Does RLVR Extend Reasoning Boundaries? Investigating Capability Expansion in Vision-Language Models

Minghe Shen[1,2], **Zhuo Zhi**[1,2], **Chonghan Liu**[3], **Shuo Xing**[4], **Zhengzhong Tu**[4], **Che Liu**[5*]

[1]University College London [2]Samsung R&D Institute UK [3]University of California, Los Angeles

[4]Texas A&M University [5]Imperial College London

minghe.shen.24@ucl.ac.uk che.liu21@imperial.ac.uk

Abstract

Recent studies posit that Reinforcement Learning with Verifiable Rewards (RLVR) primarily amplifies behaviors inherent to the pre-training distribution rather than inducing new capabilities, but these insights are predominantly limited to language-only domains, leaving the dynamics of visual-centric spatial reasoning under-explored. To examine the impact of RLVR on the capability boundaries of Vision-Language Models (VLMs), we introduce **Ariadne**, a controlled framework based on synthetic maze navigation where the reasoning difficulty is precisely regulated by path length and the number of turns. We demonstrate that applying RLVR extends the spatial reasoning boundary, achieving success on problems where the base policy VLM consistently attains 0% accuracy despite increasing pass@k sampling budgets, indicating that the optimized policy successfully navigates search spaces that were effectively unreachable by the base distribution. Furthermore, despite being trained exclusively on synthetic mazes, we evaluate the model on two real-world navigation benchmarks (MapBench and ReasonMap) in a zero-shot setting. The observed improvements in these out-of-domain tasks suggest genuine spatial reasoning capability expansion rather than mere sampling efficiency. Our code is available at: https://github.com/MingheShen/Ariadne

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a transformative paradigm for enhancing the reasoning capabilities of Large Language Models (LLMs), enabling breakthroughs in mathematics and coding without reliance on human-annotated CoT data (DeepSeek-AI et al., 2025; Shao et al., 2024; Schulman et al., 2017; Liu et al., 2025a; Wang et al., 2025b). Yet, as these methods gain prominence, a critical debate has surfaced regarding the underlying mechanisms of these improvements: *Does RLVR facilitate the acquisition of novel reasoning primitives, or does it primarily optimize the sampling efficiency of behaviors already latent within the base policy?*

Recent studies in language-dominant tasks suggest the latter (Yue et al., 2025). Research shows that on math benchmarks, base models often retain sufficient probability mass over valid trajectories, such that RLVR primarily improves sampling efficiency by amplifying existing behaviors rather than extending reasoning boundaries (Yue et al., 2025; Xu et al., 2025). We argue that this conclusion depends on the evaluation domain. Tasks such as mathematics and coding, which are well represented in pre-training and prone to contamination, constitute high-support regimes where correct solutions already lie within the base distribution (Matton et al., 2024; Liang et al., 2025; Mirzadeh et al., 2025; Wu et al., 2026), making RLVR an efficiency mechanism. In contrast, visual-centric spatial reasoning tasks often fall into low-support regimes, where prior coverage is minimal and valid solutions are effectively absent. For example, in long-horizon navigation, base models fail to produce valid trajectories even with large pass@k budgets (Feng et al., 2025; Xing et al., 2025). Success in such settings indicates that RLVR induces new reasoning trajectories, reflecting genuine capability expansion within spatial reasoning rather than improved sampling.

To probe this distinction, we introduce **Ariadne**[1], a controlled experimental framework based on synthetic maze navigation (Dao and Vu, 2025). Unlike noisy real-world benchmarks, Ariadne allows us to manually characterize the "reasoning boundary", the complexity threshold (path length and turns) where the base model's success rate diminishes to

[*]Corresponding author.

[1]Named after the mythological figure who provided the thread to navigate the Labyrinth, symbolizing our focus on guiding VLMs through structured spatial reasoning.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45317–45331

zero in high-complexity configurations. We train VLMs using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) within this setting to investigate whether the optimized policy can successfully navigate search spaces that lack effective support in the base distribution.

Our results provide decisive evidence against the efficiency-centric view in the VLM context. We demonstrate that RLVR-trained models achieve non-trivial success rates on spatial problems where the base model consistently fails (0% accuracy) despite extensive sampling. Furthermore, we show that this learned spatial logic is not an artifact of the synthetic environment; the model exhibits zero-shot transfer improvements on real-world map navigation benchmarks (MapBench and ReasonMap), suggesting the acquisition of robust spatial reasoning primitives.

In summary, our contributions are as follows:

- We introduce **Ariadne**, a controlled framework based on synthetic maze navigation designed to probe the reasoning boundaries of VLMs through verifiable rewards and precise difficulty regulation.
- We demonstrate that RLVR extends the spatial reasoning boundary beyond the effective support of the base policy, achieving success on complex instances where the base model fails despite increasing pass@k budgets, with validation via zero-shot transfer to real-world navigation tasks.

2 Related Work

RLVR and Post-Training in LLMs. Reinforcement learning post-training has become the standard paradigm for eliciting complex reasoning in Large Language Models (LLMs). Early approaches utilized PPO with learned reward models (Ouyang et al., 2022; Wang et al., 2023), but recent advances have shifted toward Reinforcement Learning with Verifiable Rewards (RLVR) in domains with deterministic correctness, such as mathematics and code (Lambert et al., 2025; Hui et al., 2024). Notable implementations, including OpenAI's o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), employ algorithms like Group Relative Policy Optimization (GRPO) to stabilize training without value networks (Shao et al., 2024). These successes have spurred extensive research into optimizing inference-time compute and self-correction strategies (Zhao et al., 2025; Deng et al., 2025; Yang et al., 2024; Ying et al., 2024).

RLVR in Vision-Language Models. The application of RLVR to the vision-language domain is an emerging frontier. Recent studies have applied RLVR to enhance multimodal reasoning in tasks such as Visual Question Answering (VQA) and chart understanding (Huang et al., 2025; Wang et al., 2025a; Wan et al., 2025). However, these evaluations remain predominantly language-centric, often focusing on tasks where visual input serves merely as context for symbolic reasoning rather than requiring executable spatial planning. Consequently, it remains an open question whether RLVR meaningfully improves spatial reasoning capabilities in vision-language settings, such as long-horizon navigation, where pre-trained models frequently fail to generate valid action sequences due to a lack of grounded spatial understanding (Zhang et al., 2025; Hou et al., 2025; Feng et al., 2025; Xing et al., 2025).

The Efficiency vs. Capability Debate. A growing body of theoretical work challenges the source of RLVR's gains. In language-only domains, recent analyses argue that RLVR primarily functions as an efficiency mechanism, optimizing the selection of correct reasoning trajectories that already exist within the pre-trained model's latent support, rather than synthesizing novel capabilities (Ye et al., 2025; Yue et al., 2025; Xu et al., 2025). Specifically, studies show that on math benchmarks, base models often achieve RLVR-level performance given sufficient sampling budgets, implying strong prior coverage (Liu et al., 2025b; AI et al., 2025). This phenomenon is likely reinforced by the fact that such domains are extensively represented in pre-training corpora and may be subject to benchmark contamination or memorization effects, further contributing to their high-support nature (Matton et al., 2024; Liang et al., 2025).

Our work serves as a critical counterpoint to this efficiency hypothesis. We investigate RLVR in a spatial reasoning setting where prior coverage is not merely low but effectively absent. By demonstrating that RLVR enables success in these regimes where the base model fails regardless of sampling budget, we provide empirical evidence that RLVR can indeed extend the reasoning boundary, distinguishing capability expansion from mere sampling efficiency.

3 Method

Our methodology is designed to examine the hypothesis that RLVR can extend the reasoning boundaries of VLMs beyond the effective reach of the base policy. To systematically investigate this capability, we introduce **Ariadne**, a minimal yet fully controllable post-training framework. Ariadne enables precise characterization of the reasoning boundary by generating synthetic maze navigation tasks where the ground-truth difficulty can be explicitly regulated, allowing us to identify regimes where the base model consistently fails to produce valid solutions.

3.1 Ariadne: A Probe for Reasoning Boundaries

3.1.1 Verifiable Maze Environment

We construct a controllable testbed based on grid-based maze navigation, adapted from AlphaMaze (Dao and Vu, 2025). Unlike natural image VQA benchmarks, where the difficulty level or reasoning boundary is hard to explicitly define, Ariadne provides a deterministic environment where the validity of a reasoning step is binary and verifiable. We generate maze images as visual inputs (see Figure 1), creating a pure spatial reasoning task where the VLM is queried to generate the action sequence in textual format (see Appendix A for prompt details). As shown in Figure 1, valid solutions require generating a sequence of discrete actions (`up`, `down`, `left`, `right`) that respect wall constraints.

This design grants us three critical analytical capabilities: **(1) Deterministic Verification:** We can explicitly verify every step of a generated chain, eliminating the need for an external judge model that might introduce bias or hallucination. **(2) Step-wise Feasibility:** We can distinguish between "valid but incorrect" paths (wrong turn) and "hallucinated" paths (walking through walls), allowing for a precise taxonomy of reasoning failures. **(3) Parametric Complexity:** By systematically varying path length (L) and number of turns (T), we can map the *exact* boundary where the base model's pass@k probability collapses to zero.

3.1.2 Curriculum for Boundary Expansion

A critical challenge in extending reasoning boundaries is ensuring the model encounters a sufficient density of successful trajectories at the frontier of its capability. Standard uniform sampling over maze difficulties is inefficient: it allocates excessive capacity to trivial instances where the policy is already competent, while frequently sampling complex instances where the model consistently fails to generate valid solutions.

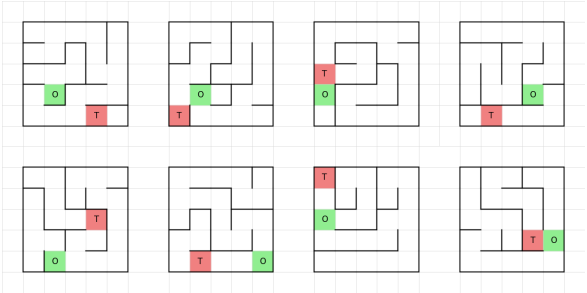

Figure 1: Representative maze instances from Ariadne. Green and red cells denote the start and goal, respectively, and black lines indicate walls. Mazes vary in path length and number of turns, enabling controlled spatial reasoning difficulty.

To bridge the gap between the model's current capability and the target difficulty zone, we design a difficulty-aware curriculum. We control difficulty via optimal path length $s \in \{1, \ldots, 5\}$ and sample instances according to an inverted Gaussian-like distribution:

$$P(s) \propto 1 - \exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right), \qquad (1)$$

where $\mu = 3$ and $\sigma = 2$. This distribution explicitly under-samples the "comfort zone" (medium difficulty) and over-samples the edges: **(1) High-Probability Anchors** ($s = 1, 2$) ensure the model retains basic valid movement primitives; and **(2) Boundary Frontiers** ($s = 4, 5$) force the model to attempt tasks at the very edge of its executable horizon. Crucially, we train exclusively on short-horizon tasks ($L \leq 5$, Turns ≤ 2) while evaluating on **unseen, out-of-distribution (OOD)** instances with significantly extended horizons ($L \leq 10$, Turns ≤ 4). This regime ensures that any observed success reflects genuine generalization to novel complexity levels rather than mere pattern memorization.

3.2 Policy Optimization via GRPO

We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for VLM post-training. Formally, for a query q, we sample a group of outputs $\{o_1, \ldots, o_G\}$ from the reference policy $\pi_{\theta_{\text{old}}}$. The advantage A_i is derived via group normaliza-

tion of the rewards:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}. \quad (2)$$

The policy π_θ is then updated to maximize the surrogate objective:

$$\mathcal{J}(\theta) = \mathbb{E}[\frac{1}{G}\sum_{i=1}^{G} \min(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i)], \quad (3)$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the importance sampling ratio. This mechanism incentivizes the model to consistently outperform its average behavior, effectively shifting the probability mass toward higher-reward trajectories.

3.3 Dense Verifiable Reward Design

In regimes where the base model lacks support for the correct solution, binary success rewards are insufficient because the model may never reach the goal during early training. To provide a learning gradient, we design a dense, prefix-matching reward function (Algorithm 1).

Let A be the ground-truth action sequence and R be the predicted sequence. We define the reward r as:

$$r = \begin{cases} \alpha_1 \cdot |A| \cdot \psi(A), & \text{if } R = A \text{ (Success)} \\ \alpha_2 \cdot k \cdot \psi(A_{1:k}), & \text{if } R \neq A \text{ (Partial)}, \end{cases} \quad (4)$$

where α_1 and α_2 are weighting factors, k is the length of the longest matching prefix ($A_{1:k} = R_{1:k}$), and $\psi(\cdot)$ is a complexity multiplier based on the number of turns. This function serves two purposes: **(1) Guidance Signal:** It rewards the model for every correct step taken, creating a dense supervision signal that extends the reasoning horizon even if the final goal is missed. **(2) Complexity Awareness:** The term $\psi(A)$ scales the reward with structural difficulty (turns), preventing the model from collapsing into simple straight-line heuristics. This dense signal is the mechanism that allows the policy to climb out of the valley of zero success.

Algorithm 1 Correctness Reward Function

```
rewards ← []
for all (r, a) in (completions, answers) do
    r_m ← moves(r);    a_m ← moves(a)
    if r_m = a_m then
        reward ← |a_m| × α_1 × turns(a_m)
    else
        k ← prefix_len(r_m, a_m)
        reward ← k × α_2 × turns(a_m[1:k])
    end if
    rewards.append(reward)
end for
return rewards
```

4 Experiments

4.1 Experimental Setup

We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) as the primary backbone, and additionally evaluate Qwen2.5-VL-3B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct (Bai et al., 2025a) to assess generality. Under the Ariadne framework, the model is post-trained with RLVR on 4,700 AlphaMaze samples (Dao and Vu, 2025) using the GRPO algorithm. We additionally train a supervised fine-tuned variant (SFT) on the same data and initialization, with identical training settings but without reinforcement learning.

The reward combines answer correctness (Algorithm 1), answer format, and reasoning format, with emphasis on correctness (α_1=0.2, α_2=0.1). Training uses 8 NVIDIA A100 (40GB) GPUs with a learning rate of 1×10^{-6} for one epoch, batch size 1, 16-step gradient accumulation (722,000 steps), and a warmup ratio of 0.05.

For each training query, we sample a group of $G = 8$ candidate responses with a temperature of 1.0. The GRPO clipping parameter is set to $\epsilon = 0.2$. During evaluation across all three benchmarks, each prompt is evaluated using independent rollouts at a temperature of 1.0, and reported results are averaged across these runs.

4.2 Benchmarks and Metrics

We evaluate capability extension across two distinct regimes: (1) internal boundary extension within the controlled Ariadne framework, and (2) zero-shot transfer to external real-world benchmarks.

Controlled Evaluation (Ariadne). To rigorously map the reasoning boundary, we utilize a held-out test set generated via Ariadne (adapted from AlphaMaze (Dao and Vu, 2025)). Crucially, this set is stratified by difficulty, including "training-distribution" instances (Length $\leq$ 5) and "out-of-distribution" (OOD) instances (Length 6–10, Turns $\leq$ 4) that were strictly excluded during post-training. The primary metric is **Success Rate**

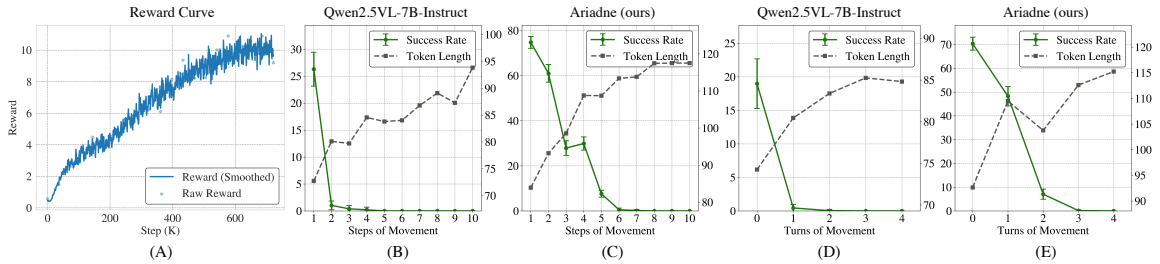

Figure 2: Training dynamics and reasoning boundary shift on AlphaMaze. **(A)** GRPO training demonstrates stable reward convergence. **(B, D)** The base VLM exhibits a sharp "capability collapse" at low complexity (exceeds 2 steps or 1 turn), failing to generate valid paths regardless of token length. **(C, E)** Ariadne triples the effective reasoning horizon (shifting the collapse point from 2 to 6 steps) and recovers non-trivial success rates in regimes where the base model has effectively zero support.

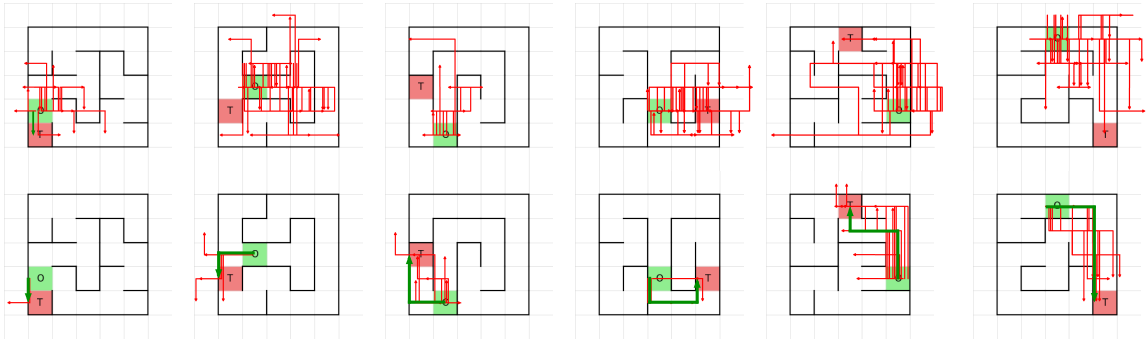

Figure 3: Qualitative comparison of search dynamics ($T = 1.0, N = 128$). **Top:** The base VLM suffers from exploration paralysis, generating diverse but invalid trajectories (red) that fail to ground in the maze structure. **Bottom:** Ariadne produces successful trajectories (green) that the base VLM fails to generate. This visualizes the transition from unstructured failure to structured spatial reasoning.

(SR), defined as the percentage of episodes where the generated action sequence exactly matches the ground-truth path.

Real-World Transfer. We evaluate generalization on two established benchmarks (see examples in Figure 8 and 9) that require similar spatial reasoning while differing substantially in visual domains:

- **MapBench** (Xing et al., 2025) evaluates navigation on realistic street maps. We report the **Shortest Path (SP) Score**, which measures the optimality of the generated path relative to the ground truth (where 1.0 indicates an optimal path).

- **ReasonMap** (Feng et al., 2025) assesses topological reasoning on complex transit networks. We utilize the weighted **Map Score**, a metric that aggregates correctness across varying difficulty levels (short vs. long-horizon queries) to reflect robust spatial planning capabilities.

4.3 RLVR Extends the Effective Reasoning Boundary

We first investigate whether RLVR genuinely expands the model's spatial reasoning capabilities or simply amplifies behaviors already present in the base policy. Figure 2 illustrates the training dynamics using Qwen2.5-VL-7B-Instruct as the base VLM, along with comparative performance on the AlphaMaze test set.

Stable Optimization and Boundary Shift. As shown in Figure 2 (A), GRPO training yields a steady ascent in reward, indicating stable policy optimization. The impact is quantified in Panels (B–E): the pretrained base VLM exhibits a distinct **reasoning boundary**, where the probability of valid navigation collapses to near zero ($< 1\%$) once task complexity exceeds 2 steps or 1 turn. This quantitative collapse is mirrored by the qualitative behavior shown in Figure 3. Under identical sampling conditions, the base model (top) generates diverse but unsuccessful trajectories, whereas Ariadne (bottom) produces structured, successful

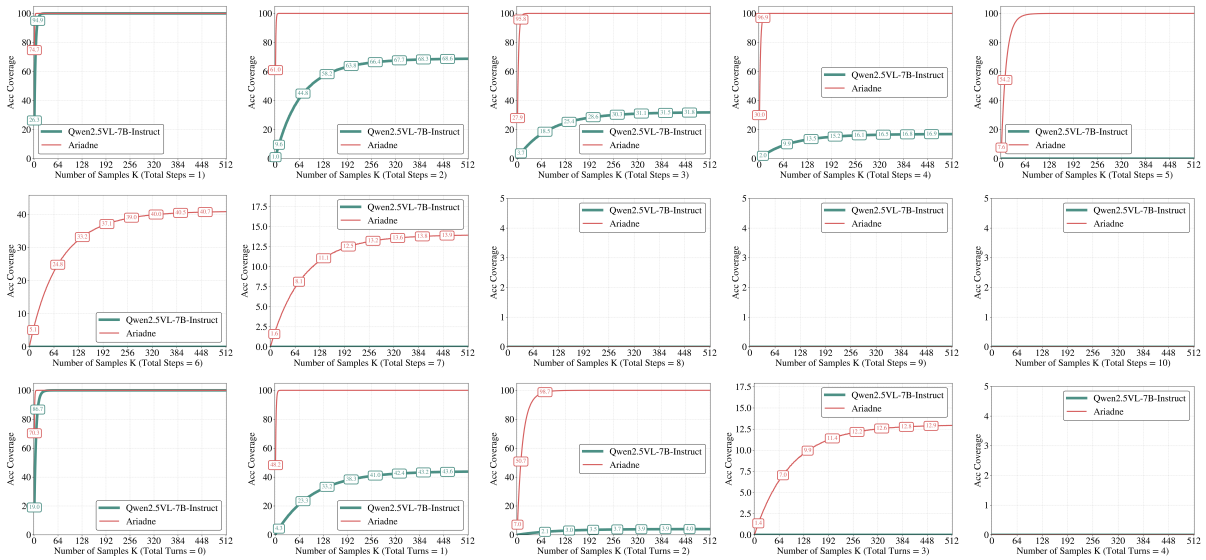

Figure 4: Pass@k coverage analysis ($k \leq 512$). **Base Model Saturation:** For complexities ≥ 5 steps or ≥ 2 turns, the base VLM saturates near 0%, indicating a total absence of valid solutions in its latent space regardless of sampling budget. **Ariadne Generalization:** Ariadne exhibits logarithmic scaling on unseen complexities (Steps 6–7, Turn 3), confirming robust generalization beyond the training data (Steps ≤ 5, Turns ≤ 2). In the far-OOD regime (Steps ≥ 8, Turn 4), the performance naturally converges to the intrinsic complexity limit of the learned representations.[2]

plans. Post-training, Ariadne fundamentally shifts this boundary, tripling the effective reasoning horizon (from 2 to ~ 6 steps) and recovering $\sim 10\%$ success on 2-turn mazes where the base model consistently fails.

RLVR Breaks the Base VLM Reasoning Boundary. A critical question regarding RLVR is whether it induces novel reasoning paths or simply improves the sampling efficiency of existing ones. Figure 4 provides decisive evidence against the pure efficiency hypothesis (Yue et al., 2025). We observe a critical distinction between "soft" and "hard" complexity regimes. In the "soft" regime (Steps < 5), the base VLM's coverage rises with sampling budget, indicating that solutions exist but are rare. However, this dynamic disappears in the "hard" regime:

- **Base Model Collapse:** For tasks beyond the initial boundary (Steps ≥ 5, Turns ≥ 2), the base VLM's coverage curve flat-lines at exactly 0% even as $k \to 512$. This "hard zero" indicates that valid trajectories are effectively absent from the model's search space, meaning no amount of sampling efficiency optimization could recover the solution.
- **OOD Generalization (Ariadne):** In contrast, Ariadne exhibits robust scaling well outside its training distribution. Despite being trained only on short horizons (Steps ≤ 5, Turns ≤ 2), it surprisingly achieves healthy coverage growth on unseen complexities like Steps 6–7 and Turn 3.

This establishes that in "hard" regimes, Ariadne is not merely optimizing efficiency, but **inducing novel behaviors** that bridge the gap between the absence of valid solutions and successful execution. Finally, we note that this capability extension is finite; in extreme out-of-distribution settings (e.g., Steps ≥ 8, Turns ≥ 4), the induced policy approaches its natural complexity horizon, reflecting the fundamental difficulty gap between the training support and far-OOD reasoning.

To further eliminate the possibility that the observed "hard zero" phenomenon arises from insufficient sampling, we extend the evaluation budget to pass@1024 in Table 1, where the results remain consistent with the trends in Figure 4.

4.4 Zero-Shot Transfer to Real-World Tasks

We next examine whether the reasoning behaviors learned in the controlled Ariadne setting transfer to real-world navigation and spatial reasoning tasks. We evaluate three VLMs in a strictly zero-shot

[2]Plots for Steps ≥ 8 and Turn 4 appear blank because both the base model and Ariadne achieve a 0% success rate, reflecting the extreme out-of-distribution difficulty.

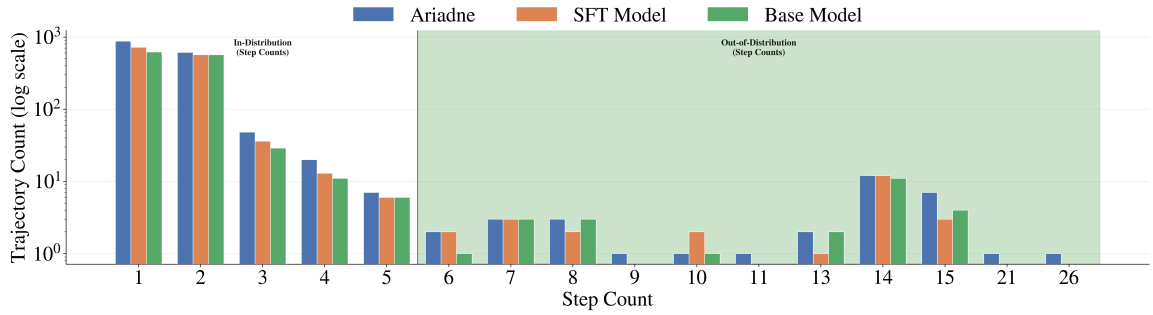

Figure 5: Mechanism of transfer on MapBench. **Left (In-Distribution):** Ariadne significantly amplifies the density of successful trajectories within the 1–5 step range (matching the Ariadne curriculum). **Right (Out-of-Distribution):** This amplified density "spills over" into the OOD regime (6+ steps). While the transfer is not a rigid boundary copy, all models achieve some success on longer real-world paths. Ariadne uniquely maintains valid solutions at extreme lengths (e.g., 21, 26 steps) where the base and SFT VLM vanish. Steps with zero success for both models are excluded from visualization.

Table 1: Pass@k coverage across step complexity under extended sampling budgets ($k \leq 1024$). Rows denote step counts, and columns report coverage for the base model (Qwen2.5-VL-7B-Instruct) and its Ariadne (RLVR-trained variant) at increasing k, including pass@512 and pass@1024.

N	pass@1		pass@4		pass@8		pass@16		pass@64		pass@128		pass@512		pass@1024	
	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne
1	26.3	74.8	70.2	99.6	90.9	100.0	99.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	1.1	61.1	4.0	97.5	7.8	99.9	14.5	100.0	41.2	100.0	56.8	100.0	68.8	100.0	69.0	100.0
3	0.4	27.9	1.5	72.6	3.0	92.3	5.7	99.3	16.9	100.0	24.7	100.0	31.9	100.0	32.0	100.0
4	0.2	29.9	0.8	75.6	1.6	93.8	3.1	99.6	9.0	100.0	13.1	100.0	16.9	100.0	17.0	100.0
5	0.0	7.6	0.0	26.9	0.0	46.5	0.0	70.9	0.0	98.9	0.0	100.0	0.0	100.0	0.0	100.0
6	0.0	0.5	0.0	2.1	0.0	4.1	0.0	7.8	0.0	22.7	0.0	32.3	0.0	40.8	0.0	41.0
7	0.0	0.2	0.0	0.7	0.0	1.3	0.0	2.4	0.0	7.4	0.0	10.7	0.0	13.9	0.0	14.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

setting based on Qwen2.5-VL-7B-Instruct: the pre-trained Base Model, a supervised fine-tuned variant (SFT Model) trained on the same AlphaMaze data, and the RLVR-trained model (Ariadne). Neither SFT nor Ariadne is exposed to MapBench or ReasonMap during training.

Table 2: Zero-shot performance comparison on MapBench (Score ↓, lower is better). **Bold** indicates the best performance.

Metric	Base VLM	SFT VLM	Ariadne
Google Map ↓	1.61	1.80	**1.30**
Mall ↓	1.43	1.44	**1.41**
Museum ↓	1.43	1.39	**1.33**
National Park ↓	1.86	1.85	**1.48**
Theme Park ↓	1.78	1.66	**1.46**
Trail ↓	2.29	1.90	**1.47**
Campus ↓	1.62	1.55	**1.29**
Urban ↓	1.93	1.94	**1.91**
Zoo ↓	1.68	1.64	**1.32**

Transfer of Route Optimality (MapBench). MapBench evaluates navigation efficiency via the *Shortest Path (SP) Score* (lower is better, $1.0 =$ optimal). As detailed in Table 2, Ariadne achieves broad improvements over the base and SFT model. While SFT yields modest improvements in some structured settings (e.g., Campus and Zoo), it fails to match Ariadne's gains in more challenging or unstructured environments. Notably, the largest improvements from Ariadne occur in terrains such as "Trail" and "National Park", which exhibit sparse landmarks and irregular connectivity. These environments more closely resemble the abstract connectivity patterns of maze navigation than grid-like "Urban" layouts, where all models perform similarly. This confirms that RLVR optimized the policy for generic search strategies rather than specific visual pattern matching.

To analyze the mechanism behind real-world transfer, Figure 5 decomposes success distributions

by path length. Within the **1–5 step in-distribution range** (Left), Ariadne substantially outperforms both the base model and SFT, closely aligning with the curriculum support of maze-based training. Beyond this range, we observe a partial generalization into the **6+ step out-of-distribution regime** (Right), indicating a spillover of learned behaviors.

Unlike the sharp reasoning boundary observed in the synthetic maze domain, real-world navigation exhibits a softer boundary: all models achieve sporadic success on longer paths (> 10 steps). Nevertheless, Ariadne consistently dominates the long-tail region, uniquely producing valid trajectories for extreme cases (e.g., 21 and 26 steps) where the base and SFT VLM fail entirely. RLVR preserves the VLM's *capacity* for coherent long-horizon reasoning in more complex settings. In contrast, SFT yields only limited improvements in this regime, suggesting that supervised fine-tuning alone is insufficient to induce robust search and path-planning behavior. Together, these results indicate that RLVR promotes more effective exploration and long-horizon coordination in spatial reasoning tasks beyond what can be acquired through imitation.

Scaling Across Model Sizes on MapBench. To verify that the observed gains generalize beyond the 7B backbone, we evaluate additional model families and scales, including Qwen2.5-VL-3B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct on MapBench. As shown in Table 3, Ariadne consistently improves route optimality across all environments and model sizes.

Table 3: Zero-shot performance on MapBench across model scales (Score ↓). All models are instruction-tuned variants. Lower is better. **Bold** indicates the better result for each pair.

Metric	Qwen2.5-VL-3B		Qwen3-VL-4B		Qwen3-VL-8B	
	Base	Ariadne	Base	Ariadne	Base	Ariadne
Google Map	2.70	**2.11**	2.33	**2.04**	1.32	**1.25**
Mall	2.99	**2.21**	2.01	**1.65**	1.22	**1.01**
Museum	2.63	**2.37**	1.93	**1.72**	1.24	**1.13**
National Park	2.24	**2.19**	2.34	**1.96**	1.58	**1.29**
Theme Park	2.61	**2.24**	2.11	**1.94**	1.42	**1.16**
Trail	2.58	**2.38**	2.45	**2.32**	2.08	**1.52**
Campus	2.57	**2.07**	2.13	**1.86**	1.64	**1.37**
Urban	3.12	**2.89**	2.39	**2.01**	1.98	**1.79**
Zoo	2.77	**2.09**	2.52	**2.19**	1.77	**1.39**

The improvements are particularly pronounced for smaller models (e.g., 3B and 4B), where the base models exhibit weaker navigation performance. This suggests that RLVR is especially effective in low-capacity regimes, where it compensates for the lack of implicit search structure. As model scale increases, the base models become stronger, but Ariadne continues to provide consistent gains across nearly all environments. Notably, the largest improvements persist in structurally complex settings such as "Trail" and "National Park", reinforcing that RLVR promotes generalizable search strategies rather than environment-specific heuristics.

Transfer of Reasoning Depth (ReasonMap). ReasonMap assesses high-level planning on schematic transit maps. Using Qwen2.5-VL-7B-Instruct as the backbone, Table 4 highlights a critical behavioral shift under RLVR: Ariadne not only improves accuracy for long questions but also substantially increases the volume of explicit reasoning. For Long Questions (L), the average token count doubles from 61 to 121, accompanied by a 1.47% absolute gain in weighted accuracy and a significant improvement in Map Score ($4.51 \rightarrow 5.15$). In contrast, SFT fails to produce similar gains in the long-question regime, yielding shorter responses and degraded performance. These results suggest that RLVR promotes explicit, step-by-step reasoning required for complex long-horizon planning, rather than merely increasing selection accuracy.

Table 4: Performance comparison on ReasonMap. S indicates short questions, L indicates long questions. **Bold** indicates the best performance.

Metric	Base VLM	SFT VLM	Ariadne
Weighted Acc. (S) ↑	13.32%	**15.44**%	14.50%
#Tokens (S)	26	25	43
Weighted Acc. (L) ↑	6.00%	4.10%	**7.47**%
#Tokens (L)	61	50	121
Weighted Map Score (S) ↑	3.73	**3.79**	3.67
Weighted Map Score (L) ↑	4.51	3.71	**5.15**

Figure 6 provides a granular breakdown of this performance scaling. As shown in Panel (a), Ariadne exhibits its largest gains on simpler instances ("Easy/Easy"), where it substantially outperforms both the base model and SFT for Short questions. Crucially, this advantage persists as task complexity increases. While absolute performance naturally declines with rising map difficulty (Panels B–D), Ariadne consistently maintains higher Map Scores than the base model across all difficulty tiers. Notably, even in the hard setting (Panel A, "Easy/Hard"), Ariadne retains meaningful accuracy ($\sim 21\%$), whereas both the base model ($\sim 12\%$)

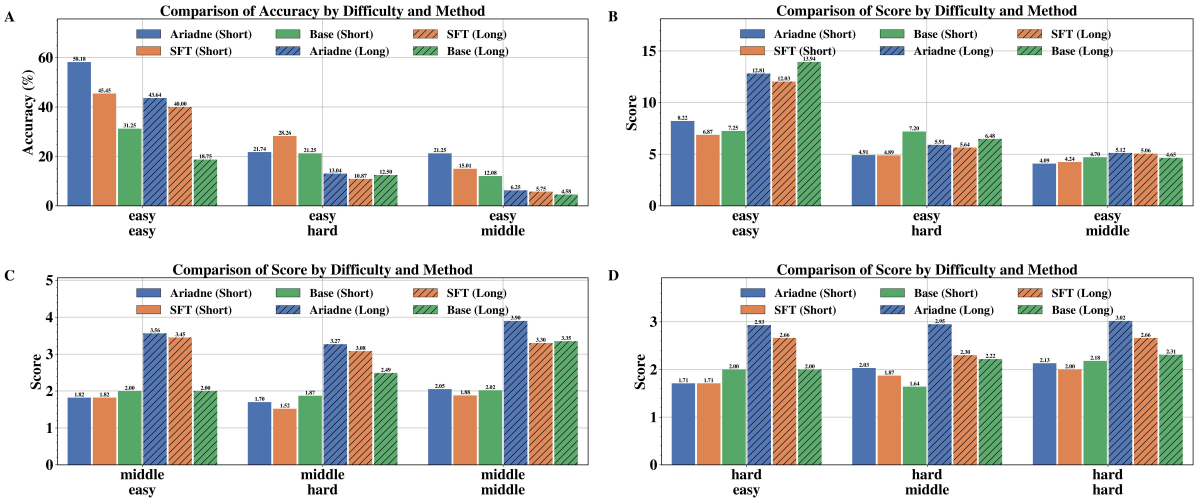

Figure 6: Performance consistency on ReasonMap. **(A)** Ariadne maintains a persistent accuracy lead over the base and SFT VLM. **(B–D)** Map Scores under increasing difficulty. Note that while absolute performance naturally degrades with complexity, Ariadne's advantage is robust, particularly in the long questions in the "Medium" difficulty band, mirroring the effectiveness of the difficulty-aware curriculum.

and SFT degrade more sharply. These results indicate that the spatial planning behaviors learned through RLVR transfer to real-world settings and support more robust long-horizon reasoning under increased difficulty.

Scaling of Reasoning Behaviors on ReasonMap. We further examine whether the reasoning improvements induced by RLVR persist across model scales beyond the 7B backbone. Table 5 shows consistent gains in both accuracy and reasoning depth across Qwen2.5-VL-3B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct.

Table 5: Performance on ReasonMap across model scales. All models are instruction-tuned variants. S: short questions, L: long questions. **Bold** indicates the better result.

Metric	Qwen2.5-VL-3B		Qwen3-VL-4B		Qwen3-VL-8B	
	Base	Ariadne	Base	Ariadne	Base	Ariadne
Weighted Acc. (S) ↑	7.92%	**8.54%**	8.05%	**8.68%**	13.88%	**14.96%**
#Tokens (S)	31	52	24	38	29	47
Weighted Acc. (L) ↑	3.48%	**4.32%**	3.62%	**4.54%**	6.24%	**7.73%**
#Tokens (L)	69	134	58	109	64	128
Weighted Map Score (S) ↑	2.19	**2.27**	2.24	**2.31**	3.82	**3.89**
Weighted Map Score (L) ↑	2.74	**3.11**	2.81	**3.20**	4.68	**5.33**

Across all models, Ariadne achieves consistent improvements, particularly in the long-question regime where planning depth is critical. Performance gains are most pronounced in this setting, and while larger models achieve higher absolute performance, the relative improvements from RLVR remain stable, suggesting that the induced behaviors are not tied to a specific capacity regime but generalize across model scales.

5 Conclusion

In this work, we investigate the prevailing view that RLVR merely amplifies existing behaviors, providing decisive evidence that it functions as a mechanism for **inducing novel behaviors** in vision-language domains. By isolating the effective reasoning boundary, where the base model's valid solution space is effectively empty regardless of sampling budget, we demonstrate that RLVR constructs novel spatial primitives that enable success in regimes characterized by the absence of valid solutions. Crucially, these learned behaviors are not overfitting artifacts but generalized logic, evidenced by their robust zero-shot transfer to the visually distinct and semantically rich environments of MapBench and ReasonMap, as well as their consistent effectiveness across model scales.

Ultimately, our findings suggest that while language-centric tasks may benefit significantly from efficiency optimization, visual-spatial reasoning requires the fundamental boundary expansion that verifiable reinforcement learning is particularly well-suited to provide. Furthermore, while RLVR extends this reasoning horizon, we observe that the precise boundary established in synthetic environments does not directly transfer to real-world tasks, where success depends on a more complex interplay of visual semantics and logical depth.

Limitations

While our results offer strong evidence for reasoning boundary extension, several limitations remain. First, our primary analysis relies on the Ariadne maze framework as a proxy for spatial intelligence. While this enables precise boundary verification, future research should incorporate a broader range of real-world scenarios with similarly rigorous definitions of difficulty levels to fully validate the transferability of these boundaries.

Second, due to computational resource constraints, our experiments primarily focus on smaller-scale VLMs from the Qwen2.5-VL and Qwen3-VL series. We leave it to future work to extend this analysis to a broader range of model variants and larger scales, in order to assess whether frontier models possess sufficient capacity to render these boundaries permeable.

Finally, while we demonstrate effective zero-shot transfer, the induced policy eventually encounters its own complexity horizon in extreme out-of-distribution settings. Moreover, although the weights of many open VLMs are accessible, their pre-training data distributions remain opaque, making it difficult to definitively map the absolute reasoning boundary of existing models beyond empirical probing.

References

Essential AI, Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Anthony Polloreno, Ashish Tanwer, Burhan Drak Sibai, Divya S Mansingka, Divya Shivaprasad, Ishaan Shah, Karl Stratos, and 9 others. 2025. Rethinking reflection in pre-training. *arXiv preprint arXiv:2504.04022*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report.

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. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wenhao Deng, Long Wei, Chenglei Yu, and Tailin Wu. 2025. Unlocking reasoning capabilities in llms via reinforcement learning exploration. *arXiv preprint arXiv:2510.03865*.

Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. 2025. Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. 2025. Do vision-language models really understand visual language? In *Proceedings of International Conference on Machine Learning*.

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

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Proceedings of Conference on Language Modeling*.

Shanchao Liang, Spandan Garg, and Roshanak Zilouchian Moghaddam. 2025. The swe-bench illusion: When state-of-the-art llms remember instead of reason. *arXiv preprint arXiv:2506.12286*.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025a. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. In *Proceedings of Conference on Language Modeling*.

Alexandre Matton, Tom Sherborne, Dennis Aumiller, Elena Tommasone, Milad Alizadeh, Jingyi He, Raymond Ma, Maxime Voisin, Ellen Gilsenan-McMahon, and Matthias Gallé. 2024. On leakage of code generation evaluation datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13215–13223.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *Proceedings of The International Conference on Learning Representations*.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the International Conference on Neural Information Processing Systems*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, Chaofan Tao, Yangfan He, Mi Zhang, and Shen Yan. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. 2025b. Emergent hierarchical reasoning in llms through reinforcement learning. *arXiv preprint arXiv:2509.03646*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. 2026. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 33944–33952.

Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. 2025. Can large vision language models read maps like a human? *arXiv preprint arXiv:2503.14607*.

Sen Xu, Yi Zhou, Wei Wang, Jixin Min, Zhibin Yin, Yingwei Dai, Shixi Liu, Lianyu Pang, Yirong Chen, and Junlin Zhang. 2025. Tiny model, big logic: Diversity-driven optimization elicits large-model reasoning ability in vibethinker-1.5b.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Guanghao Ye, Khiem Duc Pham, Xinzhi Zhang, Sivakanth Gopi, Baolin Peng, Beibin Li, Janardhan Kulkarni, and Huseyin A Inan. 2025. On the emergence of thinking in llms i: Searching for the right intuition. *arXiv preprint arXiv:2502.06773*.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, and 3 others. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Proceedings of the International Conference on Neural Information Processing Systems*.

Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From System 1 to System 2: A Survey

of Reasoning Large Language Models . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(01):1–20.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*.

A Appendix

A.1 LLM Usage Statement

Large language models (LLMs), such as ChatGPT, were used as general-purpose assistive tools during the preparation of this paper. Specifically, LLMs were employed for language refinement and improving the clarity of the manuscript. No part of the research ideation, experimental design, or core scientific contributions relied on LLMs. All scientific content, results, and conclusions were generated and verified by the authors. The authors take full responsibility for the content of this paper, including any text generated with the assistance of LLMs.

A.2 System Prompt

The following prompt defines a navigation assistant designed for visual path-finding in AlphaMaze. Given a maze image with a green starting cell ("O") and a red target cell ("T"), the assistant must infer a valid path that passes exclusively through open cells while avoiding black walls.

System Prompt for AlphaMaze

You are a navigation assistant to solve visual path-finding tasks.
Your goal is to infer a valid path from a visually marked starting point (green cell labeled 'O') to a visually marked target (red cell labeled 'T') by analyzing the maze image.
Rules:

- The maze is composed of open paths and impassable black walls.
- Movement is only allowed through open paths, not through walls.
- You can move one step at a time in the four cardinal directions: <|up|>, <|down|>, <|left|>, <|right|>.

Output Format:
Think through each step inside <think> and </think> tags.
At each step:

1. Describe your current position based on visual layout and structure (e.g., "in a corridor", "facing a wall", "at a crossroad", "turning a corner").
2. Decide the next move, and explain your reasoning.
3. Move and continue the path.

After your full reasoning, output only the final movement sequence using the allowed tokens:

<|up|><|down|><|left|><|right|>

A.3 Additional Experimental Results

To provide a more detailed view of pass@k scaling, we report finer-grained evaluations with denser sampling strides, including a tabulated version in Table 6 (focused on low-to-mid budgets) and a continuous visualization in Figure 7 (stride = 10 up to pass@100). Compared to the main text, which uses sparsely spaced evaluation budgets, these results resolve intermediate scaling behavior more precisely.

Across both the table and the figure, the observed trends remain consistent under finer resolution. In particular, the transition from low to high coverage unfolds smoothly as k increases, and the relative behavior between the base model and Ariadne is preserved across all intermediate budgets. This consistency indicates that the scaling patterns reported in the main text are not artifacts of specific evaluation points, but reflect stable properties of the underlying policies.

As shown in Figure 7, the observed trends remain consistent under finer resolution. In particular, the transition from low to high coverage unfolds smoothly as k increases, and the relative behavior between the base model and Ariadne is preserved across all intermediate budgets. This suggests that the scaling patterns reported in the main text are not artifacts of specific evaluation points, but reflect stable properties of the underlying policies.

A.4 Examples from Spatial Reasoning Benchmarks

We evaluate the generalization of the learned spatial reasoning behaviors on external benchmarks that require navigation and route planning under diverse visual layouts. These benchmarks differ significantly from the synthetic maze environment in both structure and appearance, providing a realistic testbed for assessing whether the learned spatial reasoning strategies transfer beyond the training distribution.

Specifically, we consider two representative benchmarks: 1) MapBench, which focuses on instruction-following navigation on street maps, and 2) ReasonMap, which evaluates fine-grained route planning on transit schematics. Figure 8 and Figure 9 illustrate representative examples from these datasets.

Table 6: Fine-grained pass@k coverage across step complexity ($k \leq 40$). Rows denote step counts, and columns report coverage for the base model and Ariadne at different sampling budgets.

N	pass@1		pass@5		pass@15		pass@20		pass@25		pass@30		pass@35		pass@40	
	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne	Base	Ariadne
1	26.3	74.8	77.8	99.9	98.8	100.0	99.7	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	1.1	61.1	5.0	99.0	13.8	100.0	17.7	100.0	21.3	100.0	24.6	100.0	27.6	100.0	30.4	100.0
3	0.4	27.9	1.9	80.2	5.4	99.0	6.9	99.8	8.4	99.9	9.7	100.0	10.9	100.0	12.1	100.0
4	0.2	29.9	1.0	82.7	2.9	99.4	3.7	99.9	4.5	100.0	5.2	100.0	5.9	100.0	6.5	100.0
5	0.0	7.6	0.0	32.6	0.0	68.6	0.0	78.4	0.0	85.0	0.0	89.4	0.0	92.6	0.0	94.8
6	0.0	0.5	0.0	2.6	0.0	7.3	0.0	9.5	0.0	11.4	0.0	13.2	0.0	14.9	0.0	16.4
7	0.0	0.2	0.0	0.8	0.0	2.3	0.0	3.0	0.0	3.6	0.0	4.2	0.0	4.8	0.0	5.3
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

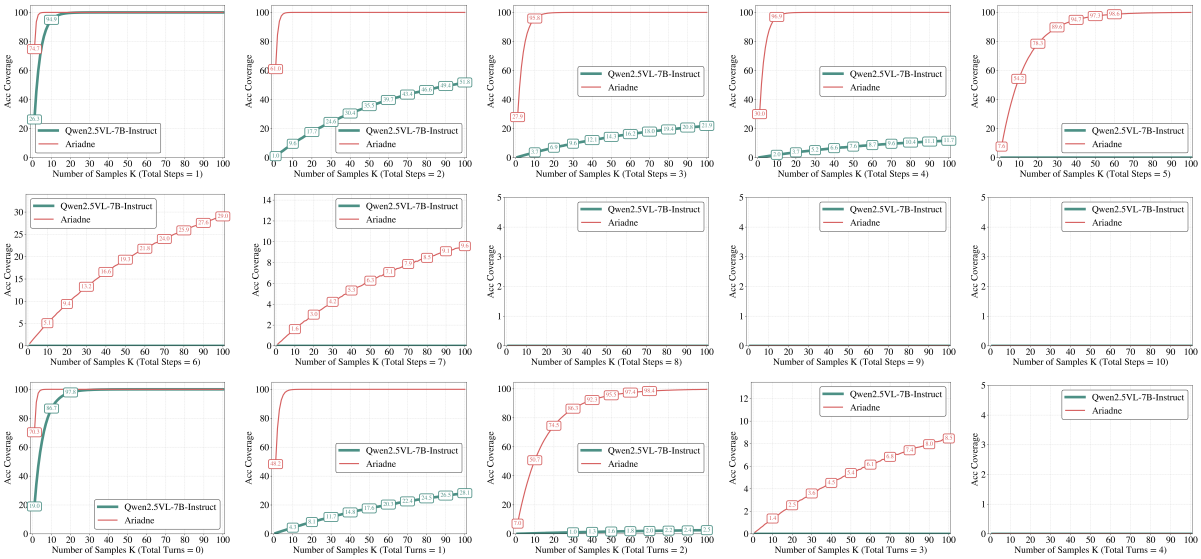

Figure 7: Fine-grained pass@k coverage analysis across step and turn complexity ($k \leq 100$, stride = 10). Coverage is shown as a function of sampling budget for both the base model and Ariadne.

Figure 8: Representative examples from **MapBench**, which evaluates instruction-following navigation on street maps.

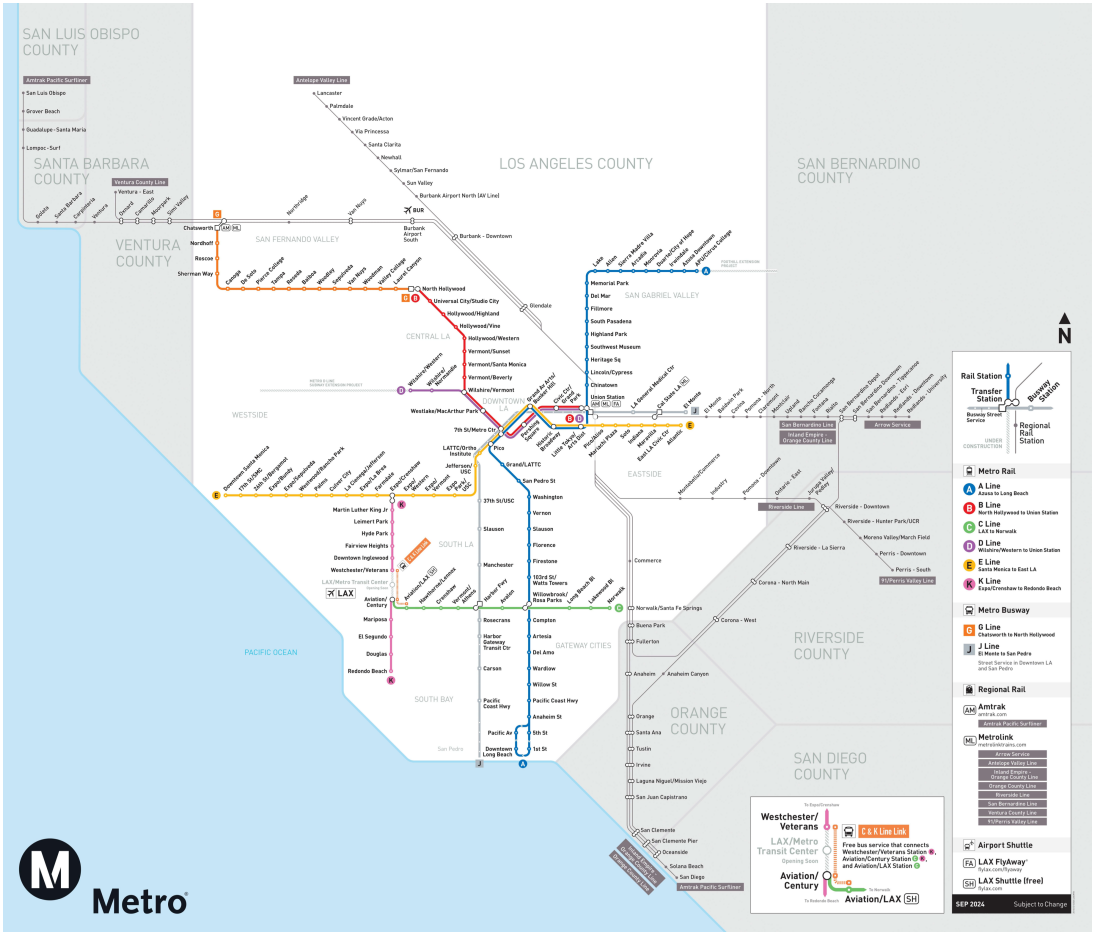

Figure 9: Representative examples from **ReasonMap**, which assesses fine-grained route planning on transit schematics.