

Visual Planning

Let's Think Only with Images

ICLR 2026 Oral

Yi Xu^{1*}, Chengzu Li^{1*}, Han Zhou^{1*}, Xingchen Wan², Caiqi Zhang¹, Anna Korhonen¹, Ivan Vulić¹

¹ Language Technology Lab, University of Cambridge ² Google



ICLR
International Conference On
Learning Representations

1

Motivation

Why think in images? The modality gap problem

2

Visual Planning Paradigm

A new reasoning paradigm without text

3

VPRL Framework

Two-stage RL for visual planning

4

Experiments & Results

FrozenLake, Maze, MiniBehavior

5

Analysis & Discussion

Why RL helps, qualitative examples, OOD

Motivation: Why Think in Images?

Key Observation: Current MLLMs process visual inputs but perform ALL reasoning in text.

This creates a modality gap that hinders spatial, geometric, and physical reasoning.

Spatial Reasoning

Route planning, maze navigation,
room layout — naturally visual,
hard to verbalize

Grounding Errors

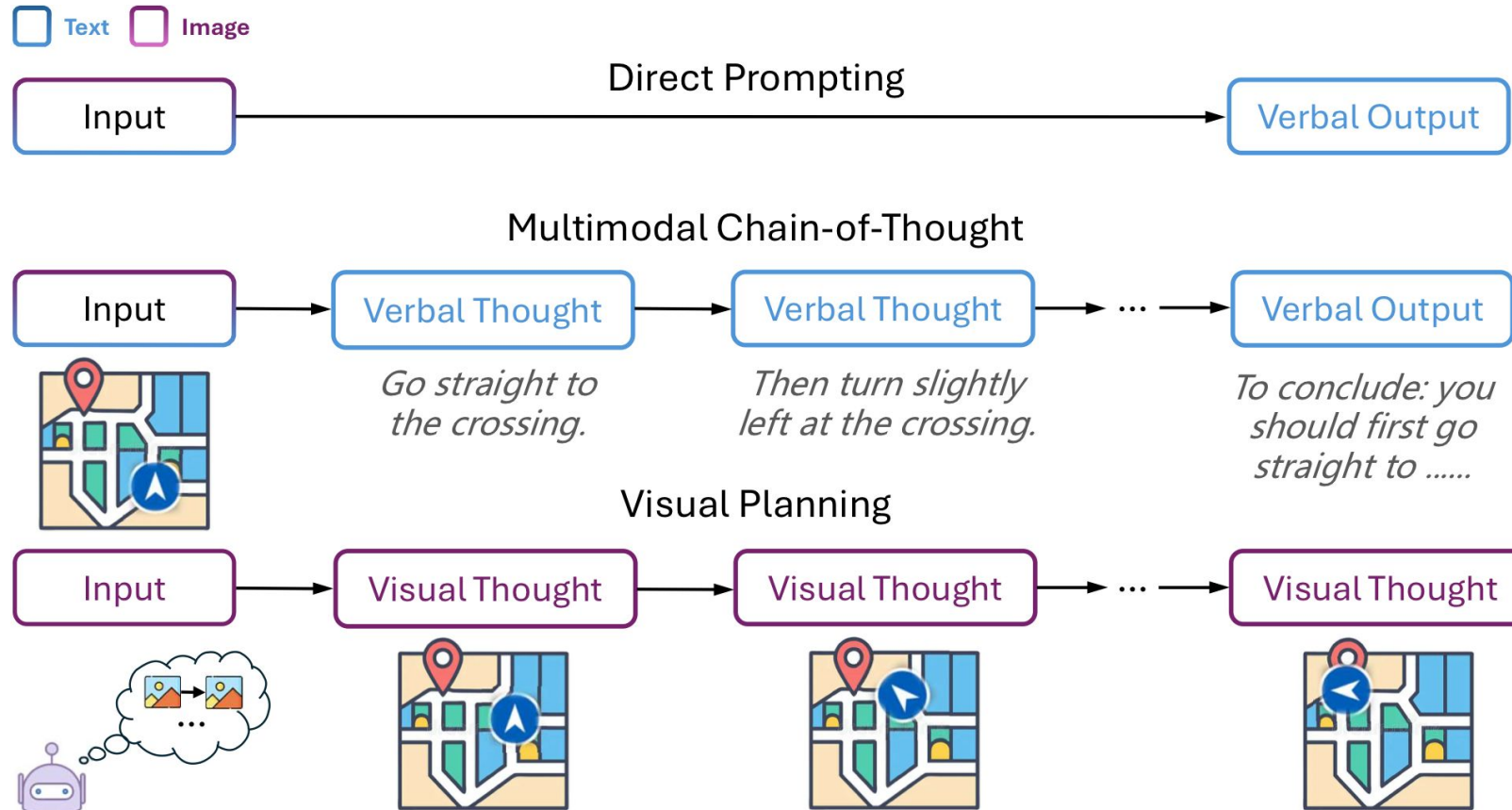
25.7% coordinate mismatches and
22.3% ASCII errors in text-based
plans even after training.

Cognitive Science

Dual Coding Theory: humans think in
both verbal and nonverbal channels

Can models directly plan in non-verbal modalities, such as images, without text mediation?

The Visual Planning Paradigm



Traditional approaches generate verbose textual plans (top & middle).

Visual Planning predicts the next visual state directly, forming a pure image trajectory without language. (bottom)

Proposal: A Two-Stage RL Framework

Stage 1: Policy Initialization

1. Train on random trajectories from random walks.
2. Learn valid visual state sequences.
3. Retain exploration capability.

Stage 2: RL for Visual Planning

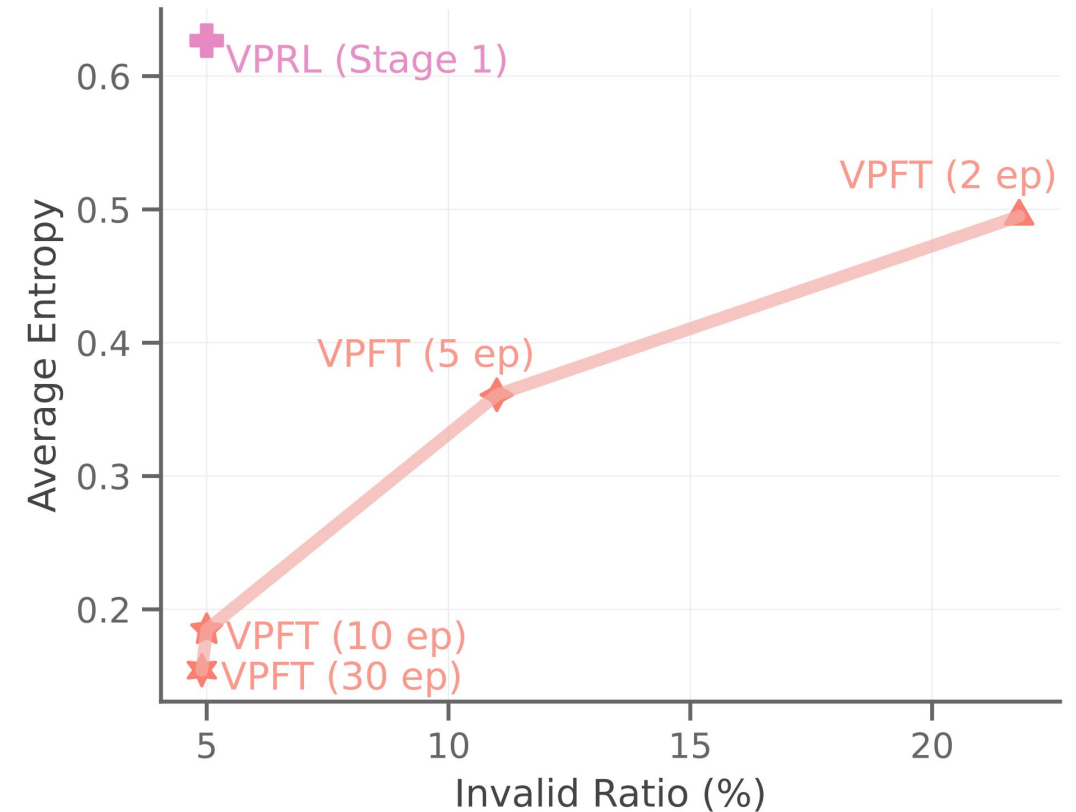
1. Simulate outcomes of potential actions.
2. Based on GRPO: rank candidate responses.
3. Crucial: progress reward function design.

Stage 1: Policy Initialization

Why not SFT on optimal trajectories?

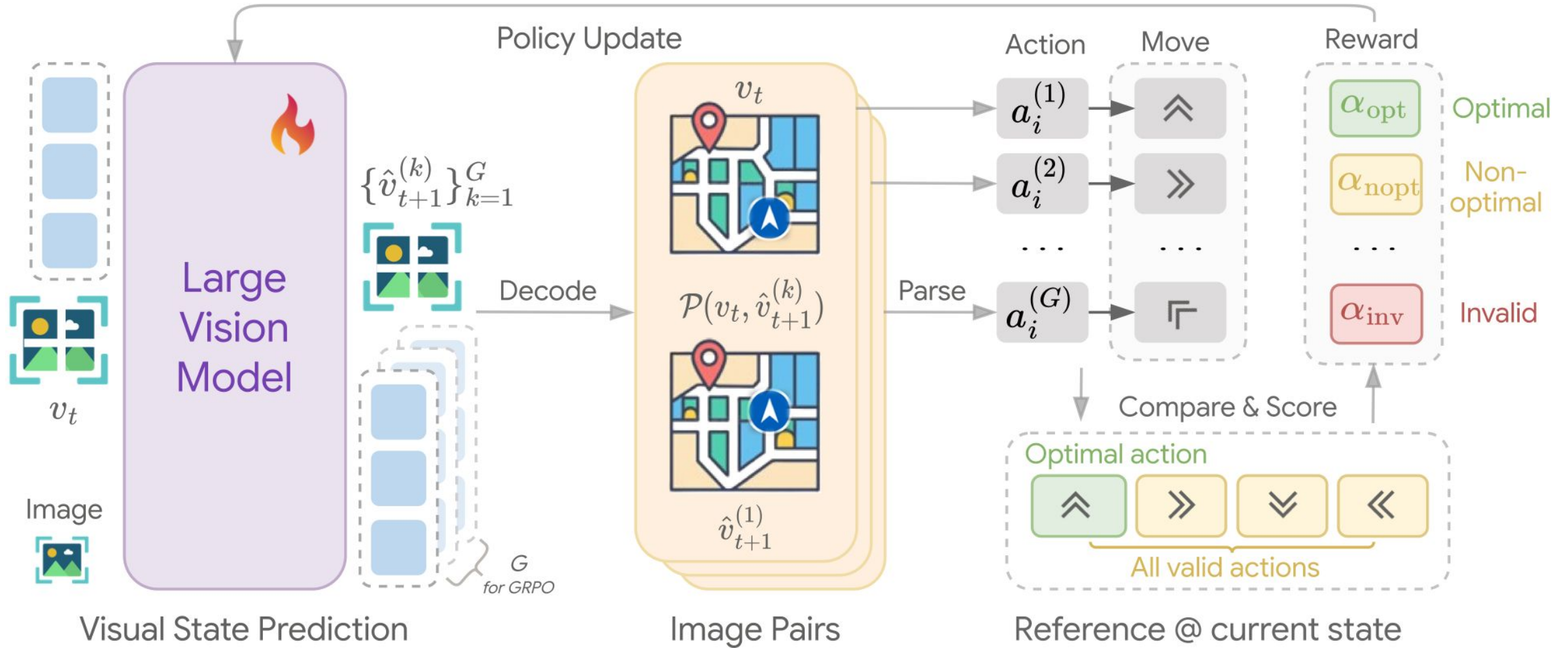
1. Train on random walk trajectories (not optimal ones).
2. Learn valid visual state transitions while retaining exploration capability.
3. Random sampling prevents overfitting to single transitions.
4. Warm-up stage ensuring visually coherent outputs.

Key: VPFT (SFT on optimal) causes *entropy collapse*, blocking RL.



VPFT entropy collapses to zero; Stage 1 maintains near-uniform entropy with low invalid ratio.

Stage 2: RL with GRPO



Stage 2: RL with GRPO – Reward

Stage-Action Parsing Function: Image Pairs \rightarrow Actions

$$\mathcal{P}(v_i, \hat{v}_{i+1}^{(k)}) = \begin{cases} a_i^{(k)}, & \text{if } a_i^{(k)} \in \mathcal{A}, \\ e_i^{(k)}, & \text{if } e_i^{(k)} \in \mathcal{E}. \end{cases}$$

Progress Reward Function

$$\underbrace{\alpha_{\text{opt}} \cdot \mathbb{I}[\mathcal{P}(v_i, \hat{v}_{i+1}^{(k)}) \in \mathcal{A}_{\text{opt}}]}_{\text{optimal}} + \underbrace{\alpha_{\text{nopt}} \cdot \mathbb{I}[\mathcal{P}(v_i, \hat{v}_{i+1}^{(k)}) \in \mathcal{A}_{\text{nopt}}]}_{\text{non-optimal}} + \underbrace{\alpha_{\text{inv}} \cdot \mathbb{I}[\mathcal{P}(v_i, \hat{v}_{i+1}^{(k)}) \in \mathcal{E}_{\text{inv}}]}_{\text{invalid}},$$

$$\alpha_{\text{opt}} = 1.$$

optimal

$$\alpha_{\text{nopt}} = 0$$

non-optimal

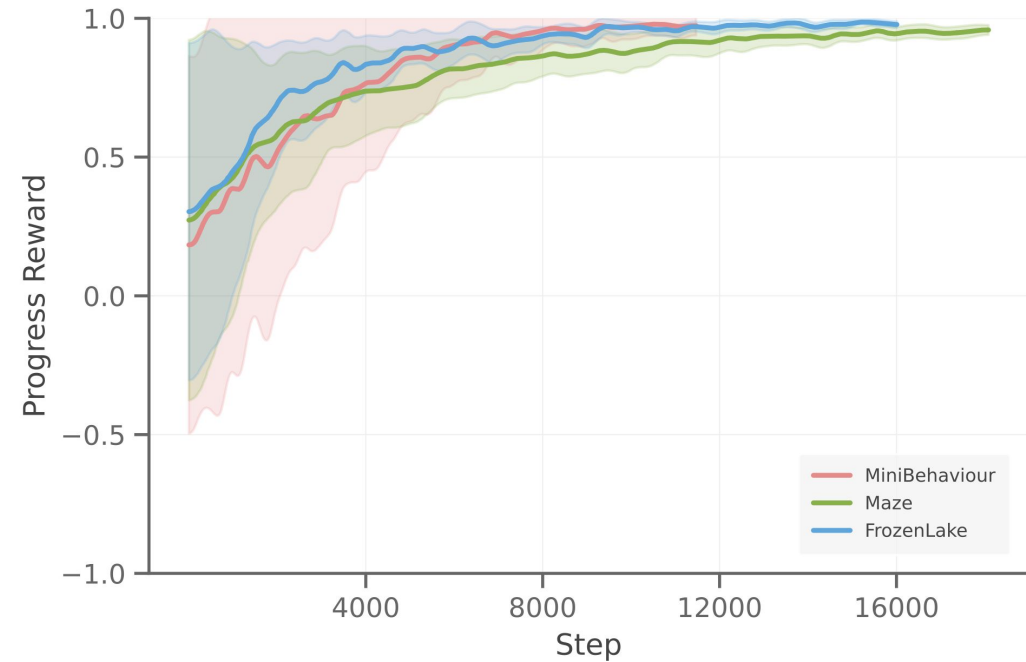
$$\alpha_{\text{inv}} = -5.$$

invalid

Stage 2: RL with GRPO — Details

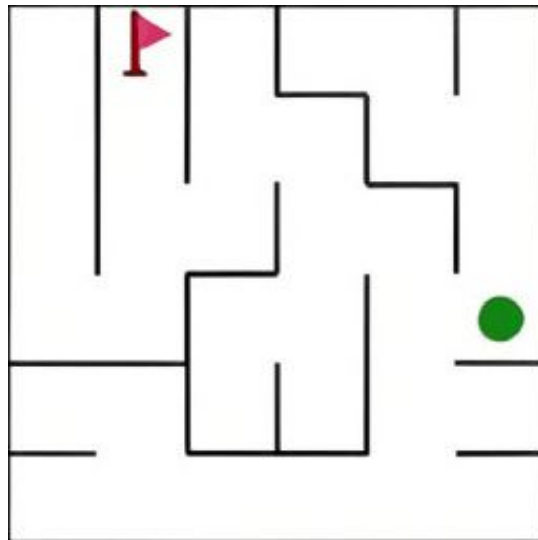
Group Relative Policy Optimization

1. Generate G candidate next visual states.
2. Rank candidates by reward within the group.
3. Relative advantage without a learned critic.
4. KL penalty ($\beta=0.001$) prevents policy collapse.



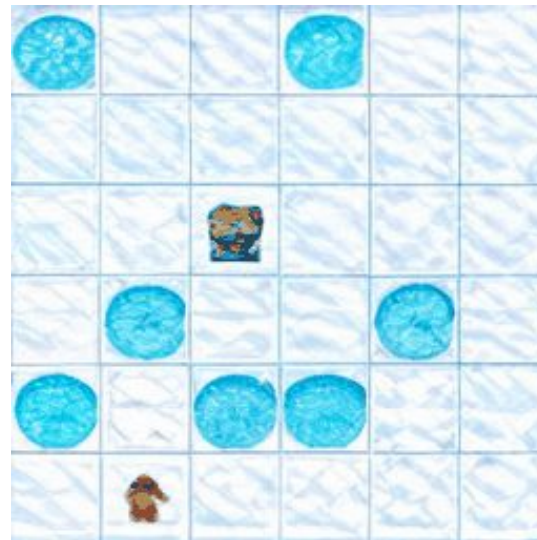
Evaluation Tasks: Grid-Based Navigation

Maze



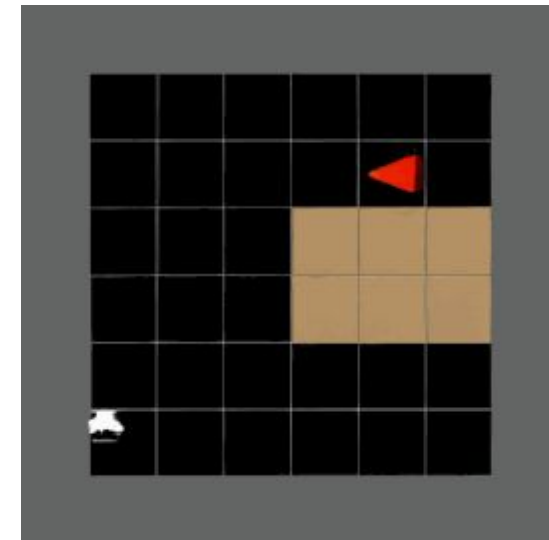
Navigate from start to goal avoiding walls

FrozenLake



Reach destination avoiding holes on ice

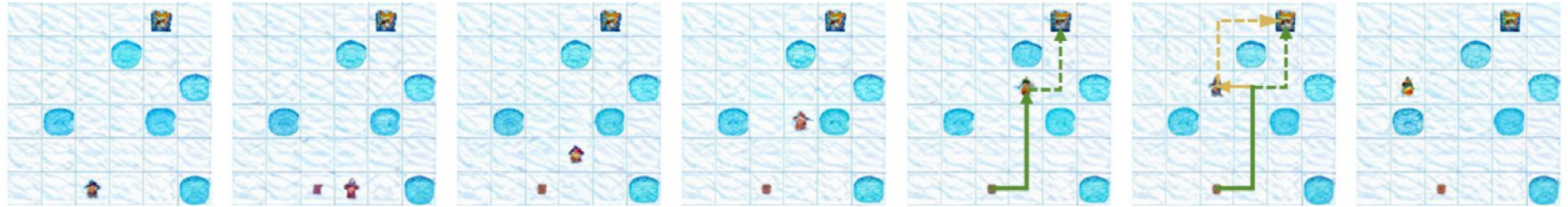
MiniBehavior



Pick up printer, carry to table, drop it

Generated Visual Planning Traces

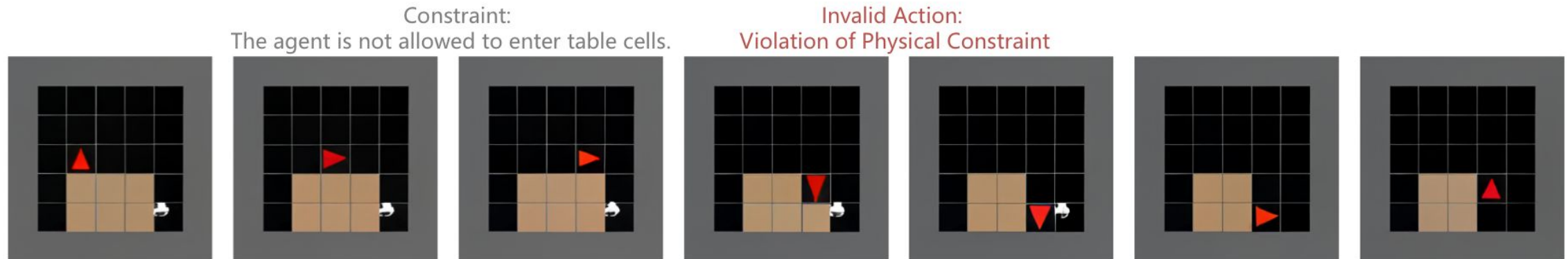
FrozenLake



Maze



MiniBehavior



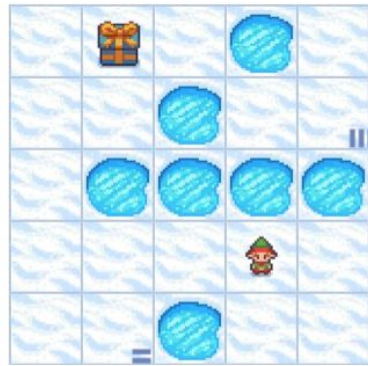
Main Results

Table 1: Performance of the closed- and open-source models on FROZENLAKE, MAZE, and MINIBEHAVIOR. VPRL performs consistently the best (**bold**) across all tasks. [†] denotes the post-trained model. **A** represents texts and represents images. The last column AVG. reports the average performance across three tasks.

Model	Input	Output	FROZENLAKE		MAZE		MINIBEHAVIOR		AVG.	
			EM (%)	PR (%)	EM (%)	PR (%)	EM (%)	PR (%)	EM (%)	PR (%)
Closed-Source Model										
Gemini 2.0 Flash										
- Direct	A+	A	21.2	47.6	8.3	31.4	0.7	29.8	10.1	36.3
- CoT	A+	A	27.6	52.5	6.9	29.8	4.0	31.2	12.8	37.8
Gemini 2.5 Pro (<i>think</i>)	A+	A	72.0	85.0	21.5	35.5	37.6	59.9	43.7	60.1
Open-Source Model										
Qwen 2.5-VL-Instruct-7B										
- Direct	A+	A	1.2	15.0	0.6	14.5	0.3	9.8	0.7	13.1
- CoT	A+	A	8.2	29.1	2.3	15.2	0.5	14.7	3.7	19.7
- SFT [†]	A+	A	68.6	84.4	60.9	70.3	31.3	56.1	53.6	69.9
LVM-7B										
- VPFT [†] (ours)			75.4	79.5	59.0	64.0	33.8	52.2	56.1	65.2
- VPRL [†] (ours)			91.6	93.2	74.5	77.6	75.8	83.8	80.6	84.9

VPRL achieves 80.6% avg EM — 27% higher than best text-based approach (SFT: 53.6%)

Visual Planning vs. Language-Based Planning



Prompt: You are given an image of a grid-based environment., your task is to generate the shortest valid sequence of actions., provide your final answer enclosed between <ANSWER> and </ANSWER>.

Gemini 2.5 Pro

<think> I frame the picture as a 5×7 grid, label the elf, gift, and every ice hole, and plan a Breadth-First Search., after expanding every reachable cell, the queue drains while row 2, a full wall of five holes, blocks all upward routes, [7000+ words omitted], I re-inspect the grid, and confirm the proof: no valid path exists. </think>
So the correct reply is an empty action list: <ANSWER></ANSWER>

Qwen 2.5 VL Instruct-SFT

<ANSWER>
left left up up up right
</ANSWER>
Violation of Constraint

Non-Optimal Action



How Does RL Help?

VPRL Reduces Invalid Actions

Invalid-failure ratio (% with ≥ 1 invalid action):

FrozenLake: VPFT 60.6% \rightarrow VPRL 36.9% (\downarrow 24%)

Maze: VPFT 73.7% \rightarrow VPRL 25.1% (\downarrow 49%)

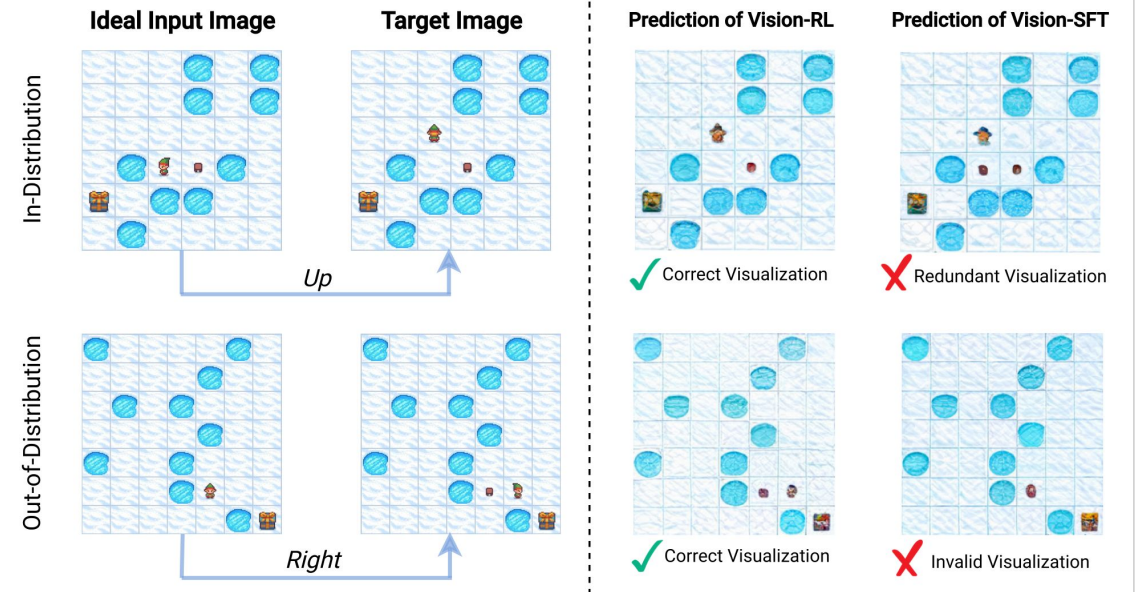
MiniBehavior: VPFT 78.3% \rightarrow VPRL 29.6% (\downarrow 49%)

RL learns constraints through exploration.

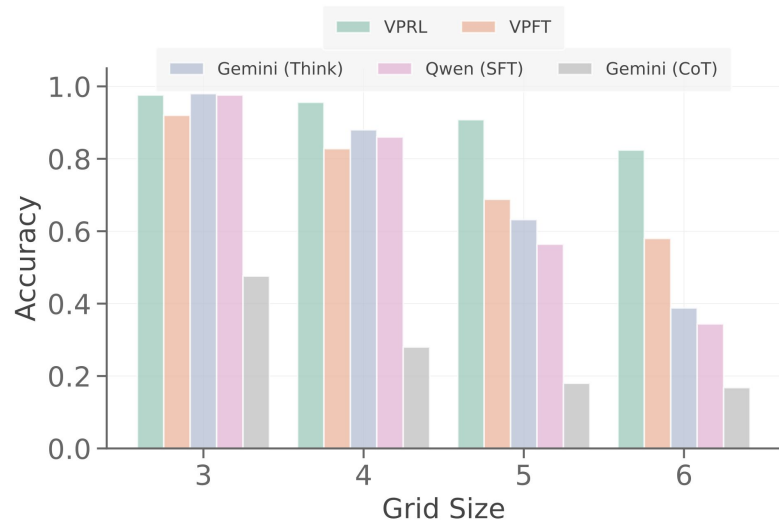
Key Insight

VPRL shows flexibility — bypasses obstacles and recovers from detours, while VPFT gets stuck repeating same actions.

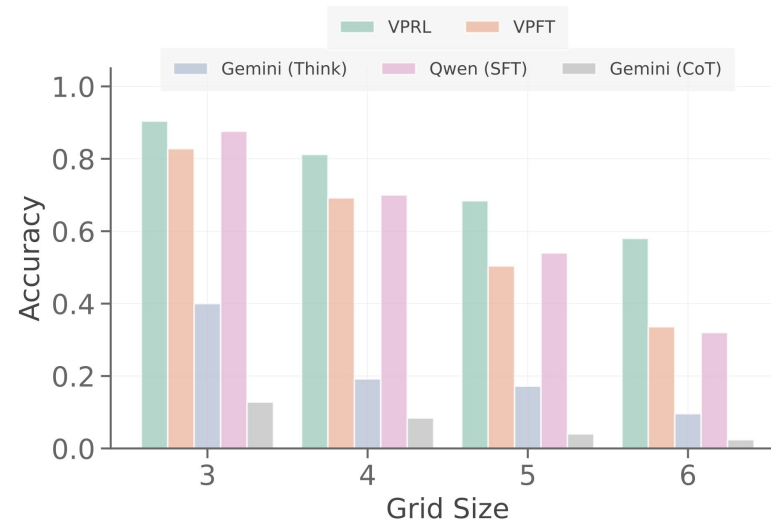
Qualitative Comparison



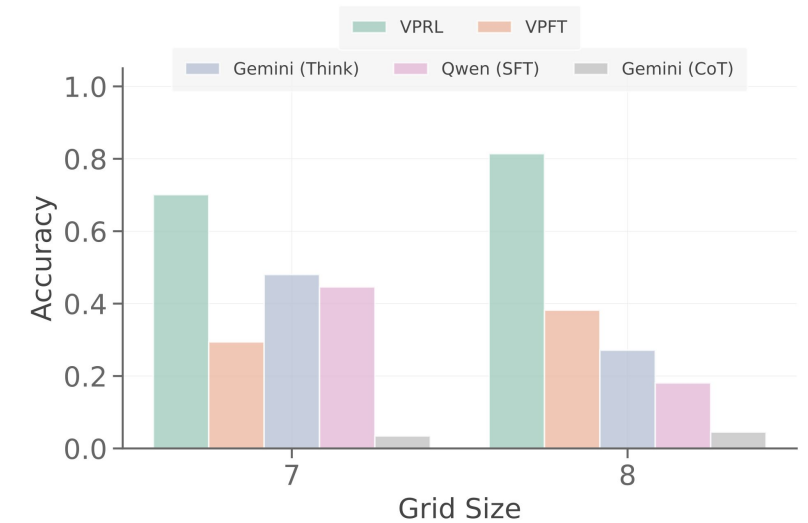
Performance by Task Complexity



FrozenLake



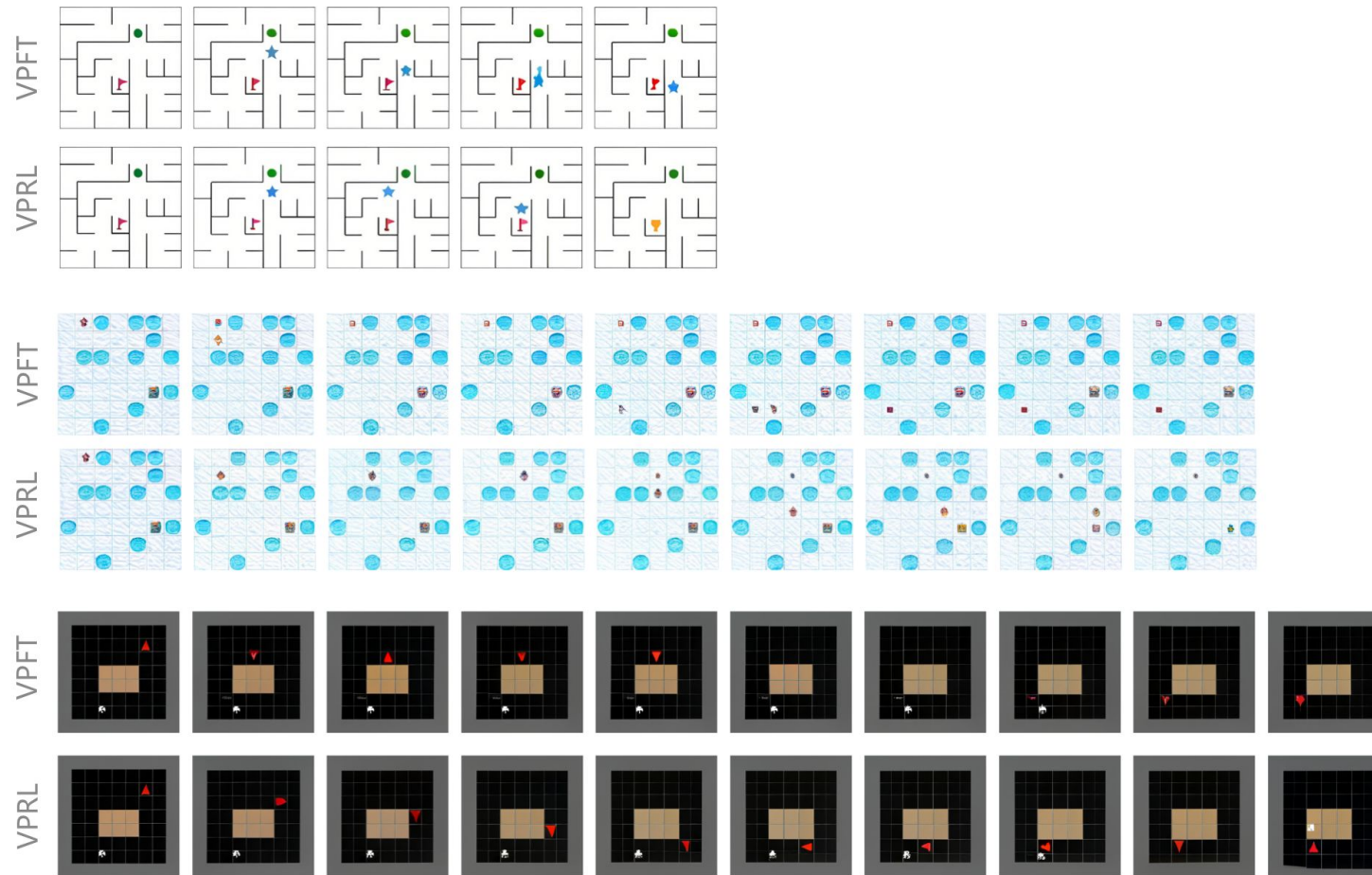
Maze



MiniBehavior

VPRL maintains robust performance as complexity increases (97.6% → 82.4% on FrozenLake)
Gemini 2.5 Pro drops sharply (98.0% → 38.8%) — text-based reasoning does not scale

Out-of-Distribution Generalization



SFT models fail completely on OOD grid sizes. VPRL preserves 20.4% EM and 31.2% PR on unseen larger grids on FrozenLake — learns generalizable strategies

Conclusion

New Paradigm

Visual Planning — reasoning as sequences of images, without any text mediation

Novel Framework

VPRL: two-stage RL with GRPO for training Large Vision Models to plan visually

Strong Results

80.6% avg EM, 27% higher than best text-based method; robust across complexity levels

Generalization

VPRL learns transferable strategies; generalizes to OOD scenarios where SFT fails

*Opening a new direction: **vision-first reasoning** for spatial, geometric, and physical tasks*

Thank You!

Visual Planning: Let's Think Only with Images

Speaker: Han Zhou

Language Technology Lab, University of Cambridge

Contact: hz416@cam.ac.uk

Paper: arxiv.org/abs/2505.11409

