathrm{all}})$

Step 1: Planning Our method takes a textual instruction $\mathbb{G}$ and dynamic images $\mathcal{I}$ as input. First, we prompt the scheduling hub to schedule a plan φ and select the correct tools e_n from the tool set $\mathbb{E}$. This step can be formalized as shown in Equation 5:

$$\varphi \leftarrow \mathcal{D}_p(\mathbb{G}, \mathbb{E}), \tag{5}$$

where $\mathcal{D}_p$ denotes the scheduling hub in this step.

Step 2: Iterative As shown in Figure 6, after completing planning, D2R invokes the tool e_n to generate the corresponding thought markers in images as drafts and fuse textual thought c_n as augmented perceptual thought $\mathcal{C}_n$. In each iteration, $\mathcal{C}_n$ will be updated as the instruction progresses. The process is formally depicted as follows:

$$\begin{cases} c_n \leftarrow \mathrm{MLLM}(\mathbb{G}, \mathcal{I}, \mathcal{C}_{<n}) \\ \quad e_n \leftarrow \mathcal{D}_p(c_n, \varphi) \\ \quad C_n \leftarrow e_n(\varphi, \mathcal{I}, c_n) \end{cases} \tag{6}$$

where $\mathcal{C}_{<n}$ denotes the set of all the augmented perceptual thoughts before n turns.

Step 3: Final Answer Iteration When the iteration ends, scheduling hub will check the last output c_{last} and determine if the answer $\mathcal{A}$ was generated. If $\mathcal{A}$ was not generated, scheduling hub will repeat the process and change the prompt strategy to instruct MLLM to output the answer As shown in Equation 7, we take the set of all CoT $\mathcal{C}_{all}$ as input and use the prompt in the appendix to arrive at the final answer $\mathcal{A}$.

$$\mathcal{A} \leftarrow \mathrm{MLLM}(\mathbb{G}, \mathcal{I}, \mathcal{C}_{\mathrm{all}}) \tag{7}$$

Model	Method	Total Acc			Average Acc
		Easy	Normal	Hard	
Maze Judgment task					
Qwen2.5VL-7B	Direct	22.5	34.0	**28.5**	28.3
	CoT	18.0(-4.5)	29.0(-5.0)	26.5(-2.0)	24.5(-3.8)
	1-shot CoT	18.0(-4.5)	20.5(-13.5)	17.0(-11.5)	18.5(-9.8)
	VAP	13.5(-9.0)	15.0(-19.0)	20.0(-8.5)	16.2(-12.1)
	D2R (ours)	**34.0**(+11.5)	**46.0**(+12.0)	28.0(-0.5)	**36.0**(+7.7)
Qwen2.5VL-72B	Direct	61.0	38.5	19.0	39.5
	CoT	67.0(+6.0)	40.0(+1.5)	23.0(+4.0)	43.3(+3.8)
	1-shot CoT	**71.0**(+10.0)	46.5(+8.0)	25.5(+6.5)	47.7(+8.2)
	VAP	15.5(-45.5)	20.0(-18.5)	15.0(-4.0)	16.8(-22.7)
	D2R (ours)	67.0(+6.0)	**49.0**(+10.5)	**41.0**(+22.0)	**52.3**(+12.8)
QwenVL-max	Direct	40.0	21.5	14.0	25.2
	CoT	36.0(-4.0)	24.0(+2.5)	11.5(-2.5)	23.8(-1.4)
	1-shot CoT	18.0(-22.0)	17.0(-4.5)	9.5(-4.5)	14.8(-10.4)
	VAP	15.0(-25.0)	9.0(-12.5)	13.0(-1.0)	12.3(-12.9)
	D2R (ours)	**46.5**(+6.5)	**35.5**(+14.0)	**28.0**(+14.0)	**36.7**(+11.5)
Maze Navigation task					
Qwen2.5VL-7B	Direct	1.0	2.0	1.0	1.3
	CoT	1.5(+0.5)	1.5(-0.5)	0.0(-1.0)	1.0(-0.3)
	1-shot CoT	2.5(+1.5)	**4.5**(+2.5)	**2.5**(+1.5)	3.2(+1.9)
	VAP(GT)	-	-	-	-
	D2R (ours)	**4.0**(+3.0)	**4.5**(+2.5)	2.0(+1.0)	**3.5**(+2.2)
Qwen2.5VL-72B	Direct	31.0	21.5	6.5	19.7
	CoT	16.5(-14.5)	17.5(-4.0)	0.5(-6.0)	11.5(-5.2)
	1Shot-CoT	17.5(-13.5)	5.0(-16.5)	1.5(-5.0)	8.0(-11.7)
	VAP(GT)	-	-	-	-
	D2R (ours)	**38.0**(+7.0)	**26.0**(+4.5)	**12.5**(+6.0)	**25.5**(+5.8)
QwenVL-max	Direct	19.5	10.5	1.5	10.5
	CoT	22.5(+3.0)	10.5 (-)	6.0(+4.5)	13.0(+2.5)
	1Shot-CoT	1.0(-18.5)	0.5(-10.0)	0.0(-1.5)	0.5(-10.0)
	VAP(GT)	-	-	-	-
	D2R (ours)	**27.5**(+8.0)	**21.5**(+11.0)	**7.0**(+5.5)	**18.7**(+8.2)

Table 3: Performance of Maze Judgment task and Maze Navigation task. The results in '(·)' represent the delta performance compared to the performance with direct prompt in each task. The best results of each dimension are **bold** and the secondary results are underlined.

5 Experiment

5.1 Experiment Setup

We construct datasets for two dynamic spatial reasoning tasks described in Section 3, encompassing three levels of complexity in environment and action spaces. We use Qwen-Max as the scheduling hub in our work, and the temperature is set to 0.1. We compare the D2R with the following reasoning methods: 1) Direct Prompt. 2)Chain-of-thought (CoT). 3) CoT with 1-shot. 4)VAP. In our experiments, we use Qwen2.5-VL-7B, Qwen2.5-VL-72B, and Qwen-VL-Max as the MLLM part of D2R.

5.2 D2R has better dynamic reasoning ability

As shown in Table 3, both two tasks show that D2R demonstrates greater stability and accuracy. In the maze judgment task, direct and language-centric CoT methods perform comparably to D2R under low-difficulty conditions, their accuracy declines significantly as task complexity increases. This suggests that textual Chain-of-Thought reasoning is insufficient for handling more complex scenarios. In contrast, the performance gap widens in favor of D2R as difficulty increases, highlighting its robustness and effectiveness under challenging conditions. Furthermore, D2R also shows higher

Method	Total Acc			Average Acc
	Easy	Normal	Hard	
Qwen2.5VL-72B(D2R)	67.0	49.0	41.0	52.3
w/o Textual Thought	52.0(-15.0)	44.3(-4.7)	32.0(-9.0)	42.7(-9.6)
w/o Drafts over dynamic images	45.1(-21.9)	33.3(-15.7)	17.1(-23.9)	31.8(-20.5)

Table 4: The accuracy of the removal of drafts over dynamic images or textual thought in D2R of the maze judgment task. The results in ‘(·)’ represent the delta performance compared to D2R with both two modalities.

Model	Method	Acc
Qwen2.5VL-7B	Draft CoT(GT)	32.5
	D2R	28.0
Qwen2.5VL-72B	Draft CoT(GT)	41.0
	D2R	41.0
QwenVL-max	Draft CoT(GT)	22.0
	D2R	28.0

Table 5: Performance of hard maze judgment between D2-CoT(GT) and D2R among three models.

accuracy in the maze navigation task. It is important to note that our method achieves performance improvements across all models and difficulties. This underscores the crucial role of integrating both textual thought and their drafts in dynamic planning tasks, as such collaboration enhances the model's ability to effectively handle complex reasoning scenarios.

5.3 How D2R is effective?

Can D2R be effective with different MLLMs' abilities? To further explore the effectiveness of our method with different models' ability, we conduct experiments on three MLLMs and the results are shown in Table 3. Although the effect varies with basic model ability and task difficulty, we can still enhance the capabilities of different models: all three MLLMs can perform better than the basics in most cases. However, Qwen2.5-VL-72B and QwenVL-max gain substantially more from D2R than Qwen2.5-VL-7B, highlighting the challenges faced by less capable models in fully utilizing our method. In other words, while D2R can help externalize the reasoning process of MLLM, it cannot fundamentally improve the inherent reasoning capacity of the model.

Are drafts and texts equally important? To further validate the contribution of the textual thought and drafts over dynamic images to D2R, we experiment by removing textual thoughts and corresponding drafts in the maze judgment task, respectively. As shown in Table 4, the removal of any component from either part leads to a performance decline across all difficulty levels, reflecting the importance of integrating both textual thought and its drafts in reasoning. Notably, performance drops more significantly when the drafts are removed than when the textual thoughts are removed, further proving the crucial role of draft processing in dynamic spatial reasoning.

Can D2R be as effective as Draft CoT(GT)? To explore whether our methods can reach the same performance with draft DoT(GT), we compare the experimental results between the D2R and Draft CoT(GT). As shown in Table 5, compared to the results with Draft CoT(GT), all three models can obtain comparable performance using our methods. The results show that our method can successfully make the MLLMs detect the current position and output the next action to accomplish different tasks in most cases, resulting in only a small gap from the ground truth.

6 Conclusion

In this paper, we introduce GRASSLAND and present two tasks to evaluate the performance on dynamic multimodal spatial reasoning: Maze Judgment and Maze Navigation. Through experiments, we observe that the combination of the textual thoughts and their drafts over dynamic input images, termed Draft CoT, significantly outperforms other approaches in these tasks, providing new insights into the dynamic spatial reasoning process. To make Draft CoT more widely applicable in existing MLLMs, we propose the Dynamic Draft Augmented Reasoning Framework, a training-free framework that generates intermediate thoughts by combining both textual thoughts and their drafts over dynamic input images. Experimental results show that D2R delivers exceptional performance across various dynamic spatial reasoning tasks.

Limitation

While D2R significantly outperforms other methods that do not require training under multiple tasks, the performance gains are different among various models, especially the weaker models gain less than the stronger models. This discrepancy suggests that D2R's benefits are more pronounced in models with a higher baseline capacity, highlighting its potential to enhance the performance of more powerful architectures more effectively. Moving forward, we plan to explore strategies for improving D2R's applicability to weaker models, aiming to achieve more excellent performance across a broader range of architectures.

Ethic Consideration

Our data is generated through open-source software and our own proprietary code. All models used are open-source, and their sources are clearly credited. The entire process follows transparent and ethical guidelines, ensuring there are no ethical concerns or issues with the data generation. We are committed to maintaining high standards of integrity and transparency in our work.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M^{3}cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *Preprint*, arXiv:2405.16473.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. *Preprint*, arXiv:2411.19488.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. *Preprint*, arXiv:2202.04800.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *Preprint*, arXiv:2502.09621.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *Preprint*, arXiv:2501.07542.

Hongcheng Liu, Zhe Chen, Hui Li, Pingjie Wang, Yanfeng Wang, and Yu Wang. 2024a. Msg-bart: Multi-granularity scene graph-enhanced encoder-decoder language model for video-grounded dialogue generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10516–10520.

Hongcheng Liu, Yusheng Liao, Siqv Ou, Yuhao Wang, Heyang Liu, Yanfeng Wang, and Yu Wang. 2024b. Med-pmc: Medical personalized multi-modal consultation with a proactive ask-first-observe-next paradigm. *ArXiv*, abs/2408.08693.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, and 8 others. 2025. Nvila: Efficient frontier visual language models. *Preprint*, arXiv:2412.04468.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. *Preprint*, arXiv:2311.17076.

OpenAI. 2025. Introducing openai o3 and o4-mini.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun

Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Ziyang Xiao, Dongxiang Zhang, Xiongwei Han, Xiaojin Fu, Yin Yu, Tao Zhong, Sai Wu, Yuan Wang, Jianwei Yin, and Gang Chen. 2024. Enhancing llm reasoning via vision-augmented prompting. In *Advances in Neural Information Processing Systems*, volume 37, pages 28772–28797. Curran Associates, Inc.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025a. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025b. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023a. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *Preprint*, arXiv:2309.17421.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *Preprint*, arXiv:2303.11381.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Li Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *ArXiv*, abs/2501.13106.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Preprint*, arXiv:2310.16436.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Details on MLLMs

Table 6 shows the hyperparameters for generating with MLLM and size information for each model. For QwenVL-Max and Qwen-Max, we use the 2025-01-25 version through Aliyun platform.

Model	Max tokens	Size
QwenVL2.5	700	72B, 32B, 7B
InternVL2.5	700	78B, 38B, 8B
QwenVL-Max*	700	-
VideoLLaMA3	700	7B
Qwen-Max*	400	-

Table 6: Hyperparameters for model generation. Model called via API has been marked by *

B Metric

We use the accuracy as the evaluation metric for both two tasks. For the maze judgment task, the accuracy aims to detect whether the model can obtain the final state. For the maze navigation task, the accuracy aims to detect whether the model can reach the final position according to the model's response.

C Other results

The other results about our methods are presented in Table 7 and Table 8. Specifically, Table 7 and Table 8 presents the detailed performance across various methods for each task. For maze judgment task, we observe a clear uneven distribution of answer accuracies on other methods, with answer "D. Action Failed: Agent Safe but Fail to Reach Destination" being significantly more accurate than the other three options. It reflects the shortcomings of inadequate ability to judge complex states on these methods. In contrast, D2R outperforms other methods, optimizing accuracy on the complex options A and B.

For the maze navigation task, we notice an interesting feature in all methods that in the correct path answer, the effective length is shorter than the full path length. It means the goal point is reached at the halfway. Even D2R can only make the gap smaller, not eliminate it completely. This reflects a possible deficiency in the model to perform spatial planning tasks.

D Prompt

D.1 Basic Prompt

Table 9 shows the prompting template of direct reasoning and D2R task prompt for each task. Table 10 and Table 11 shows the prompt for each task with different reasoning methods.

D.2 Method Prompt

Table 12 shows the example of prompt for scheduling hub. Table 13 shows the prompt in iteration process for each task in D2R.

E Case Study

E.1 Maze Judgment

Figure 7 presents the thought process of D2R in maze judgment task. In each step, after receiving the action instruction, D2R mark the original frame with the position staying now, then searches the grids in the action direction to judge the state after the action is executed.

E.2 Maze Navigation

Figure 8 provides an example of the thought process of D2R in maze navigation task. In each step, D2R receives the original frame, then mark it with the current position. According to the marked frame and full video, D2R judges the dangerous position and generates a safe move direction for now, until it reaches the destination.

Model	Method	Choice Acc.				Total Acc.
		A	B	C	D	
Easy Level						
Qwen2.5VL-7B	Direct	45.0	-	13.2	21.8	22.5
	CoT	35.0	-	23.3	13.3	18.0
	1-shot CoT	75.0	-	15.7	10.5	18.0
	VAP	75.0	-	18.4	3.5	13.5
	D2R (ours)	65.0	-	23.7	32.4	34.0
Qwen2.5VL-72B	Direct	25.0	-	15.8	78.2	61.0
	CoT	40.0	-	10.5	85.9	67.0
	1-shot CoT	10.0	-	5.3	97.2	71.0
	VAP	10.0	-	5.3	19.0	15.5
	D2R (ours)	30.0	-	21.1	84.5	67.0
QwenVL-Max	Direct	35.0	-	10.5	48.6	40.0
	CoT	35.0	-	10.5	43.0	36.0
	1-shot CoT	30.0	-	7.9	19.0	18.0
	VAP	35.0	-	5.3	14.8	15.0
	D2R (ours)	40.0	-	2.6	59.2	46.5
Normal Level						
Qwen2.5VL-7B	Direct	46.7	325.3	22.2	40.6	34.0
	CoT	46.7	23.5	25.0	30.2	29.0
	1-shot CoT	13.3	5.9	5.6	26.0	20.5
	VAP	80.0	0.0	8.3	12.5	15.0
	D2R (ours)	33.3	0.0	9.7	83.3	46.0
Qwen2.5VL-72B	Direct	6.7	17.6	11.1	67.7	38.5
	CoT	6.7	11.8	8.3	74.0	40.0
	1-shot CoT	13.3	5.9	6.9	88.5	46.5
	VAP	20.0	0.0	4.2	35.4	20.0
	D2R (ours)	33.3	47.1	12.5	79.2	49.0
QwenVL-Max	Direct	26.7	0.0	8.3	34.4	21.5
	CoT	40.0	5.9	4.2	39.6	24.0
	1-shot CoT	52.6	20.8	4.9	19.1	17.0
	VAP	13.3	0.0	11.1	8.3	9.0
	D2R (ours)	26.7	11.8	2.8	65.6	35.5
Hard Level						
Qwen2.5VL-7B	Direct	21.1	54.7	8.6	36.2	28.5
	CoT	15.8	43.4	18.5	25.5	26.5
	1-shot CoT	52.6	20.8	4.9	19.1	17.0
	VAP	89.5	7.5	13.6	17.0	20.0
	D2R (ours)	15.8	1.9	11.0	91.0	28.0
Qwen2.5VL-72B	Direct	10.5	9.4	4.9	57.4	19.0
	CoT	10.5	5.7	9.9	70.2	23.0
	1-shot CoT	5.3	0.0	8.6	91.5	25.5
	VAP	10.5	0.0	1.2	57.4	15.0
	D2R (ours)	15.8	39.6	19.8	89.4	41.0
QwenVL-Max	Direct	31.6	1.9	9.9	27.7	14.0
	CoT	42.1	0.0	6.2	21.3	11.5
	1-shot CoT	21.1	7.5	8.6	8.5	9.5
	VAP	42.1	0.0	14.8	12.8	13.0
	D2R (ours)	21.1	20.8	6.2	76.6	28.0

Table 7: Detailed performance on maze judgment task.

Model	Method	Arrived	Failed	Unfinished	Ave. Step (Effective)	Ave. Step (Answer)
Easy Level						
Qwem2.5VL-7B	Direct	1.0	4.0	95.0	4.50	6.00
	CoT	1.5	9.5	89.0	4.00	5.67
	1-shot CoT	2.5	62.0	35.5	4.40	5.80
	D2R (ours)	4.0	38.0	58.0	3.88	6.00
Qwen2.5VL-72B	Direct	31.0	40.0	29.0	3.55	5.84
	CoT	16.5	43.5	40.0	3.48	5.97
	1-shot CoT	17.5	35.5	47.0	3.54	5.91
	D2R (ours)	38.0	38.0	24.0	3.72	5.74
QwenVL-Max	Direct	19.5	30.5	50.0	3.31	5.97
	CoT	22.5	39.5	38.0	3.56	5.93
	1-shot CoT	1.0	41.0	58.0	6.00	6.00
	D2R (ours)	27.5	41.0	31.5	4.04	5.47
Normal Level						
Qwem2.5VL-7B	Direct	2.0	8.5	89.5	4.00	5.75
	CoT	1.5	13.5	85.0	3.67	6.00
	1-shot CoT	4.5	52.5	43.0	3.78	5.56
	D2R (ours)	4.5	52.5	43.0	4.67	6.00
Qwen2.5VL-72B	Direct	21.5	58.5	20.0	3.72	5.77
	CoT	17.5	48.5	34.0	3.89	6.06
	1shot-CoT	5.0	56.5	38.5	3.50	6.00
	D2R (ours)	26.0	51.0	23.0	3.94	5.98
QwenVL-Max	Direct	10.5	40.0	49.5	3.52	6.00
	CoT	10.5	38.0	51.5	3.71	6.05
	1-shot CoT	0.5	50.5	49.0	6.00	6.00
	D2R (ours)	21.5	54.0	24.5	3.93	5.93
Hard Level						
Qwem2.5VL-7B	Direct	1.0	9.5	89.5	4.00	5.50
	CoT	0.0	12.0	88.0	0.00	0.00
	1-shot CoT	2.5	62.0	35.5	4.40	5.80
	D2R (ours)	2.0	63.0	35.0	5.25	6.00
Qwen2.5VL-72B	Direct	6.5	74.5	19.0	4.69	5.69
	CoT	0.5	72.0	27.5	4.00	6.00
	1-shot CoT	1.5	67.5	31.0	4.67	6.00
	D2R (ours)	12.5	75.5	12.0	4.48	6.00
QwenVL-Max	Direct	1.5	61.0	37.5	4.67	6.00
	CoT	6.0	59.5	34.5	4.17	6.00
	1-shot CoT	0.0	62.5	37.5	0.00	0.00
	D2R (ours)	7.0	77.5	15.5	4.86	5.85

Table 8: Detailed performance on maze navigation task.

Direct

Task:Maze Judgment
Tile info: character can move pass the green tile(grass). The red tile(lava) and the blue tile(water) is dangerous, when player move into them, game fail. The yellow and brown tile is wall which can not be passed.
The white flag is start point, the red flag is destination point
Player can't move off the map, considering it as air walls
Actions: the lava tile change position every second, and player also move every second. Consider player move first in same time, which mean if player and lava tile move to same position, the game fail.
Determine whether the agent (elf character) can safely reach the destination following the action sequence without falling into the lava or water. If not, identify the failure reason shortly. The definitions of the actions are as below.
* In the video, the red line shows the movement path of the agent.
* Go up/left/down/right: move one grid space in the absolute up/left/down/right direction.
After analyse the actions, return A, B, C or D.
Full Action Sequence: action_sequence
A. Action Success.
B. Action Failed: Fall into the water.
C. Action Failed: Fall into the lava.
D. Action Failed: Agent Safe but Fail to Reach Destination.

Task: Route Plan
Tile Info: The character can move across the green tile (grass). The red tile (lava) and the blue tile (water) are dangerous. If the player moves onto them, the game fails. The yellow and brown tiles are walls, which cannot be passed. The white flag represents the starting point, and the red flag represents the destination.
The player cannot move off the map; treat the edges as air walls. Actions: The lava tiles change position every second, and the player also moves every second.
Consider the player moving first in the same time step, which means if the player and a lava tile move to the same position, the game fails.
You will receive a 6-second video showing the dynamic map. Your task is to analyze this video, apply the rules mentioned above, then determine a route that allows the player to reach the destination safely within 6 steps.
The answer should follow this format: "Action: [START] Go right, Go up, Go down, ... [END]"
Each command corresponds to one move. And put it at the end of your answer.
Move Commands: Go up/left/down/right: Move one grid space in the absolute up/left/down/right direction.

Table 9: Example of input for Direct reasoning

CoT reasoning

Task:Maze Judgment
Tile info: character can move pass the green tile(grass). The red tile(lava) and the blue tile(water) is dangerous, when player move into them, game fail. The yellow and brown tile is wall which can not be passed.
The white flag is start point, the red flag is destination point
Player can't move off the map, considering it as air walls
Actions: the lava tile change position every second, and player also move every second. Consider player move first in same time, which mean if player and lava tile move to same position, the game fail.
Determine whether the agent (elf character) can safely reach the destination following the action sequence without falling into the lava or water. If not, identify the failure reason shortly. The definitions of the actions are as below.
* In the video, the red line shows the movement path of the agent.
* Go up/left/down/right: move one grid space in the absolute up/left/down/right direction.
After analyse the actions, return A, B, C or D.
Full Action Sequence: action_sequence
A. Action Success.
B. Action Failed: Fall into the water.
C. Action Failed: Fall into the lava.
D. Action Failed: Agent Safe but Fail to Reach Destination.
Let's think it step-by-step and make right choice.

Task: Route Plan
Tile Info: The character can move across the green tile (grass). The red tile (lava) and the blue tile (water) are dangerous. If the player moves onto them, the game fails. The yellow and brown tiles are walls, which cannot be passed. The white flag represents the starting point, and the red flag represents the destination.
The player cannot move off the map; treat the edges as air walls. Actions: The lava tiles change position every second, and the player also moves every second.
Consider the player moving first in the same time step, which means if the player and a lava tile move to the same position, the game fails.
You will receive a 6-second video showing the dynamic map. Your task is to analyze this video, apply the rules mentioned above, then determine a route that allows the player to reach the destination safely within 6 steps.
The answer should follow this format: "Action: [START] Go right, Go up, Go down, ... [END]"
Each command corresponds to one move. And put it at the end of your answer.
Move Commands: Go up/left/down/right: Move one grid space in the absolute up/left/down/right direction.
Let's think it step-by-step and make right choice.

Table 10: Example of input for CoT reasoning

CoT with 1-shot prompting

Task:Maze Judgment
Tile info: character can move pass the green tile(grass). The red tile(lava) and the blue tile(water) is dangerous, when player move into them, game fail. The yellow and brown tile is wall which can not be passed.
The white flag is start point, the red flag is destination point
Player can't move off the map, considering it as air walls
Actions: the lava tile change position every second, and player also move every second. Consider player move first in same time, which mean if player and lava tile move to same position, the game fail.
Determine whether the agent (elf character) can safely reach the destination following the action sequence without falling into the lava or water. If not, identify the failure reason shortly. The definitions of the actions are as below.
* In the video, the red line shows the movement path of the agent.
* Go up/left/down/right: move one grid space in the absolute up/left/down/right direction.
After analyse the actions, return A, B, C or D.
Full Action Sequence: action_sequence
A. Action Success.
B. Action Failed: Fall into the water.
C. Action Failed: Fall into the lava.
D. Action Failed: Agent Safe but Fail to Reach Destination.
Here is an example, consider video follow behind the text. The action sequence is: Go down, Go up, Go up, Go left. First, the agent move down. Check the tile agent move to, it is grass with no trap, so agent can move to. Then agent move up, it is start point, agent can move to here. Then agent move up again, it is grass, agent can move to here. Then agent move left, it is the end point, so agent arrive at the destination. So the answer is: A. Action Success.
Video: <example_video>

Task: Route Plan
Tile Info: The character can move across the green tile (grass). The red tile (lava) and the blue tile (water) are dangerous. If the player moves onto them, the game fails. The yellow and brown tiles are walls, which cannot be passed. The white flag represents the starting point, and the red flag represents the destination.
The player cannot move off the map; treat the edges as air walls. Actions: The lava tiles change position every second, and the player also moves every second.
Consider the player moving first in the same time step, which means if the player and a lava tile move to the same position, the game fails.
You will receive a 6-second video showing the dynamic map. Your task is to analyze this video, apply the rules mentioned above, then determine a route that allows the player to reach the destination safely within 6 steps.
The answer should follow this format: "Action: [START] Go right, Go up, Go down, ... [END]"
Each command corresponds to one move. And put it at the end of your answer.
Move Commands: Go up/left/down/right: Move one grid space in the absolute up/left/down/right direction.
Here is an example, consider video follow behind the text. To move safely, we check the position of destination, make choice, and review the traps position in video to conform the action safe. In this example, the best action is: [START] Go right, Go right, Go right, Go right, Go right, Go down [END]
Video: <example_video>

Table 11: Example of input for CoT reasoning with 1-shot prompting

Planning prompt for manager

You are controlling the VideoProcessing agent, PositionGet agent, DrawPosition agent and MLLM-Reply agent.
1.Each time you need to extract and save the video by VideoProcessing agent, get the postion by PositionGet agent, and draw the position by DrawPosition agent.
2.you need to complete the task by MLLMReply agent until the MLLMReply agent output final answer.
3.If the MLLMReply agent do not output the final answer, you need to continue completing the task by MLLMReply agent.
4.You should follow the order in examples, and don't make any superfluous execution.
5.When MLLMReply agent output <finish>, you need to finish the task.

Table 12: Example of input for manager LLM

Iteration prompt for MLLM

Additionly, there are the <visualization of the thought>, you should output next action until the final answer is obtained. Here are some tips you should follow:
1.<visualization of the thought> consist of the image of now position and the text of next action.
2.The black square represents your current position.
3.Based on the <Task Description> and <visualization of the thought>, follow the next action in <Full Action Sequence>. Each action in <visualization of the thought> has already been executed, don't execute them again.
4.If the next action is wall or move off the map, append <can_not_pass> after the next action.
5.As soon as you can get the final answer, you will immediately output the final answer and append <finish> after it.
6.Each time you can only output one action or one final answer.

Table 13: Example of input system prompt for MLLM in iteration process

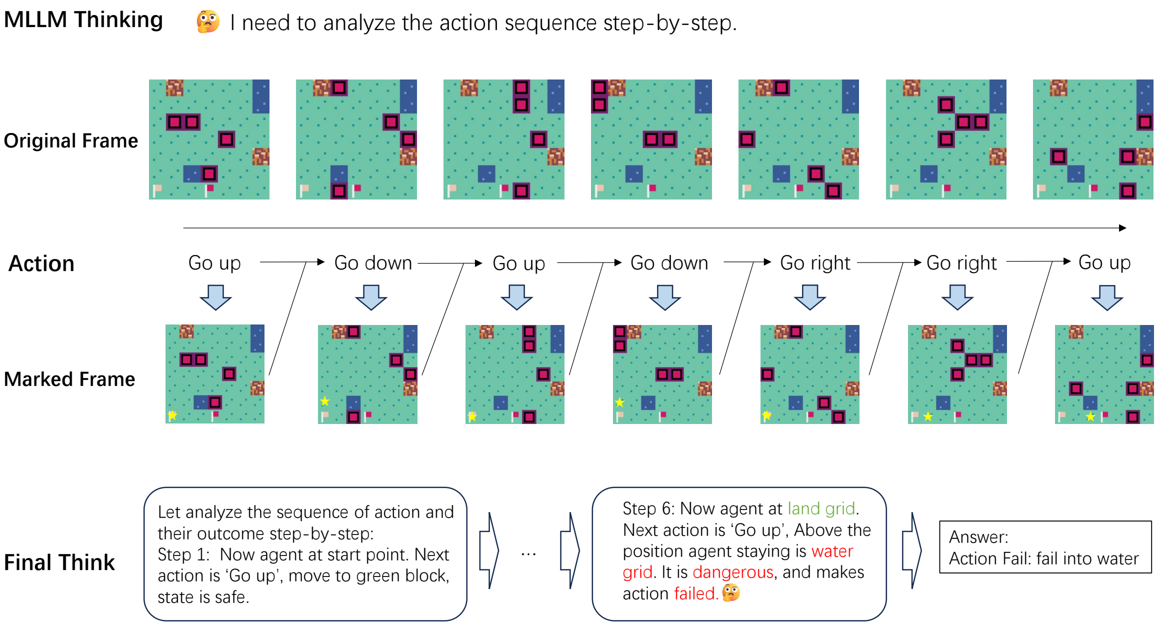

Figure 7: An example of the thought process for D2R in maze judgment task.

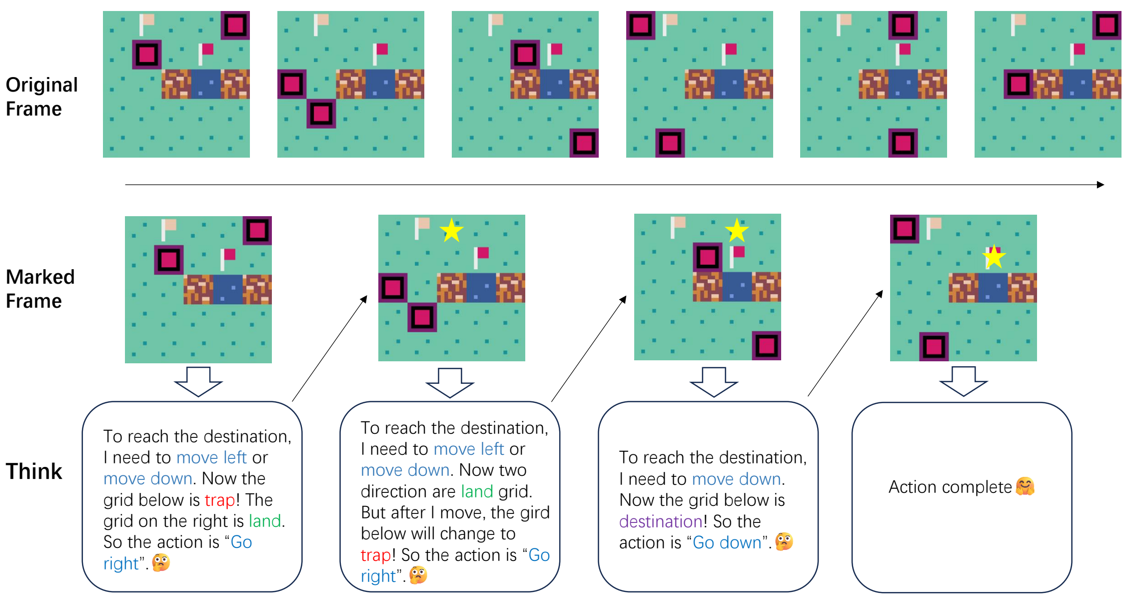

Figure 8: An example of the thought process for D2R in maze navigation task.