Deconstructing Spatial Complexity: Hierarchical Decomposition for LLM Spatial Reasoning

Yi Wang[1]*, Haojie Lu[1]*, Li Chen[1], Sihong Xie[2]†
[1]College of Future Technology, [2]Artificial Intelligence Thrust
The Hong Kong University of Science and Technology (Guangzhou)
ywang778@connect.hkust-gz.edu.cn, sihongxie@hkust-gz.edu.cn

Abstract

LLMs have shown remarkable proficiency in general language understanding and reasoning. However, they consistently underperform in spatial reasoning that severely limits their application, particularly in embodied intelligence. Inspired by the success of hierarchical reinforcement learning, this paper introduces a novel method for hierarchical task decomposition in LLM spatial reasoning. Our approach guides LLMs to decompose complex tasks into manageable sub-tasks by identifying key intermediate states and generating simplified sub-environments. However, we identify that LLMs often fail to derive optimal intermediate states due to their insufficient spatial prior, leading to sub-optimal task decomposition. To address this limitation and enhance its planning capability, we propose the MCTS-Guided Group Relative Policy Optimization (M-GRPO), where we reformulate the UCT formula by incorporating the LLM's prior predictive probabilities alongside its epistemic uncertainty. Furthermore, we implement a more fine-grained advantage function, enabling the model to learn optimal path planning. Experimental results demonstrate that our method substantially improves LLM performance on spatial tasks, including navigation, planning, and strategic games, achieving state-of-the-art results. This work paves the way for LLMs in real-world applications.

1 Introduction

Large Language Models (LLMs) have revolutionized the landscape of artificial intelligence, achieving remarkable breakthroughs across various domains, including natural language processing and scientific reasoning (Zhao et al., 2023). However, as LLMs transition into the era of embodied AI, a critical and persistent bottleneck has emerged: their inherent limitations in spatial reasoning. While

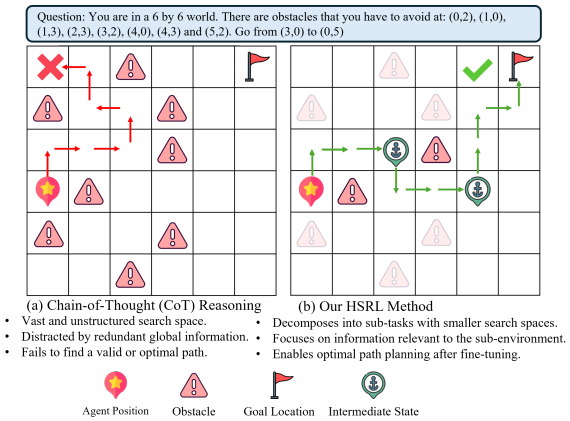

Figure 1: Comparison between CoT Reasoning and our HSRL method. **(a)** Standard CoT reasoning fails on complex spatial planning by inefficiently exploring a vast search space amidst distracting information. **(b)** In contrast, our HSRL framework succeeds by decomposing the task via key intermediate states and constructing focused sub-environments, which enables efficient and optimal planning.

LLMs excel at manipulating abstract concepts and language, they often struggle with understanding complex spatial relationships, performing efficient path planning, and engaging in sequential action reasoning(Ma et al., 2025; Chen et al., 2024). This severely limits their development and practical deployment in embodied systems.

Existing research has explored several avenues to address this challenge, yet each faces significant limitations. Prompt engineering methods like CoT(Wei et al., 2022), ToT(Yao et al., 2023a) and ProgPrompt(Singh et al., 2023) aim to elicit reasoning through specialized prompts, but their effectiveness is capped by the model's often flawed intrinsic spatial capabilities. Fine-tuning approaches (Dao and Vu, 2025; Deng et al., 2025; Aghzal et al., 2024b) show promise but typically demand vast, expensive task-specific datasets and suffer from poor generalization to novel environments. Task decomposition strategies like HyperTree (Gui et al.,

*Equal contribution.
†Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 59298–59319
October 24-29, 2026 ©2026 Association for Computational Linguistics

2025) and Plan-and-Act (Erdogan et al., 2025) are primarily designed for tasks with clear, language-based "logical breaks". For example, "cooking" naturally includes sub-steps like "buying ingredients" and "chopping." In such tasks, sub-goals have distinct semantic boundaries, allowing models to decompose them using pre-trained semantic priors. Consequently, these strategies are rendered ill-suited for spatial reasoning problems like pathfinding that lack such linguistic segmentation. Finally, offloading planning to external, non-differentiable tools breaks the end-to-end optimization paradigm, as these tools cannot be jointly trained with the LLM's representation layer and may not be universally deployable at test time.

To overcome these limitations, we introduce Hierarchical Spatial Reasoning with LLM (HSRL). The core innovation of HSRL lies in its state- and environment-based hierarchical mechanism, which is fundamentally different from prior language-based decomposition methods. The framework employs a two-level hierarchy. A high-level LLM planner decomposes complex tasks by generating a structured sequence of intermediate states, which serve as sub-goals to guide the overall planning process. Then, for each state-to-state transition, we decouple the process into two distinct components: first, an Environment Processor constructs a localized sub-environment, and subsequently, a Low-Level Action Generator generates the precise actions required to reach the target sub-goal. While this hierarchical structure provides a powerful framework for decomposition, the high-level planner's effectiveness is constrained by the LLM's insufficient spatial priors, leading to the generation of sub-optimal intermediate states. To address this, we propose an innovative online fine-tuning framework, M-GRPO, designed to enhance the high-level planner. Our approach improves planning optimality by tackling two fundamental challenges: effective exploration of the solution space and precise credit assignment for training. To achieve robust exploration, we draw inspiration from Monte Carlo Tree Search (MCTS), where the high-level LLM generates multiple candidate sequences of intermediate states, building a search tree to systematically explore diverse planning strategies. Unlike standard LLM-MCTS integration paradigms (Xie et al., 2024; Guan et al., 2025) rely on the vanilla UCT formula, we propose incorporating the LLM's intrinsic confidence and uncertainty as internal reward signals. This integration facilitates the selection of trajectories that exhibit both high environmental rewards and high internal certainty, thereby mitigating the instability caused by high-reward but low-confidence paths—which often represent stochastic hallucinations that hinder convergence.(Kuhn et al., 2023; Azaria and Mitchell, 2023) Furthermore, we leverage perplexity-based uncertainty as an informative metric to evaluate whether the model has sufficiently explored the state space within its semantic knowledge, leading to a more principled and efficient search process. For precise credit assignment, we introduce a fine-grained advantage function, a significant departure from traditional Group Relative Policy Optimization (GRPO) which evaluates whole-trajectory values without detailed supervision. Our method calculates the advantage of each intermediate state relative to its "sibling" states (i.e., those that share a common prefix state sequence). This provides a focused and accurate training signal, enabling the LLM to learn which specific sub-goals are the most effective. Our method requires only a small amount of data and can be flexibly applied to multi-level planning tasks. In summary, this work makes the following key contributions:

- **A Novel State- and Environment-Based Hierarchical Reasoning Framework:** We introduce HSRL, a framework that presents a novel state- and environment-based decomposition paradigm for LLM spatial reasoning, departing from prevalent language-based methods. This paradigm is specifically designed to address continuous spatial problems where traditional language-based decomposition is ineffective.

- **A Novel Fine-Tuning Framework for Planning Optimality:** To address the sub-optimal planning inherent in pre-trained LLMs, we develop M-GRPO, a new fine-tuning algorithm. By integrating a modified Monte Carlo Tree Search exploration mechanism with a fine-grained, node-level advantage function, our method substantially improves planning optimality with high data efficiency.

- **Comprehensive Empirical Validation of Superiority:** Through extensive experiments on large-scale navigation, object planning, and strategy game benchmarks, we demonstrate that HSRL achieves state-of-the-art performance. The results validate its significant

gains over existing methods and its strong generalization across diverse task modalities.

2 Related Work

2.1 Spatial Reasoning in Large Language Models

Many researchers have pointed out that LLMs have weaknesses in spatial reasoning or spatial planning(Aghzal et al., 2024a,b). To address these issues, some methods leverage in-context examples and prompting techniques, such as Chain-of-Thought (CoT)(Wei et al., 2022) and Tree-of-Thought (ToT)(Yao et al., 2023a), which have demonstrated remarkable reasoning abilities in various tasks. However, for spatial reasoning tasks, in-context learning often fails because LLMs lack spatial reasoning knowledge or their knowledge even conflicts with it.

To overcome this challenge, some studies utilize LLMs for general-purpose reasoning, converting spatial information into logical forms(Yang et al., 2023) or using them as a general pattern machine for sequence transformation(Mirchandani et al., 2023; Gong et al., 2025). Recently, other works have evaluated LLMs as a cognitive capability in navigation and planning tasks(Momennejad et al., 2023). However, these methods perform poorly in tasks requiring continuous action reasoning.

Another mainstream approach introduces closed-loop feedback mechanisms. Some works, like (Renze and Guven, 2024), use self-reflection for self-evaluation and replanning, while others adopt external feedback for reflection (Kumar et al., 2024). Furthermore, the Vision-of-Thought (VoT) method (Wu et al., 2024) materializes intermediate states to assist with reasoning. Nevertheless, this iterative feedback loop often results in high costs and inefficiency in querying or interactions.

2.2 Hierarchical Method

Hierarchical reasoning breaks down decision-making tasks into multiple levels, from high-level strategic planning to low-level specific control. This decomposition reduces computational complexity by solving several less difficult sub-tasks, thus enabling the handling of tasks more challenging than direct complex reasoning. Hierarchical reasoning has achieved notable results in many reinforcement learning tasks, especially in embodied AI scenarios. For example, (Duan et al., 2020) has applied hierarchical methods to autonomous driving, allowing for smooth and safe decision-making on highways. (Lu et al., 2023) and (Zhu and Hayashibe, 2023) separate decision-making tasks into different layers, such as global path planning and local motion control. These models benefit from breaking down the decision-making process into simpler, more tractable components, enabling each layer to focus on a specific task. This enhances computational efficiency and decision accuracy in complex environments.

In recent years, hierarchical reasoning methods have also been successfully introduced into the planning tasks of LLMs. For instance, DeAR(Xue et al., 2024) imitates the human reasoning cycle by using a tree-based question decomposition approach to organize the reasoning process and break down problems into simpler sub-questions. Hyper-Tree Planning(Gui et al., 2025) is a new paradigm that enhances LLM reasoning with a hypertree structure. It effectively breaks down intricate reasoning steps using a flexible divide-and-conquer strategy to handle diverse constraints and manage multiple distinct sub-tasks, demonstrating superior performance in complex tasks like travel planning. Plan-and-Act(Erdogan et al., 2025) explicitly separates high-level planning from low-level execution. This framework includes a PLANNER model for generating structured high-level plans and an EXECUTOR model for translating these plans into environment-specific actions, thereby improving performance on complex multi-step tasks such as web navigation.

However, these methods primarily focus on high-level, coarse-grained task planning, failing to fully leverage hierarchical reasoning for fine-grained, low-level spatial and motion control. Furthermore, while prior MCTS-based semi-online methods (e.g., SEEA-R1; Tian et al., 2025) have explored search-based planning, they are largely restricted to discrete action spaces and suffer from distribution shifts on static trees. To bridge these gaps, our study introduces M-GRPO to solve complex action planning problems. By employing strictly on-policy online optimization with dynamic iterative expansion, our method efficiently navigates the combinatorially explosive, high-dimensional search spaces inherent in spatial planning.

3 Method

In this work, we introduce the HSRL framework, as illustrated in Figure 2, to address the limitations

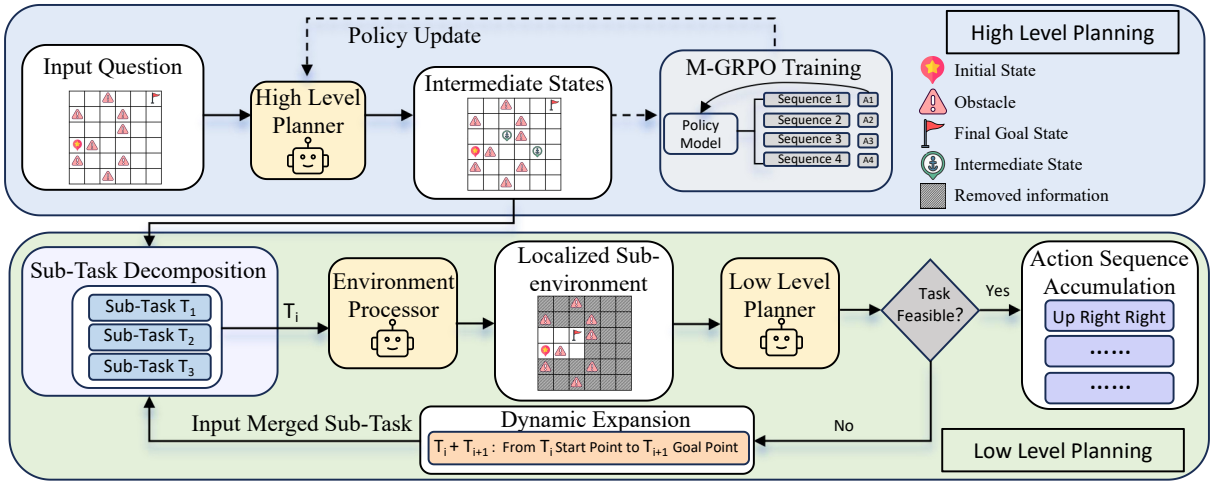

Figure 2: An overview of the HSRL framework. This framework employs a two-level hierarchical strategy. An M-GRPO trained high-level planner first identifies key intermediate states, decomposing the task into a series of sub-tasks. A low-level planner selects relevant information for the sub-task and then generates action sequences for each sub-task within a localized sub-environment. If a sub-task is unsolvable, it is merged with the next one (e.g., from the start of Subtask 1 to the end of Subtask 2) and replanned.

of existing LLM-based planning methods (see Section 2). Our approach consists of two key components: a novel two-level hierarchical framework that decomposes complex tasks into a series of manageable sub-problems, and an innovative MCTS-guided finetuning method designed to enhance the optimality of the generated plans.

3.1 Hierarchical Planning with State and Environment Decomposition

Our framework leverages a two-level hierarchical decomposition strategy to break down complex planning tasks. This decomposition is applied at both the state level and the environmental level, effectively managing the complexity of the problem space.

State-Level Decomposition via LLM. Prior research in LLM-based path planning has shown promising results by manually decomposing tasks into sub-goals (Aghzal et al., 2024b). We extend this concept by enabling the LLM to autonomously generate these key intermediate states. Given a task's initial and final states, our method prompts the LLM to reason and generate a concise sequence of critical intermediate states. This process transforms a high-level goal into a series of state-to-state transitions, effectively simplifying the planning horizon for subsequent steps.

Environmental-Level Decomposition and Dynamic Expansion. After decomposing the task at the state level, much of the global environmental information becomes irrelevant noise for solving a specific sub-task, which can hinder the reasoning process. Following the generation of the state sequence, we define a sub-task for each consecutive pair of intermediate states. For each sub-task, we create a corresponding regional environment by identifying information that is relevant to the current sub-problem (e.g., obstacles or landmarks within a localized area). This hierarchical representation allows the model to focus on a smaller, more manageable sub-environment, thereby improving efficiency and reducing the search space. If the model is unable to find a valid path within the localized environment, the scope of the sub-task is expanded. The end state is extended to the next intermediate state in the sequence, creating a larger sub-task that encompasses a broader area. This process is repeated until a solution is found or, in the worst case, the problem reverts to the original, full-scale task, ensuring robust and complete coverage of the problem space.

3.2 Optimizing Planning with M-GRPO

Due to an inherent lack of pre-training in spatial reasoning, LLMs often struggle to generate optimal sequences of intermediate states. However, the generation of correct intermediate states is a critical prerequisite for the success of subsequent environment decomposition and low-level action planning. To address this, we propose an online learning approach, as illustrated in Figure 3, that integrates the exploratory power of MCTS with

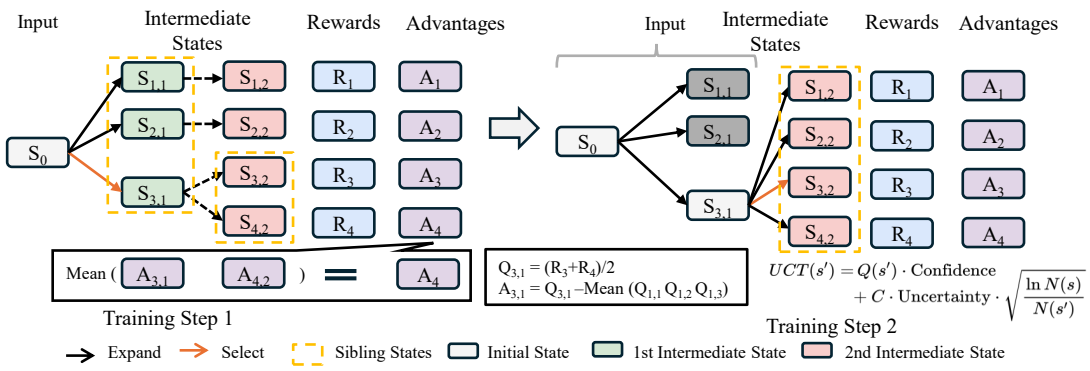

Figure 3: Overview of the M-GRPO algorithm. Starting from the initial state S_0, the framework expands four simulation sequences in parallel. The **advantage values** of nodes within these sequences are calculated to guide the model training step. Post-training, the **pivotal intermediate waypoint** is selected based on the reformulated UCT formula, serving as the root for the subsequent expansion and training cycle. This iterative process continues until the MCTS termination criterion is satisfied.

the fine-tuning process of GRPO. This approach enables the LLM to learn and improve its planning policy during exploration.

MCTS-Guided Exploration with LLM-Induced Prior Preference. We formulate intermediate-state generation as a tree search problem, where each node denotes a partial reasoning state and each root-to-leaf path defines a candidate trajectory. Starting from the initial state, the search tree is iteratively expanded through selection, expansion, simulation, and backpropagation.

A central challenge is how to incorporate the LLM's internal confidence into the search process without treating it as a calibrated probability of correctness. To this end, we use the generation likelihood of the newly fine-tuned LLM policy to construct a prior preference over candidate states. For a generated child state s' with token sequence $x_{1:|T|}$, we compute the average log-likelihood:

$$\ell(s') = \frac{1}{|T|} \sum_{t=1}^{|T|} \log P(x_t \mid x_{<t}, \text{context}). \quad (1)$$

Based on this quantity, we define the LLM-induced prior confidence as

$$c(s') = \exp(\tau \cdot \ell(s')), \quad (2)$$

where τ denotes the temperature parameter. Intuitively, candidates that are more semantically coherent and more supported by the model receive a larger weight in the exploitation term. Let $Q(s')$ denote the empirical value estimate of state s', computed from Monte Carlo returns, then the prior-modulated value estimate can be obtained by $\tilde{Q}(s') = c(s')Q(s')$, where the empirical node value is calibrated by an LLM-induced prior.

Meanwhile, to encourage exploration toward low-likelihood regions where the LLM policy is less certain, we define an exploration factor

$$u(s') = 1 + \gamma \cdot \text{clamp}(-\ell(s'), 0, u_{\max}), \quad (3)$$

which serves as an uncertainty-aware exploration amplifier. Larger uncertainty leads to a stronger exploration bonus, encouraging the search to gather additional evidence in the model's epistemically weak regions.

We then define the refined selection score as

$$U(s'; s) = \tilde{Q}(s') + C\, u(s') \sqrt{\frac{\ln N(s)}{N(s')}}, \quad (4)$$

considering exploitation and exploration for key points' distribution and further LLM rollout. The next state is selected by

$$s_{\text{next}} = \arg \max_{s' \in \text{Children}(s)} U(s'; s). \quad (5)$$

By considering the uncertainty and LLM prior preference over plausible candidates, this formulation augments the standard optimism-under-uncertainty structure of UCTs.

Fine-Grained Advantage Function for Precise Policy Updates In standard policy optimization frameworks like GRPO, the advantage function is typically computed based on the cumulative return of an entire trajectory. This coarse-grained signal poses a significant credit assignment challenge, as it fails to disambiguate the individual contributions

Algorithm 1 M-GRPO Training Algorithm

Require: Policy π_θ, initial state s_0, iterations N_{max}, group size M
Ensure: Optimized policy π_θ
$T \leftarrow$ InitializeTree(s_0)
for it = 1 **to** N_{max} **do**
  // MCTS-Guided Exploration
  $L \leftarrow$ SELECTLEAF(T) using Refined UCT (Eq. 4)
  $\{\tau_m\}_{m=1}^{M} \leftarrow$ EXPANDANDSIMULATE(π_θ, L)
  for each trajectory τ_m **do**
    $R_m \leftarrow$ SIMULATETOGOAL(τ_m)
    BACKPROPAGATE(τ_m, R_m) {Update Q-values}
  end for
  // Fine-Grained Advantage Estimation
  $A_{\text{all}} \leftarrow [\,]$
  for each $\tau_m = \{s_{m,1}, \ldots, s_{m,T}\}$ **do**
    for $n = 1$ **to** T **do**
      $\mathcal{S}_{\text{sib}} \leftarrow$ GETSIBLINGS($s_{m,n}$)
      $\bar{Q}_{\text{sib}} \leftarrow$ MEAN$\{Q(s') \mid s' \in \mathcal{S}_{\text{sib}}\}$
      $A_{m,n} \leftarrow Q(s_{m,n}) - \bar{Q}_{\text{sib}}$ {Node-level adv.}
    end for
    $A(\tau_m) \leftarrow \frac{1}{T}\sum_{n=1}^{T} A_{m,n}$ {Trajectory adv.}
    A_{all}.APPEND($A(\tau_m)$)
  end for
  $\pi_\theta \leftarrow$ UPDATEPOLICY($\pi_\theta, A_{\text{all}}$) {GRPO Loss Update}
end for
return π_θ

of intermediate states. Consequently, it is difficult for the model to pinpoint which specific choices are most critical for achieving success.

To overcome this limitation, we introduce a fine-grained advantage function calculated at the intermediate state level. Our approach is tailored for a tree-search process wherein a LLM generates a set of M candidate sequences (or completions), $\{\tau_1, \ldots, \tau_M\}$, for a given planning problem. Each trajectory τ_m is composed of a sequence of intermediate states, $\tau_m = (s_{m,1}, s_{m,2}, \ldots, s_{m,T_m})$.

Let $s_{m,n}$ be the n-th intermediate state in the m-th generated sequence. We estimate its corresponding state-value, or Q-value $Q_{m,n}$, as the mean empirical return from all Monte Carlo simulations that traverse this state. Specifically, if $W_{m,n}$ is the sum of cumulative rewards from all visits to state $s_{m,n}$ and $N_{m,n}$ is its total visit count, the Q-value is given by:

$$Q_{m,n} = \frac{W_{m,n}}{N_{m,n}} \tag{6}$$

We then define the advantage of a specific state, $A_{m,n}$, relative to its "sibling" states—i.e., the set of other candidate states $\{s_{j,n}\}_{j=1}^{M}$ that share a common prefix sequence. The state-level advantage is formulated as:

$$A_{m,n} = Q_{m,n} - \text{Mean}(Q_{\text{siblings}}) \tag{7}$$

where the second term represents the mean Q-value across all sibling states at depth n. This formulation directly quantifies how much better the choice leading to $s_{m,n}$ is compared to the average of alternative choices at that decision point. A deliberate design choice is the omission of reward normalization (i.e., skipping division by the standard deviation). As the LLM often generates identical optimal completions, forgoing normalization prevents "reward hacking," where the value of a superior path could be artificially deflated due to its high frequency of generation. The necessity of this non-normalization approach in such scenarios has been formally demonstrated in (Liu et al., 2025).

Finally, to align with the GRPO framework, we compute a single advantage value for each trajectory by averaging the advantages of all its constituent intermediate states. For a trajectory τ_m of length T_m, its overall advantage $A(\tau_m)$ is calculated as:

$$A(\tau_m) = \frac{1}{T_m}\sum_{n=1}^{T_m} A_{m,n} \tag{8}$$

This trajectory-level advantage $A(\tau_m)$ is then used as the training signal within the GRPO loss function. This fine-grained approach to advantage calculation provides a more precise and informative signal, enabling the model to learn not only which overall sequences are effective, but also to discern the value of the specific intermediate states that are most critical for constructing an optimal plan. We present the full pseudo-code in Algorithm 1

4 Experiments

4.1 Experimental Setup

Given that spatial reasoning in embodied intelligence often operates under stringent computational constraints, we focus on open-source models with manageable parameter scales. Specifically, we employ Qwen3-4B-Instruct-2507 as our base model, which is further optimized via our M-GRPO framework to serve as the high-level planner. The untrained version of the same model served as both the environment decomposition model and the action generation model. We deliberately exclude large-scale proprietary closed-source models as they are often unsuitable for real-time embodied tasks, where low-latency local execution, robust offline functionality, and strict data privacy are paramount. Furthermore, to ensure

our framework's generalizability across different open-source architectures, we evaluate it on the DeepSeek-R1-Distill-Llama-8B model family; due to space limits, these detailed results are provided in Appendix A.

Datasets We evaluate our HSRL framework across four planning benchmarks with increasing difficulty to assess its performance and generalization capability. First, we use Maze Navigation(Valmeekam et al., 2023), a classical task consisting of 1,090 10×10 grids. To create highly challenging scenarios, we configured 40% of the cells in each map as obstacles. The dataset is partitioned into 668 training and 422 testing instances. To validate the real-world applicability of our approach, inspired by (Zhao et al., 2025), we leverage the R2V dataset (Liu et al., 2017), which comprises 815 authentic architectural floorplans. We randomly selected a subset of 50 maps from this collection. Each floorplan was converted into a textual representation and scaled proportionally to a 20×20 resolution, followed by the random sampling of three pairs of start and goal positions for each layout. This procedure resulted in the evaluation benchmark for real-world indoor navigation tasks. Second, to assess out-of-distribution (OOD) generalization, we employ the Blocksworld benchmark(Saha et al., 2025), whose test set is intentionally more complex, featuring more blocks and requiring longer plans (7–10 steps) than the training set (1–6 steps). Finally, we validate our framework on the novel and highly challenging GameTraversalBenchmark (GTB)(Nasir et al., 2024). This benchmark contains 150 diverse maps with multiple objectives and paths exceeding 100 steps. As GTB lacks a training set, we evaluate our Maze-trained model in a zero-shot transfer setting to test its capabilities on complex, unseen tasks.

Baselines We compare HSRL against a diverse set of representative baselines. First, we compare it with foundational reasoning strategies, including the classic Chain-of-Thought (CoT)(Wei et al., 2022) and ReAct(Yao et al., 2023b), which interleaves reasoning traces with actions for improved synergy. We also include advanced reasoning and self-reflection methods like Inner Monologue(Huang et al., 2023), which enhances internal thought processes, and Reflexion(Shinn et al., 2023), which uses iterative self-correction to refine plans. For direct planning, we use ProgPrompt(Singh et al., 2023) as a strong representative of in-context learning-based approaches. Furthermore, we contrast HSRL with search-based methods like Tree Planner(Hu et al., 2024) and the hierarchical planner HyperTree(Gui et al., 2025), the latter of which is known to have limitations on spatial reasoning tasks. Finally, we include System-1.x(Saha et al., 2025), a powerful baseline meticulously fine-tuned on tasks similar to ours, which employs a controller to switch between "fast-thinking" and "slow-thinking" modes.

Evaluation metrics We evaluate our model's planning ability using metrics tailored to each benchmark. For the classical Maze and Blocksworld tasks, we measure the Completion Rate (CR), which is the percentage of successfully solved instances, and the Optimal Rate (OR), defined as the percentage of completed tasks where the plan length matches the shortest path computed by an A* search. For the more complex GameTraversalBenchmark (GTB), we adopt its official metrics. The primary metric is the GTB Score, a composite measure that assesses performance based on goal proximity, path length, and generation errors (see Appendix B.1). Additionally, we report Top-5 Accuracy, the fraction of tasks where the agent ends within five tiles of the target, to evaluate success in large-scale maps.

4.2 Implementation Detail

M-GRPO Finetuning. We fine-tuned the model using the trl library on eight NVIDIA A800 GPUs. The optimization was conducted using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We employed a learning rate of 1×10^{-6} following a cosine decay schedule. The hyperparameters τ is 1.0 and γ is 0.4. The Epoch Number was 1 and the Batch Size was 1. For the inference phase, the Temperature was set to 1.0 and the Num generations was 8.

Reward Function. To ensure training stability and prevent reward hacking, we implement a comprehensive reward function that weights task success against path length through multiple shaping terms. Specifically, to mitigate the tendency of RL agents to over-generate—a common phenomenon where models produce excessively long responses to maximize cumulative rewards—we introduce a stringent length penalty to encourage concise and effective reasoning. Detailed formulations of these composite reward functions are provided in Appendix C.1.

Table 1: Main results: Comparison of HSRL against baseline methods across various spatial reasoning benchmarks. **CR** and **OR** denote Completion Rate and Optimality Rate, respectively. The best performance in each category is highlighted in **bold**.

Method	Maze (10×10)		R2V (Real-World)		Blocksworld (5-7)		GTB	
	CR (%) ↑	OR (%) ↑	CR (%) ↑	OR (%) ↑	CR (%) ↑	OR (%) ↑	Score ↑	Top-5 Acc. ↑
Direct Answer	23.69	23.45	1.33	1.13	6.50	6.00	23.61	25.96
CoT	43.12	38.39	16.67	12.33	10.50	8.00	26.58	31.61
Reflexion	45.02	37.91	22.00	21.33	15.00	8.50	29.34	37.76
ReAct	53.80	26.06	16.00	13.33	8.00	3.00	20.40	39.85
ProgPrompt	34.60	33.41	43.33	42.00	9.50	6.00	22.45	26.12
Inner Monologue	54.03	34.60	27.33	16.67	4.00	0.00	19.18	21.41
System-1.x	54.74	36.02	50.67	49.33	27.00	14.50	27.73	30.01
HyperTree	37.91	22.98	25.33	19.67	8.00	3.50	25.81	26.67
Tree Planner	39.10	27.01	6.33	3.00	7.00	4.00	25.28	25.46
HSRL (Ours)	**61.37**	**47.87**	**62.67**	**55.33**	**30.50**	**21.00**	**32.69**	**44.41**

4.3 Results Analysis

The experimental results in Table 1 clearly demonstrate the effectiveness of our framework.

HSRL Significantly Improves Task Performance. HSRL achieves state-of-the-art (SOTA) performance across all tasks. In the Maze (10×10) task, it reaches a 61.37% completion rate, markedly outperforming System-1.x (54.74%) and ReAct (53.80%). On the large-scale real-world R2V dataset, HSRL achieves a 62.67% completion rate, maintaining high efficacy by decomposing long-horizon navigation to overcome the reasoning deterioration typical of monolithic methods. This advantage extends to the complex Blocksworld (30.50% completion rate) and the GameTraversal-Benchmark (GTB), where HSRL achieves the highest GTB Score (32.69 vs. Reflexion's 29.34) and Top-5 Accuracy (44.41%).

Superior Solution Optimality. Generating optimal paths is a crucial metric of planning intelligence. HSRL achieves the highest optimality rates in Maze (47.87%), R2V (55.33%), and Blocksworld (21.00%). By combining MCTS's forward-search with the LLM's general knowledge, our framework thoroughly explores the solution space, effectively avoiding local optima to devise concise and efficient paths.

Cross-Task Robustness and Generalization. The value of a general-purpose planning model lies in its cross-task generalization capability. As shown in the table, our method performs exceptionally well on the classical spatial reasoning tasks of Maze and Blocksworld, and the complex world knowledge required for the Game Travel Benchmark. In contrast, some baselines exhibit strong task-specific biases; for example, while ReAct performs reasonably well in Maze, its completion rate falls to 8.00% in Blocksworld, with an optimality rate of 3.00%. This comparison validates the robustness of our hierarchical framework, which consistently decomposes complex problems into manageable sub-goals for effective problem-solving, irrespective of the task modality. Moreover, our excellent performance in Blocksworld demonstrates strong out-of-distribution generalization capabilities.

4.4 Further Analysis

Ablation Study. To validate the contribution of each component, we conducted a comprehensive ablation study (Table 2). The results reveal a clear hierarchy of importance. Replacing the modified UCT with the standard UCT significantly degrades performance on the challenging GTB task. Removing the MCTS module—thereby reverting M-GRPO to a standard GRPO training paradigm (HSRL w/o MCTS)—leads to a notable decline in both success and optimality, confirming that its systematic, forward-looking search is crucial for exploring diverse solution pathways. Further removing the M-GRPO policy optimization, where the base model is used directly as the high-level planner without further training (HSRL Untrained) causes a precipitous performance collapse, especially in optimality (e.g., Maze optimality plummets from 47.87% to 14.22%). This demonstrates that M-GRPO is the core engine that translates the rich search experience from MCTS into a refined planning intuition, endowing the LLM with the ability to generate high-quality, task-aligned

Table 2: Ablation results across four benchmarks. **CR** and **OR** denote completion and optimality rates, respectively; bold indicates the best result.

Model Configuration	Maze (10×10)		R2V (Real-World)		Blocksworld (5-7)		GTB	
	CR (%) ↑	OR (%) ↑	CR (%) ↑	OR (%) ↑	CR (%) ↑	OR (%) ↑	Score ↑	Top-5 Acc. ↑
State-Hierarchical Only	50.24	13.03	52.33	25.67	10.00	9.50	26.16	29.85
HSRL (Untrained)	54.50	14.22	60.33	28.00	12.50	9.50	26.96	32.38
HSRL (w/o MCTS)	55.21	45.97	58.00	54.67	28.00	15.00	27.80	38.09
HSRL (Standard UCT)	60.43	46.44	59.33	55.00	29.50	18.00	30.65	40.29
HSRL (Ours)	**61.37**	**47.87**	**62.67**	**55.33**	**30.50**	**21.00**	**32.69**	**44.41**

sub-goals. Finally, the performance of the State-Hierarchical Only configuration—which eliminates environment-level hierarchy and relies on the base model for both intermediate waypoint identification and low-level action generation—still significantly surpasses direct answering methods, and the inclusion of environment-hierarchical approaches effectively improves task completion.

Fair Comparison. To ensure a fair evaluation and address concerns regarding unequal computational budgets, we validate our framework across both inference-only and training-based settings. In the inference-only track, our zero-shot "HSRL (Untrained)" relies solely on the proposed state and environment decomposition prompting, yet it achieves a 54.50% completion rate on the Maze task, significantly outperforming strong baselines like CoT (43.12%), ProgPrompt (34.60%), and ReAct (53.80%). Furthermore, in the training-based track, we compare HSRL against System-1.x under identical training datasets and interaction budgets. Under these strictly controlled conditions, HSRL still yields substantial improvements. These dual-track results confirm that our performance gains stem fundamentally from the architectural advantages of hierarchical decomposition and the superior optimization logic of M-GRPO, rather than computational disparities.

Robustness and Qualitative Insights. As shown in Figure 4, HSRL's performance degrades far more gracefully with increasing task difficulty than baselines, maintaining a substantial advantage on challenging Maze instances. This resilience stems fundamentally from improved high-level planning. Qualitatively (Figure 5), while an untrained model generates ill-conceived sub-goals causing failure, the M-GRPO trained HSRL produces a strategic path. This synergy of quantitative robustness and qualitative intelligence validates HSRL's effectiveness in complex, long-horizon planning.

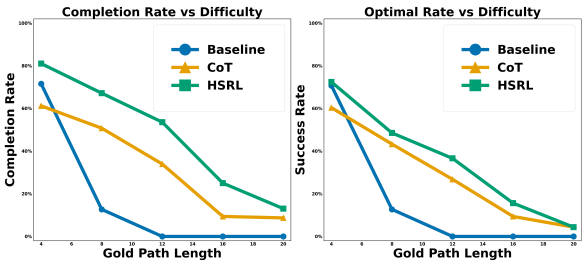

Figure 4: Performance degradation as task difficulty increases. HSRL shows greater robustness.

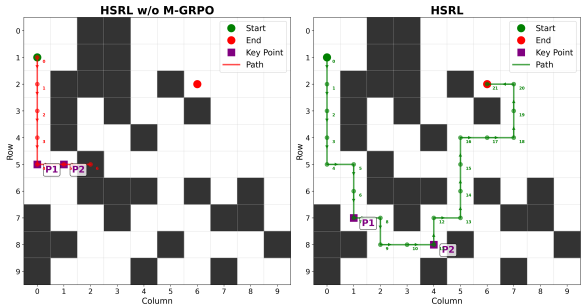

Figure 5: Qualitative comparison of a generated plan before (left) and after (right) M-GRPO training.

5 Conclusion

This paper introduced HSRL, a hierarchical framework that overcomes LLM spatial reasoning limitations through state and environment decomposition. By combining MCTS-guided GRPO algorithm, HSRL achieves SOTA performance across navigation and strategic benchmarks. Our results show substantial gains in completion rates and path optimality over existing methods. This work establishes a scalable approach for integrating LLMs into complex physical environments and embodied intelligence tasks.

Limitations

Despite the advances in spatial reasoning achieved by this study, several limitations remain. First, M-GRPO incurs additional training overhead because integrating MCTS into policy optimization requires

extensive path simulations, making it more time-consuming than conventional fine-tuning. However, this cost is confined to offline training and does not affect inference efficiency. Second, because large-scale datasets containing purely textual descriptions of 3D environments are scarce, our evaluation focuses mainly on 2D indoor navigation tasks, e.g., R2V. Extending the framework to complex 3D representations and multimodal environments remains an important direction for future work. Third, although HSRL excels at complex spatial reasoning, its applicability is inherently task-dependent. The framework relies on spatial topology to identify geometric intermediate states and construct localized sub-environments, reducing noise from tightly coupled actions. While effective in continuous domains without clear semantic boundaries, HSRL is not a universal planner. In abstract logic or semantic planning tasks, where sub-goals are loosely coupled and naturally separated linguistically, geometric decomposition offers limited benefits. Future work could explore adaptive mechanisms that switch between geometric and semantic decomposition according to task context.

Acknowledgments

We would like to express my deepest gratitude to the RedBirdMPhil Program at the HongKong University of Science and TechnologyGuangzhou for providing me with generous support, resources, and funding, which have been instrumental in the successful completion of my research.

We would like to express my sincere gratitude to Guangzhou SmarTech Automation Co., Ltd. for their strong technical support, valuable real-world data, and practical application scenarios, which were instrumental in translating this research into practice.

Sihong Xie is supported by the Department of Science and Technology of Guangdong Province (Grant No.2023CX10X079), and the Guangdong Basic and Applied Basic Research Foundation (Grant No.2025A1515220171), in part by Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence.

References

Mohamed Aghzal, Erion Plaku, and Ziyu Yao. 2024a. Can large language models be good path planners? a benchmark and investigation on spatial-temporal reasoning. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Mohamed Aghzal, Erion Plaku, and Ziyu Yao. 2024b. Look further ahead: Testing the limits of gpt-4 in path planning. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 1020–1027. IEEE.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976. Association for Computational Linguistics.

Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *2024 IEEE International conference on robotics and automation (ICRA)*, pages 6695–6702. IEEE.

Alan Dao and Dinh Bach Vu. 2025. Alphamaze: Enhancing large language models' spatial intelligence via grpo. *Preprint*, arXiv:2502.14669.

Hourui Deng, Hongjie Zhang, Jie Ou, and Chaosheng Feng. 2025. Can llm be a good path planner based on prompt engineering? mitigating the hallucination for path planning. In *International Conference on Intelligent Computing*, pages 3–15. Springer.

J. Duan, S. E. Li, Y. Guan, Q. Sun, and B. Cheng. 2020. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data. *IET Intelligent Transport Systems*, 14(5):297–305.

Lutfi Eren Erdogan, Hiroki Furuta, Sehoon Kim, Nicholas Lee, Suhong Moon, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. 2025. Plan-and-act: Improving planning of agents for long-horizon tasks. In *Forty-second International Conference on Machine Learning*.

Nanxu Gong, Chandan K. Reddy, Wangyang Ying, Haifeng Chen, and Yanjie Fu. 2025. Evolutionary large language model for automated feature transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16844–16852.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rStar-math: Small LLMs can master math reasoning with self-evolved deep thinking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 20640–20661.

Runquan Gui, Zhihai Wang, Jie Wang, Chi Ma, Huiling Zhen, Mingxuan Yuan, Jianye Hao, Defu Lian, Enhong Chen, and Feng Wu. 2025. Hypertree planning: Enhancing LLM reasoning via hierarchical thinking. In *Forty-second International Conference on Machine Learning*.

Mengkang Hu, Yao Mu, Xinmiao Chelsey Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. 2024. Tree-planner: Efficient close-loop task planning with large language models. In *The Twelfth International Conference on Learning Representations*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2023. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N. Rafferty, John Stamper, and Michael Liut. 2024. Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, pages 86–97, New York, NY, USA. Association for Computing Machinery.

Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. 2017. Raster-to-vector: Revisiting floorplan transformation. In *Proceedings of the IEEE international conference on computer vision*, pages 2195–2203.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

Xinyang Lu, Flint Xiaofeng Fan, and Tianying Wang. 2023. Action and trajectory planning for urban autonomous driving with hierarchical reinforcement learning. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*.

Runyu Ma, Jelle Luijkx, Zlatan Ajanović, and Jens Kober. 2025. Explorllm: Guiding exploration in reinforcement learning with large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9011–9017. IEEE.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2498–2518. PMLR.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with cogeval. In *Advances in Neural Information Processing Systems*, volume 36, pages 69736–69751. Curran Associates, Inc.

Muhammad Umair Nasir, Steven James, and Julian Togelius. 2024. Gametraversalbenchmark: Evaluating planning abilities of large language models through traversing 2d game maps. In *Advances in Neural Information Processing Systems*, volume 37, pages 31813–31827. Curran Associates, Inc.

Matthew Renze and Erhan Guven. 2024. Self-reflection in LLM agents: Effects on problem-solving performance. *Preprint*, arXiv:2405.06682.

Swarnadeep Saha, Archiki Prasad, Justin Chen, Peter Hase, Elias Stengel-Eskin, and Mohit Bansal. 2025. System 1.x: Learning to balance fast and slow planning with language models. In *The Thirteenth International Conference on Learning Representations*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE.

Wanxin Tian, Shijie Zhang, Kevin Zhang, Xiaowei Chi, Chun-Kai Fan, Junyu Lu, Yulin Luo, Qiang Zhou, Yiming Zhao, Ning Liu, Siyu Lin, Zhiyuan Qin, Xiaozhu Ju, Shanghang Zhang, and Jian Tang. 2025. SEEA-R1: Tree-structured reinforcement fine-tuning for self-evolving embodied agents. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 78458–78499. Curran Associates, Inc.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models - a critical investigation. In *Advances in Neural Information Processing Systems*, volume 36, pages 75993–76005. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's

eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 90277–90317. Curran Associates, Inc.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Shangzi Xue, Zhenya Huang, Jiayu Liu, Xin Lin, Yuting Ning, Binbin Jin, Xin Li, and Qi Liu. 2024. Decompose, analyze and rethink: Solving intricate problems with human-like reasoning cycle. In *Advances in Neural Information Processing Systems*, volume 37, pages 357–385. Curran Associates, Inc.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5186–5219, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yubo Zhao, Qi Wu, Yifan Wang, and Yu-Wing Tai. 2025. Navigating motion agents in dynamic and cluttered environments through LLM reasoning. *Preprint*, arXiv:2503.07323.

Wei Zhu and Mitsuhiro Hayashibe. 2023. A hierarchical deep reinforcement learning framework with high efficiency and generalization for fast and safe navigation. *IEEE Transactions on Industrial Electronics*, 70(5):4962–4971.

Appendix

A Generalization on DeepSeek Model

To further evaluate the architecture-agnostic nature and generalization capability of our HSRL framework, we conduct additional experiments using the DeepSeek-R1-Distill-Llama-8B model as the backbone. We maintain the same experimental settings and evaluation metrics (CR, OR, Score, and Top-5 Acc.) across four diverse spatial reasoning benchmarks: Maze, R2V, Blocksworld, and GTB.

As shown in Table 3, HSRL consistently outperforms other competitive baselines when integrated with the DeepSeek architecture. Notably, on the Blocksworld (5-7) task, HSRL achieves a CR of 33.00% and an OR of 25.00%, marking a significant improvement over the Direct Answer and CoT methods. These results demonstrate that the hierarchical task decomposition and M-GRPO optimization in our framework are not biased toward a specific model family (e.g., Qwen) but provide a robust enhancement for spatial planning across different Large Language Models.

B More Background

B.1 GTB Score

$$\text{GTB_Score} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\Delta R^{(m)}}\Big(R^{(m)} - \text{LLM}_{PL}^{(m)} - \varepsilon^{(m)} - R_{\min}^{(m)}\Big) \tag{9}$$

where:

- M = total number of maps in the dataset.
- $R^{(m)}$ = reward obtained for map m, determined by the final distance d to the objective:

$$R^{(m)} = \begin{cases} +200, & d = 0 \\ +100, & d = 1 \\ +50, & d \in [2,3] \\ +25, & d \in [3,5] \\ -50, & d \in [5,8] \\ -100, & d \geq 8 \end{cases}$$

- $\text{LLM}_{PL}^{(m)}$ = path length taken by the LLM agent on map m.
- $\varepsilon^{(m)}$ = total generation errors made by the LLM on map m.
- $R_{\max}^{(m)} = 200 - A_{PL}^{*}(m)$, the maximum achievable reward (perfect path with no errors), where $A_{PL}^{*}(m)$ is the optimal path length computed by an A* agent.
- $R_{\min}^{(m)} = -100 - A_{PL}^{*}(m) - \varepsilon_{\max}^{(m)}$, the minimum achievable reward (farthest position, maximal path cost, and maximal errors).

Table 3: Generalization results on the DeepSeek model across different spatial reasoning benchmarks. Bold denotes the best performance.

Method	Maze (10×10)		R2V (Real-World)		Blocksworld (5-7)		GTB	
	CR (%) ↑	OR (%) ↑	CR (%) ↑	OR (%) ↑	CR (%) ↑	OR (%) ↑	Score ↑	Top-5 Acc. ↑
Direct Answer	11.85	8.77	2.67	2.67	8.50	8.50	20.76	22.41
CoT	13.50	13.03	10.67	8.67	13.50	12.50	24.11	24.59
System-1.x	18.72	15.40	20.67	19.33	29.50	23.00	25.89	32.97
HSRL (Ours)	**23.70**	**20.38**	**28.00**	**22.67**	**33.00**	**25.00**	**27.92**	**34.82**

C MORE IMPLEMENTATION DETAILS

C.1 M-GRPO Reward

For a sampled completion $completion_i$, we parse its anchor list $A_i = [a_{i1}, a_{i2}, \ldots, a_{in}]$. If the anchor list cannot be parsed, we directly assign a fixed penalty; otherwise, the reward score is computed by a signed power transformation to enlarge the margin between high- and low-quality completions:

$$R_i = \begin{cases} \texttt{PARSE_FAIL_PENALTY}, & \text{if } A_i = \emptyset, \\ \operatorname{sign}(z_i)\,|z_i|^p, & \text{otherwise}, \end{cases} \tag{10}$$

where $\operatorname{sign}(z_i)$ preserves the direction of z_i, and $p > 1$ amplifies its magnitude non-linearly.

The raw score z_i aggregates the quality of anchors visited by a trajectory, while discouraging the use of overly many anchors through a penalty term:

$$z_i = \sum_{a \in A_i} \overline{r}(a) - \alpha \cdot \max\big(0,\, |A_i| - A_{\text{expected}}\big). \tag{11}$$

To evaluate each anchor consistently, we first assign each completion an initial reward r_i according to its alignment with the Manhattan distance of the optimal A^* path. :

$$r_i = BASIC_QUALITY_SCORE - \Big|\sum_{a_i \in A_i} |a_{i+1} - a_i| - \sum_{\hat{a_i} \in A^*} |\hat{a_{i+1}} - \hat{a_i}|\Big|. \tag{12}$$

Each anchor reward $\overline{r}(a)$ is then defined as the average quality of all completions that pass through it, reflecting a consensus measure across different trajectories:

$$\overline{r}(a) = \frac{\sum_{i:\, a \in A_i} r_i}{|\{\, i \mid a \in A_i \,\}|}. \tag{13}$$

D Declaration of Use of AI Assistants

During the preparation of this work, the authors used Gemini-3-Flash to polish the manuscript for language clarity and to assist with LaTeX formatting. Following the use of this tool, the authors reviewed and edited the content as needed and take full responsibility for the final version of the paper.

E Quantitative Analysis of Computational Costs

Training Time Comparison. Table 4 reports the training time for Standard GRPO and our M-GRPO under identical hardware and batch size configurations.

Table 4: Training time comparison (same hardware and batch size).

Environment	Standard GRPO	M-GRPO (Ours)
Maze	1h 35min	4h 49min
Blocksworld	29min	2h 05min

Inference Efficiency. Crucially, inference remains highly efficient and comparable to standard Chain-of-Thought (CoT). Table 5 presents the average inference time per task.

Table 5: Average inference time comparison.

Task	CoT	HSRL
Blocksworld	19.20s	20.18s
Maze	22.34s	28.23s

F Prompts and Examples

High-level Planning Prompt for Maze

```
### Role
You are an expert high-level path planner. You must strictly adhere to the requirements outlined in
the system instructions and tasks I have provided to you.

### Instructions
1. Your task is to plan a feasible, obstacle-free path for a single agent in a given 10x10 grid
environment, from a start to an end point.
2. The path should be defined by a series of key anchor point coordinates.
3. You must identify exactly {{num_anchors}} feasible intermediate anchor points for the given task.
These anchor points should be the key turning points of the path used to navigate around obstacles
or toward the goal.

### Anchor Point Selection Strategy
- The path does not have to be the shortest path. The priority is feasibility and safety (avoiding
all obstacles).
- Explore multiple valid paths and select a reasonable one to define your anchor points.
- Anchor points should be strategically located at important positions around obstacles.

### Output Format
- You must strictly follow the format below to output the list of anchor points.
- Do not provide any explanation or text other than the final trajectory list.
- Directly output the result in the given format:
<trajectory for planning> = [(start_x, start_y), (anchor_1_x, anchor_1_y), ..., (end_x, end_y)]
---
### Examples

**Example 1:**
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (4,7), (8,6), (3,3),
(9,5), (8,9), (1,1), (5,4), (1,3), (9,9), (4,1), (5,7), (1,6), (9,0), (8,3), (0,0), (7,1), (4,6),
(5,0), (2,5) and (4,0). Go from (2,1) to (0,2).
<trajectory for planning> = [(2,1),(2,2),(1,2),(0,2)]

**Example 2:**
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (0,7), (3,2), (0,4),
(3,4), (4,6), (7,2), (7,3), (2,0), (3,9), (9,3), (8,2), (9,5), (8,4), (7,5), (4,8), (5,2), (5,5),
(7,8), (6,3) and (9,8). Go from (6,8) to (6,1).
<trajectory for planning> = [(6,0),(6,4),(5,3),(6,1)]

**Example 3:**
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (8,5), (7,2), (1,7),
(2,0), (3,2), (5,0), (1,9), (3,3), (3,6), (4,7), (0,3), (5,7), (5,3), (4,6), (2,8), (4,3), (9,0),
(7,5), (5,5) and (8,9). Go from (0,8) to (7,1).
<trajectory for planning> = [(0,8),(0,4),(1,1),(7,1)]
---

### Task to Solve
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (9,6), (1,0), (3,7),
(4,4), (9,1), (4,0), (3,4), (8,9), (7,1), (5,1), (3,6), (4,9), (4,8), (0,1), (6,4) and (0,0). Go
from (4,1) to (6,8).
<trajectory for planning> = [(4,1),(5,3),(5,6),(6,8)]
```

Figure 6: High-level Planning Prompt for Maze.

Environment Decomposition Prompt for Maze

```
Given the following list of coordinate points:
[(9,6), (1,0), (3,7), (4,4), (9,1), (4,0), (3,4), (8,9), (7,1), (5,1), (3,6), (4,9), (4,8), (0,1),
(6,4), (0,0)]


Please select all points from the list that satisfy the following condition:
The x-coordinate (x) of the point must be within the closed interval from 4 to 5.


Please strictly follow this format for the output, including only the selected points:
### Output Format
<obstacles> = [ (x1, y1), (x2, y2), ..., (xn, yn)]


<assistant>
<obstacles> = [(4,4),(4,0),(5,1),(4,9),(4,8)]
```

Environment Decomposition Prompt for Maze

```
Given the following list of coordinate points:
[(9,6), (1,0), (3,7), (4,4), (9,1), (4,0), (3,4), (8,9), (7,1), (5,1), (3,6), (4,9), (4,8), (0,1),
(6,4), (0,0)]


Please select all points from the list that satisfy the following condition:
The y-coordinate (y) of the point must be within the closed interval from 1 to 3.


Please strictly follow this format for the output, including only the selected points:
### Output Format
<obstacles> = [ (x1, y1), (x2, y2), ..., (xn, yn)]


<assistant>
<obstacles> = [(9,1),(7,1),(5,1),(0,1)]
```

Figure 7: Environment Decomposition Prompt for Maze.

Low-level Execution Prompt for Maze

```
### Role
You are an low-level path planner located in a 10 by 10 world. You must strictly adhere to the requirements
of the tasks I have provided to you.

### Environment
Provide a sequence of actions to navigate a world to reach a goal. (0,0) is located in the upper-left corner
and (9,9) lies in down-right corner.

### Rules
- <left = (0,-1)>
- <right = (0,+1)>
- <up = (-1,0)>
- <down = (+1,0)>

### Output Format
Actions = [action_0 action_1 … action_n]

Here are some examples:

###
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (2,1). Go from (0,1) to (3,4).

Actions = [right right right down down down]

###
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (1,5) and (1,2). Go from (5,4)
to (0,5).

Actions = [up up up up up right]

###
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (0,3), (2,5) and (5,2). Go
from (4,2) to (0,5)

Actions = [up up up right right up right]

### Task to Solve subtask_1

Task: You are in a 10 by 10 world. There are obstacles that you must avoid
at:(4,4),(4,0),(5,1),(4,9),(4,8),(9,1),(7,1),(5,1),(0,1). Go from (4,1) to (5,3).

Actions = [right right down]
```

Figure 8: Low-level Execution Prompt for Maze.

High-level Planning Prompt for Blocksworld

```
### Role
You are a high-level Blocksworld planner. You must strictly adhere to the requirements outlined in
the system instructions and tasks I have provided to you.

### Instructions
1. Plan a feasible sequence of block configurations from an initial state to a goal state.
2. Define the plan with exactly {{num_anchors}} intermediate stack states.
3. Intermediate states should be key subgoals (e.g., clearing a block or forming partial stacks).

### Strategy
- The sequence need not be shortest; feasibility and clarity are the priority.
- Choose anchor states that mark meaningful progress toward the goal.

### Output Format
- You must strictly follow the format below to output the list of anchor states.
- Do not provide any explanation or text other than the final output list.
- Directly output the result in the given format:
Output = [initial_state, anchor_state1, ..., goal_state]
---
### Examples

**Example 1:**
The initial state:
A is on the table. B is on A. B is clear.
The goal is:
B is on the table. A is on B. A is clear.

Output = ['A is on the table. B is on A. B is clear.','A is on the table. A is clear. B is on the
table. B is clear.','B is on the table. A is on B. A is clear.']

**Example 2:**
The initial state:
C is on the table. D is on C. D is clear.
The goal is:
C is on D. D is on the table. D is clear.

Output = ['C is on the table. D is on C. D is clear.','C is on the table. C is clear. D is on the
table. D is clear.','C is on D. D is on the table. D is clear.']

**Example 3:**
The initial state:
B is on the table. C is on B. A is on C. A is clear.
The goal is:
B is on the table. B is clear.
C is on B. A is on C. A is clear.

Output = ['B is on the table. C is on B. A is on C. A is clear.','B is on the table. B is clear. C
is on the table. C is clear. A is on the table. A is clear.','B is on the table. C is on B. C is
clear. A is on the table. A is clear.']
---
### Task to Solve
The initial state:
B is on the table. B is clear.
D is on the table. C is on D. E is on C. A is on E. A is clear.

The goal is:
C is on the table. E is on C. D is on E. B is on D. A is on B. A is clear.

Output = ['The hand is empty. B is on the table. B is clear. D is on the table. C is on the table. C
is clear. E is on the table. E is clear. A is on the table. A is clear.', 'C is on the table. E is
on C. D is on E. B is on the table. B is clear. A is on the table. A is clear.']
```

Figure 9: High-level Planning Prompt for Blocksworld.

Environment Decomposition Prompt for Blocksworld

```
Given the 2 state of the same group of Blocks:
<1>'B is on the table. B is clear. D is on the table. C is on the table. C is clear. E is on the
table. E is clear. A is on the table. A is clear.’
<2>'C is on the table. E is on C. D is on E. B is on the table. B is clear. A is on the table. A is
clear.'

Please select all state pairs (including block-block pairs, block-table pairs, and block-clear
pairs) whose relative positions remain unchanged.

Please strictly follow this format for the output, including only the selected pairs:
### Output Format
<pairs> = [(Object1,Object2),(Object3,Object4),...]


<assistant>
<pairs> = [(B,table),(A,table),(C,table)]
```

Figure 10: Environment Decomposition Prompt for Blocksworld.

Low-level Execution Prompt for Blocksworld

```
You are a Blocks World action planner. You will be given an initial state and a goal state.
Provide a sequence of actions to move the blocks to reach a goal similarly to the examples below.
Do not include any extra text or explanations.

Here are some examples:
###
The initial state:
The hand is empty.
B is on the table. A is on B. C is on A. C is clear.

The goal is:
B is on the table. A is on B. A is clear.
C is on the table. C is clear.

<Observation>: B is still on table; A is still on B
Actions: Move C from A to table

###
The initial state:
The hand is empty.
B is on the table. C is on B. D is on C. A is on D. A is clear.

The goal is:
A is on the table. A is clear.
C is on the table. B is on C. D is on B. D is clear.

<Observation>: A is still clear
Actions: Move A from D to table | Move D from C to table | Move C from B to table | Move B from
table to C | Move D from table to B


Now, here is your task:
###
The initial state:
B is on the table. B is clear.
D is on the table. D is clear.
C is on the table. C is clear.
E is on the table. E is clear.
A is on the table. A is clear.

The goal is:
C is on the table. E is on C. D is on E. D is clear.
B is on the table. B is clear.
A is on the table. A is clear.

<Observation>: B is still on table; A is still on table; C is still on table
Actions: Move E from table to C | Move D from table to E
```

Figure 11: Low-level Execution Prompt for Blocksworld.

High-level Planning Prompt for GTB

```
### Role
You are an expert high-level planner for a 2D-gird game. You must strictly adhere to the
requirements outlined in the system instructions and tasks I have provided to you.

### Instructions
1. Your task is to plan a feasible sequence of moves in a {{world_height}}×{{world_length}} grid
environment, from a start state to a goal state.
2. The plan should be defined by a series of key anchor coordinates.
3. You must identify exactly {{num_anchors}} feasible anchor states for the given task. These
should be important turning points or subgoals used to navigate around obstacles or toward
objectives.

### Anchor State Selection Strategy
- The plan does not need to be the shortest; the priority is feasibility and safety (avoiding all
obstacles).
- Anchor states should be strategically located at key subgoals: clearing an obstacle, moving
around blocking tiles, or forming partial progress toward objectives.
- Explore multiple valid strategies and select one reasonable plan.

### Reward Context
- Rewards are given as follows:
  {{reward_design}}
  {{reward_feedback}}
- You are also given information about your previous attempt:
  - Actions generated: {{total_actions[list(objective_tile_dict.keys())[i]]}}
  - Start position: {{prev_protagonist_position}}
  - End position: {{protagonist_position}}
  - Distance from objective: {{distance_from_objective}}
  - Objective location: {{list(objective_tile_dict.values())[i]}}
  - GTB Reward received: {{reward_this_objective[list(objective_tile_dict.keys())[i]]}}

### Output Format
- Strictly output the result in the following format, without any explanation:
<trajectory for planning> = [(start_x, start_y), (anchor_1_x, anchor_1_y), ..., (end_x, end_y)]
---
### Examples

**Example 1:**
Task: You are in a 10 by 10 world. There are obstacles that you have to avoid at: (4,7), (8,6),
(3,3), (9,5), (8,9), (1,1), (5,4), (1,3), (9,9), (4,1), (5,7), (1,6), (9,0), (8,3), (0,0), (7,1),
(4,6), (5,0), (2,5) and (4,0). Go from (2,1) to (0,2).
<trajectory for planning> = [(2,1),(2,2),(0,2)]

**Example 2:**
Task: You are in a 10 by 10 world. There are obstacles that you have to avoid at: (0,7), (3,2),
(0,4), (3,4), (4,6), (7,2), (7,3), (2,0), (3,9), (9,3), (8,2), (9,5), (8,4), (7,5), (4,8), (5,2),
(5,5), (7,8), (6,3) and (9,8). Go from (6,8) to (6,1).
<trajectory for planning> = [(6,0),(6,4),(5,3),(4,2),(6,1)]

**Example 3:**
Task: You are in a 10 by 10 world. There are obstacles that you have to avoid at: (8,5), (7,2),
(1,7), (2,0), (3,2), (5,0), (1,9), (3,3), (3,6), (4,7), (0,3), (5,7), (5,3), (4,6), (2,8), (4,3),
(9,0), (7,5), (5,5) and (8,9). Go from (0,8) to (7,1).
<trajectory for planning> = [(0,8),(0,4),(1,1),(7,1)]
---
### Task to Solve
Task: {{task}}
<trajectory for planning> =
```

Figure 12: High-level Planning Prompt for GTB.

Environment Decomposition Prompt for GTB

```
Given the following list of coordinate points:
[{obstacles_this_object}]


Please select all points from the list that satisfy the following condition:
The x-coordinate (x) of the point must be within the closed interval from {x_this_object_min} to
{x_this_object_max}.


Please strictly follow this format for the output, including only the selected points:
### Output Format
<obstacles> = [ (x1, y1), (x2, y2), ..., (xn, yn)]


<assistant>
<obstacles> =
```

Environment Decomposition Prompt for GTB

```
Given the following list of coordinate points:
[{obstacles_this_object}]


Please select all points from the list that satisfy the following condition:
The y-coordinate (y) of the point must be within the closed interval from {y_this_object_min} to
{y_this_object_max}.


Please strictly follow this format for the output, including only the selected points:
<obstacles> = [ (x1, y1), (x2, y2), ..., (xn, yn)]


<assistant>
<obstacles> =
```

Figure 13: Environment Decomposition Prompt for GTB.

Low-level Execution Prompt for GTB

```
### Role
You are an low-level path planner located in a {world_height} by {world_width}world. You must
strictly adhere to the requirements of the tasks I have provided to you.

### Environment
Provide a sequence of actions to navigate a world to reach a goal. (0,0) is located in the upper-
left corner and {(world_height,world_width)} lies in down-right corner.

### Rules
- <left = (0,-1)>
- <right = (0,+1)>
- <up = (-1,0)>
- <down = (+1,0)>

### Output Format
Actions = [action_0 action_1 … action_n]

Here are some examples:

###
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (2,1). Go from
(0,1) to (3,4).

Actions = [right right right down down down]

###
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (1,5) and (1,2).
Go from (5,4) to (0,5).

Actions = [up up up up up right]

###
Task: You are in a 10 by 10 world. There are obstacles that you must avoid at: (0,3), (2,5) and
(5,2). Go from (4,2) to (0,5)

Actions = [up up up right right up right]

### Task to Solve subtask_1

Task: {GTB_sub_task}

Actions =
```

Figure 14: Low-level Execution Prompt for GTB.