GenPlan: Generative Sequence Models as Adaptive Planners

Akash Karthikeyan, Yash Vardhan Pant
Department of Electrical and Computer Engineering,
University of Waterloo.

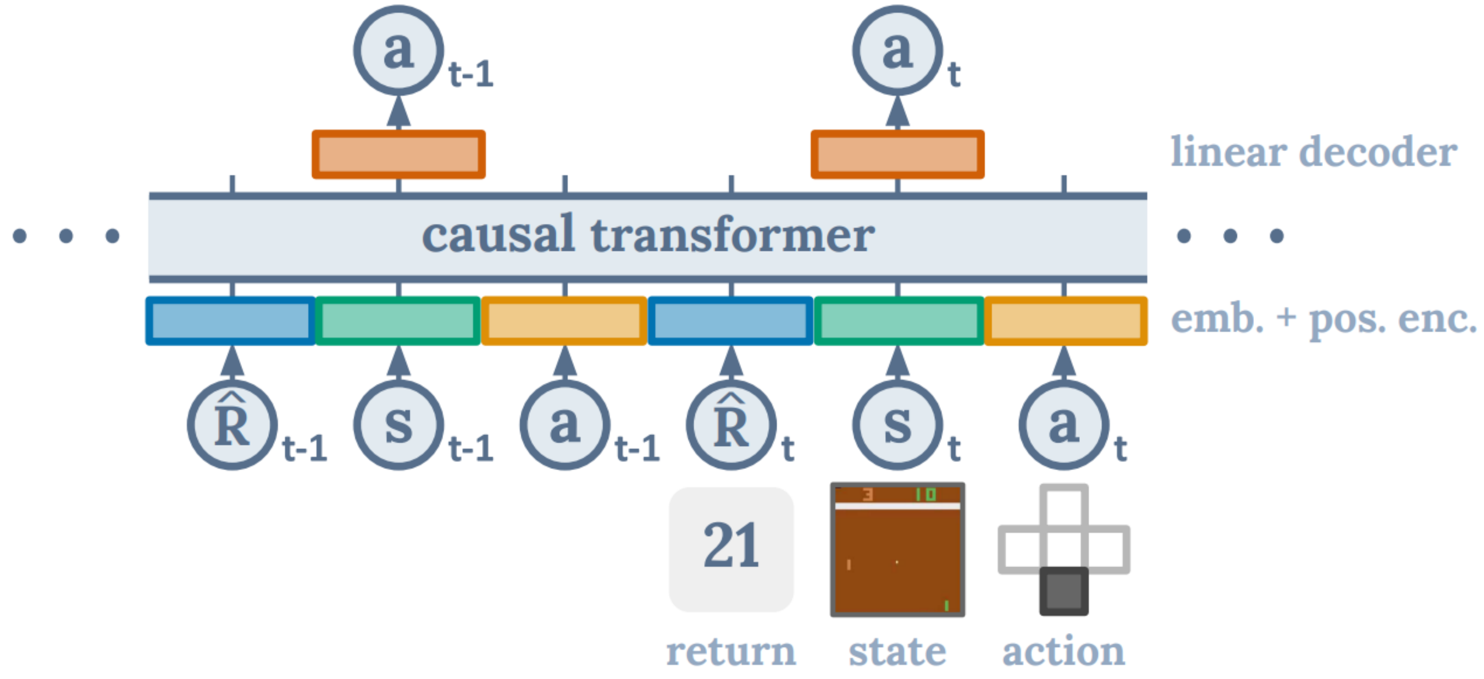
Accepted at AAAI Conference on Artificial Intelligence, 2025

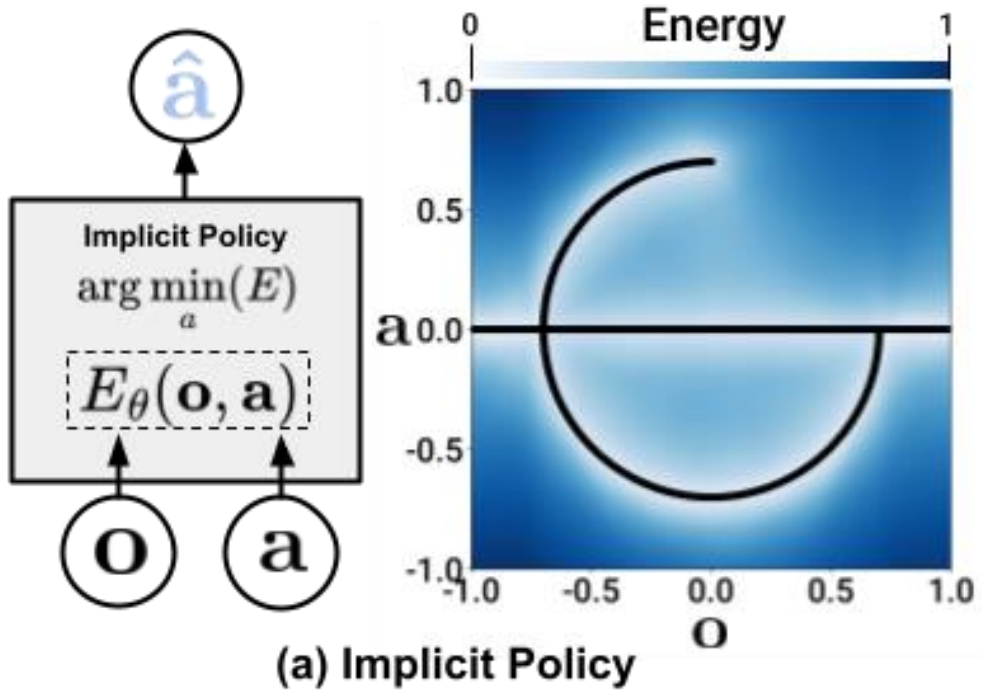
AAAI-25 / IAAI-25 / EAAI-25

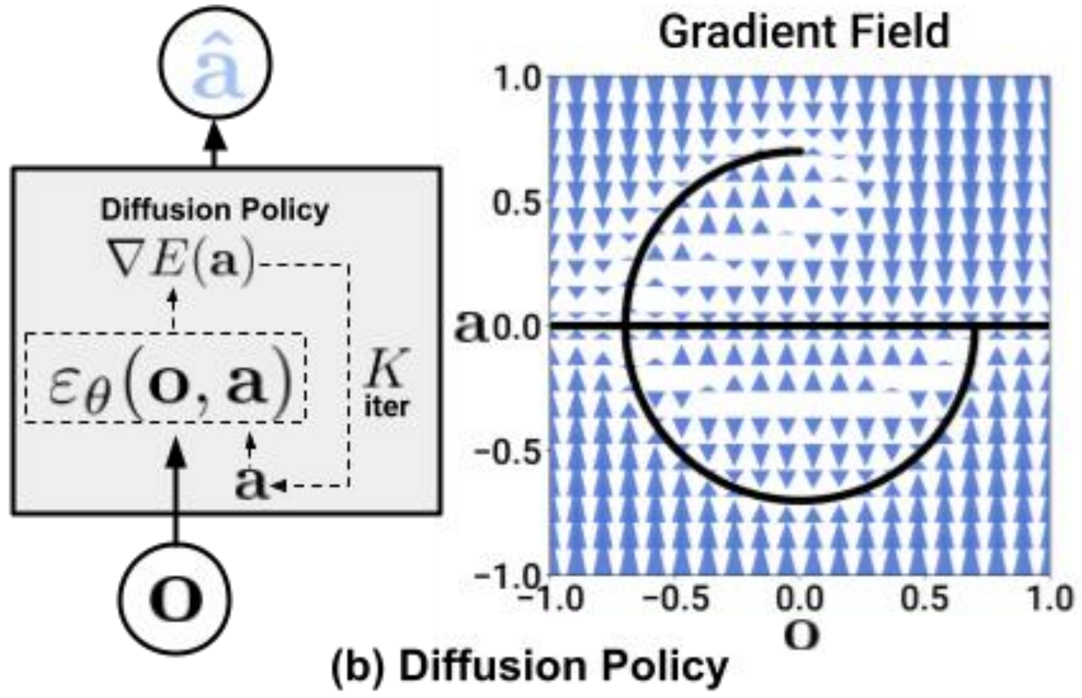
01 INTRODUCTION

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Planning as Behavioral Cloning



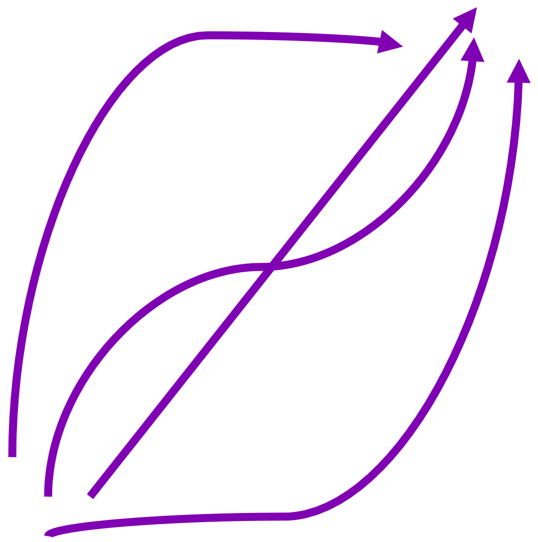




Decision Transformer [1] **Implicit Behavioral Cloning [2a]** **Diffusion Policy [2b]**

[1] L. Chen *et al.*, “Decision Transformer: Reinforcement Learning via Sequence Modeling,” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 15084–15097.
[2a] P. Florence *et al.*, “Implicit Behavioral Cloning,” Sep. 01, 2021, *arXiv*: arXiv:2109.00137. doi: 10.48550/arXiv.2109.00137.
[2b] C. Chi *et al.*, “Diffusion Policy: Visuomotor Policy Learning via Action Diffusion,” Jun. 01, 2023, *arXiv*: arXiv:2303.04137. doi: 10.48550/arXiv.2303.04137.

Learning alphabet of actions



Action Dataset

$\Rightarrow$

This is continuous

Hard to learn multi-modal distributions!

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Learning alphabet of actions

$\mathbf{e}_{12}$
$\mathbf{e}_{2}$
$\mathbf{e}_{9}$
$\mathbf{e}_{5}$
$\mathbf{e}_{15}$

$\Rightarrow$

This is discrete

Easy to learn multi-modal distributions!

Latent Actions/ Embeddings

S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto, "Behavior Generation with Latent Actions," Jun. 28, 2024, *arXiv*: arXiv:2403.03181.
N. M. M. Shafiullah, Z. J. Cui, A. Altanzaya, and L. Pinto, "Behavior Transformers: Cloning k modes with one stone," Oct. 11, 2022, *arXiv*: arXiv:2206.11251.

Planning ?

Problem …..

Prior works often require well-represented train demonstrations, and fail to generalize to harder tasks.

Given …..

Demonstrations of only sub-tasks (partial goals)
(noisy, unlabelled, short (temporal) horizon)

Goal …..

- Learn a planner that optimizes at the sequential level.
- Abstract reasoning to generalize across tasks.
- Using simple demonstrations to adapt to harder tasks.

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

02 MOTIVATION AND HIGHLIGHTS

02 MOTIVATION AND HIGHLIGHTS

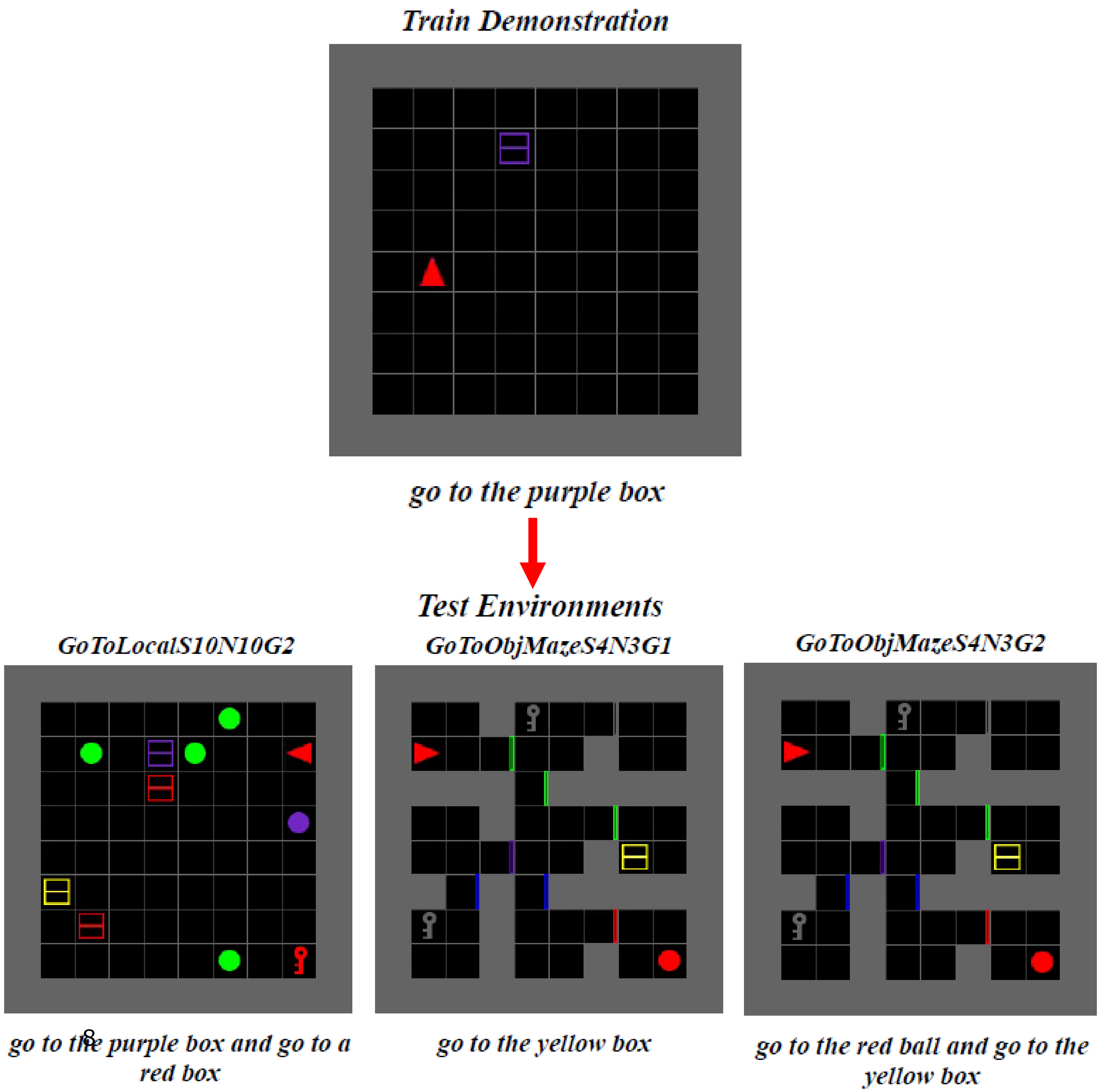
2A Adaptation to Harder Task

2B Unconditional Rollouts

2C Skill Composition / Multi-Modality

2D Intermediate Representations for Trajectory Optimization

2E Can we learn all in a sample-efficient manner?



02 MOTIVATION AND HIGHLIGHTS

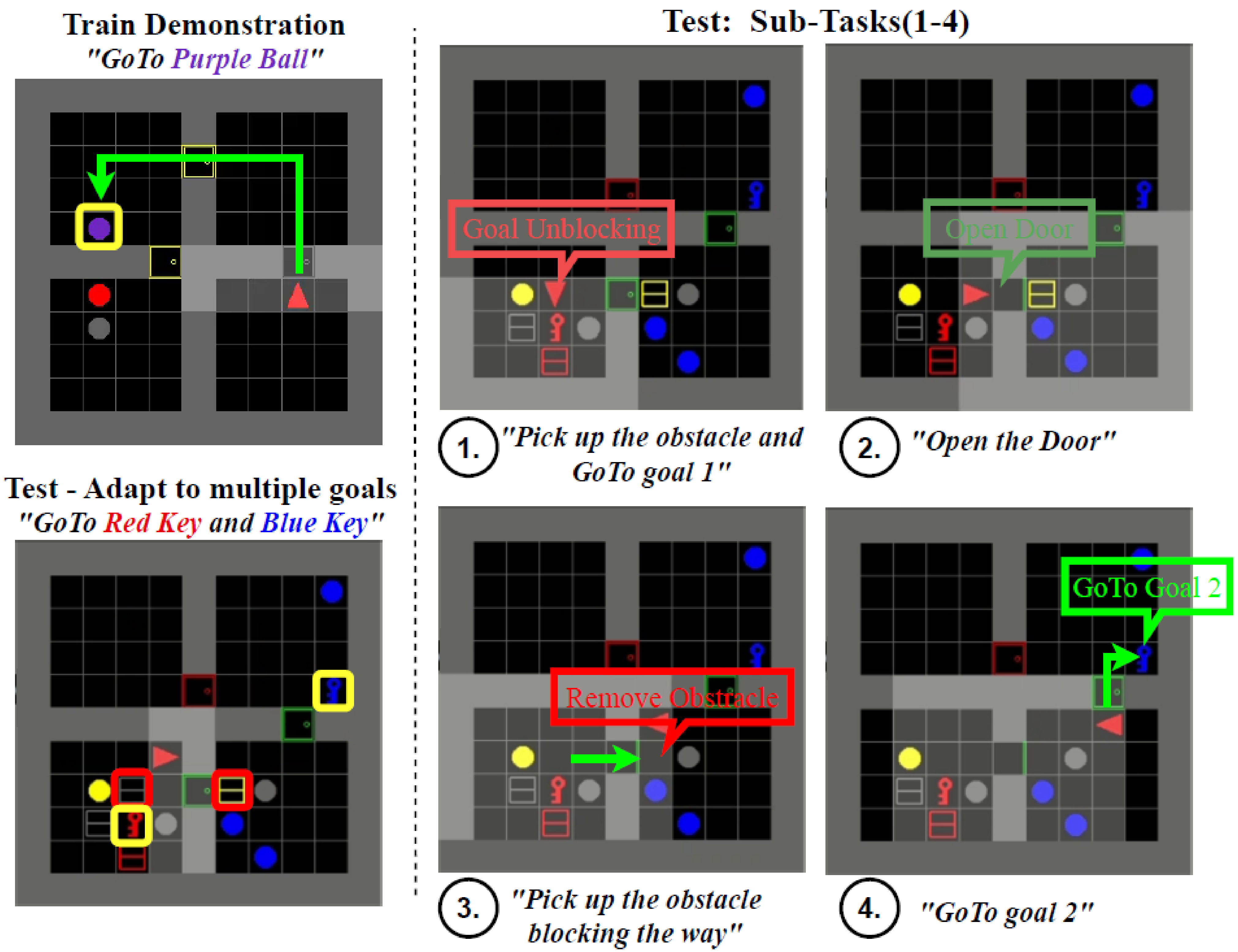
2A Adaptation to Harder Task

2B Unconditional Rollouts

2C Skill Composition / Multi-Modality

2D Intermediate Representations for Trajectory Optimization

2E Can we learn all in a sample-efficient manner?



UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

02 MOTIVATION AND HIGHLIGHTS

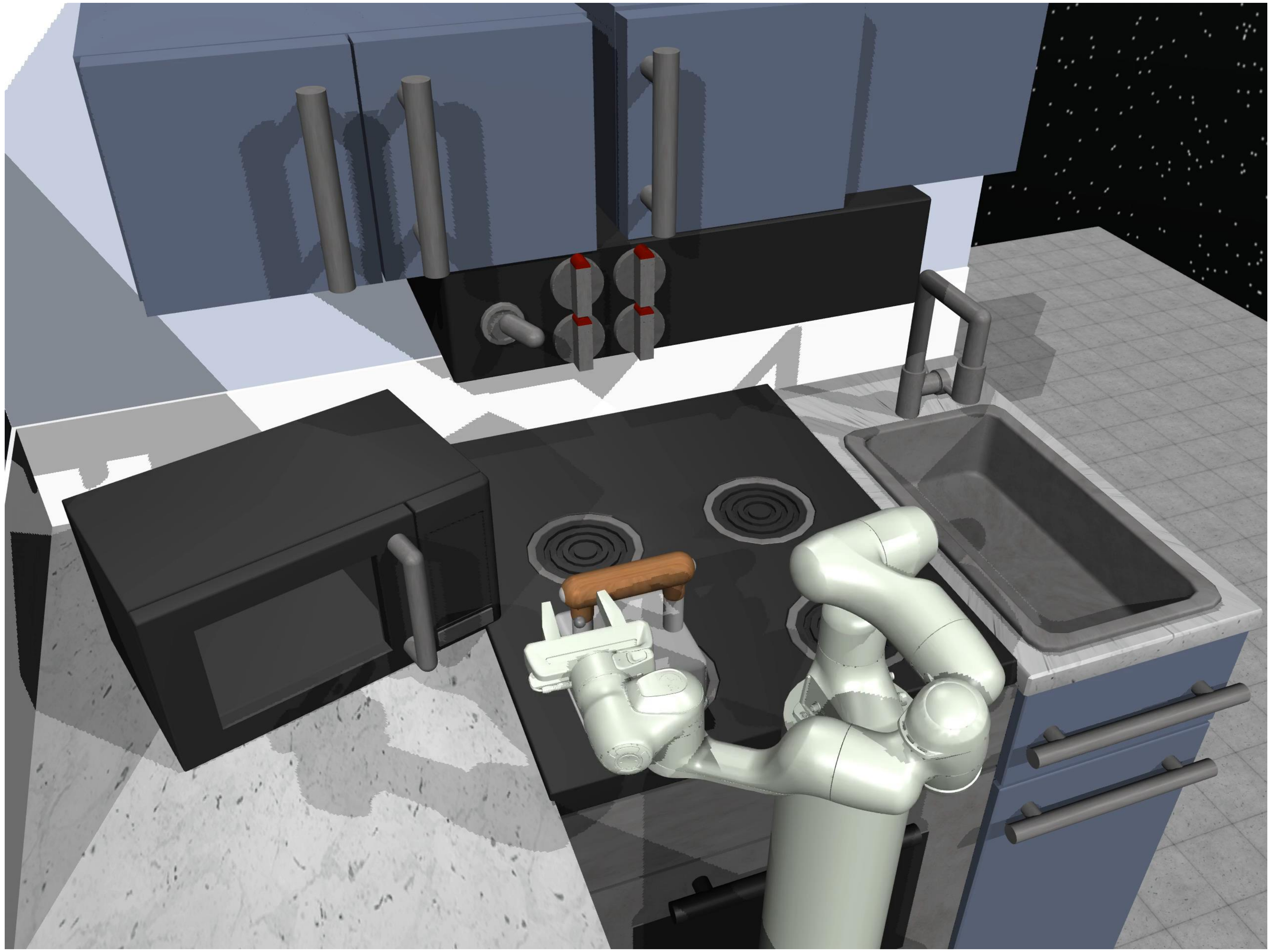
2A Adaptation to Harder Task

2B Unconditional Rollouts

2C Skill Composition / Multi-Modality

2D Intermediate Representations for Trajectory Optimization

2E Can we learn all in a sample-efficient manner?



Gupta, A.; Kumar, V.; Lynch, C.; Levine, S.; and Hausman, K. 2019. *Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning*. arXiv preprint arXiv:1910.11956

02 MOTIVATION AND HIGHLIGHTS

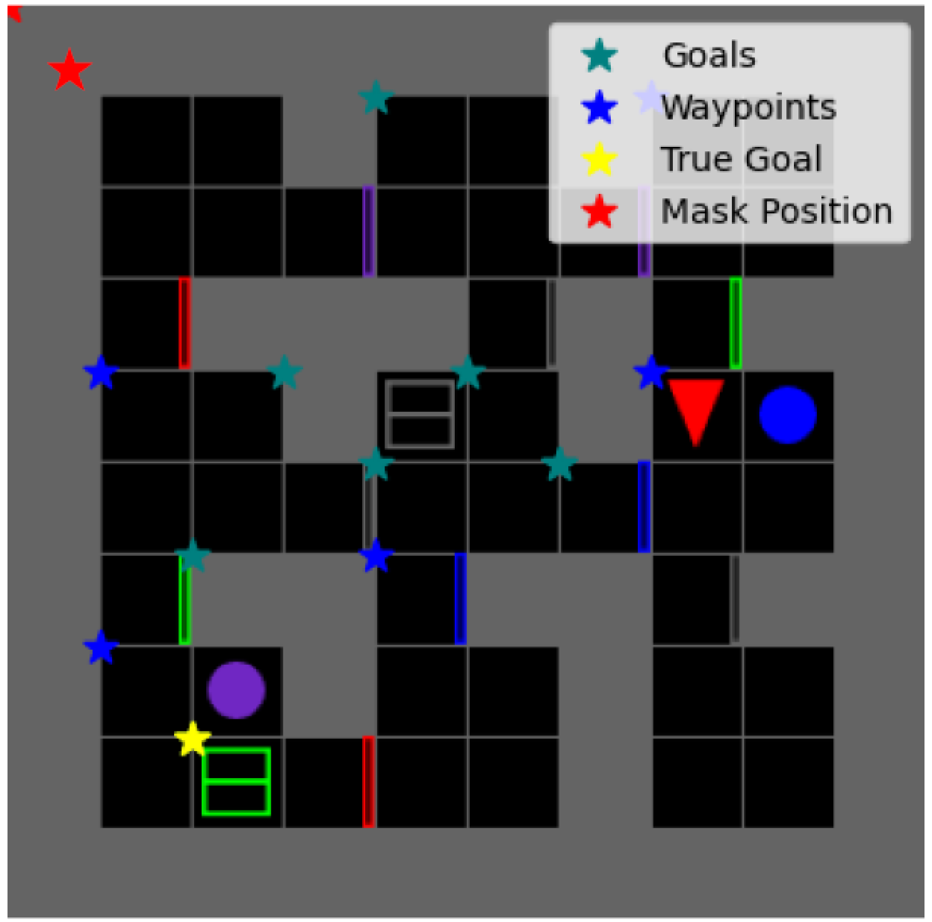
2A Adaptation to Harder Task

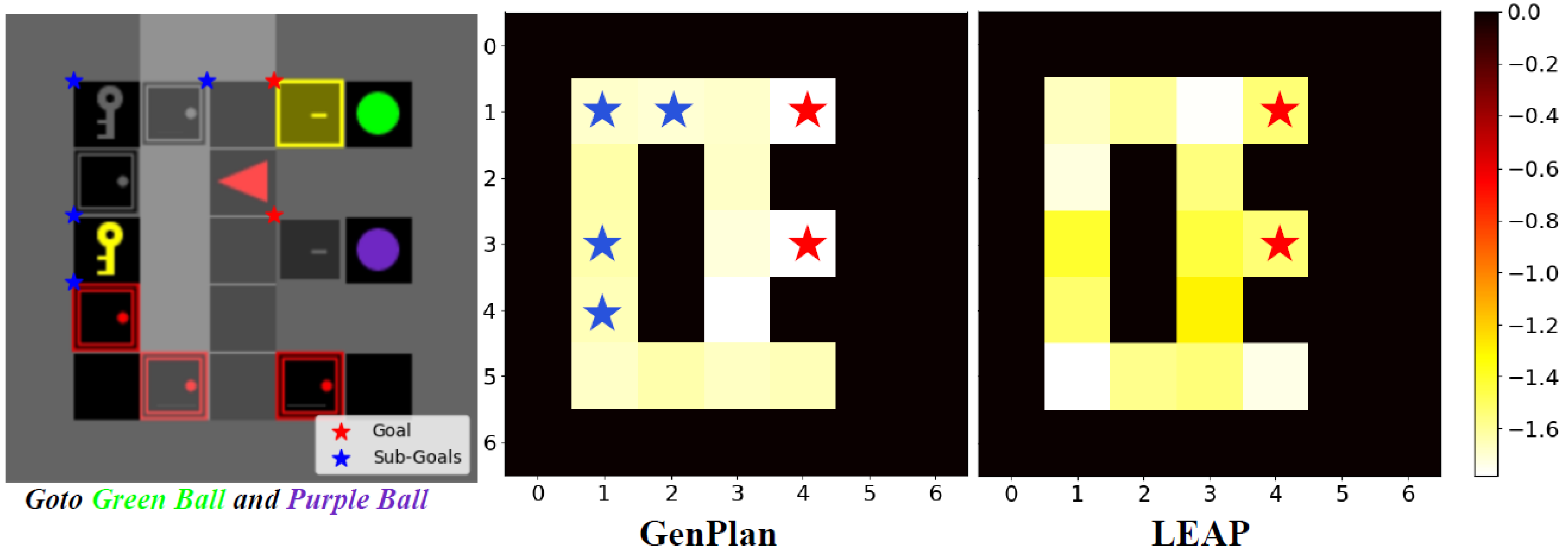
2B Unconditional Rollouts

2C Skill Composition / Multi-Modality

2D Intermediate Representations for Trajectory Optimization

2E Can we learn all in a sample-efficient manner?





02 MOTIVATION AND HIGHLIGHTS

2A Adaptation to Harder Task

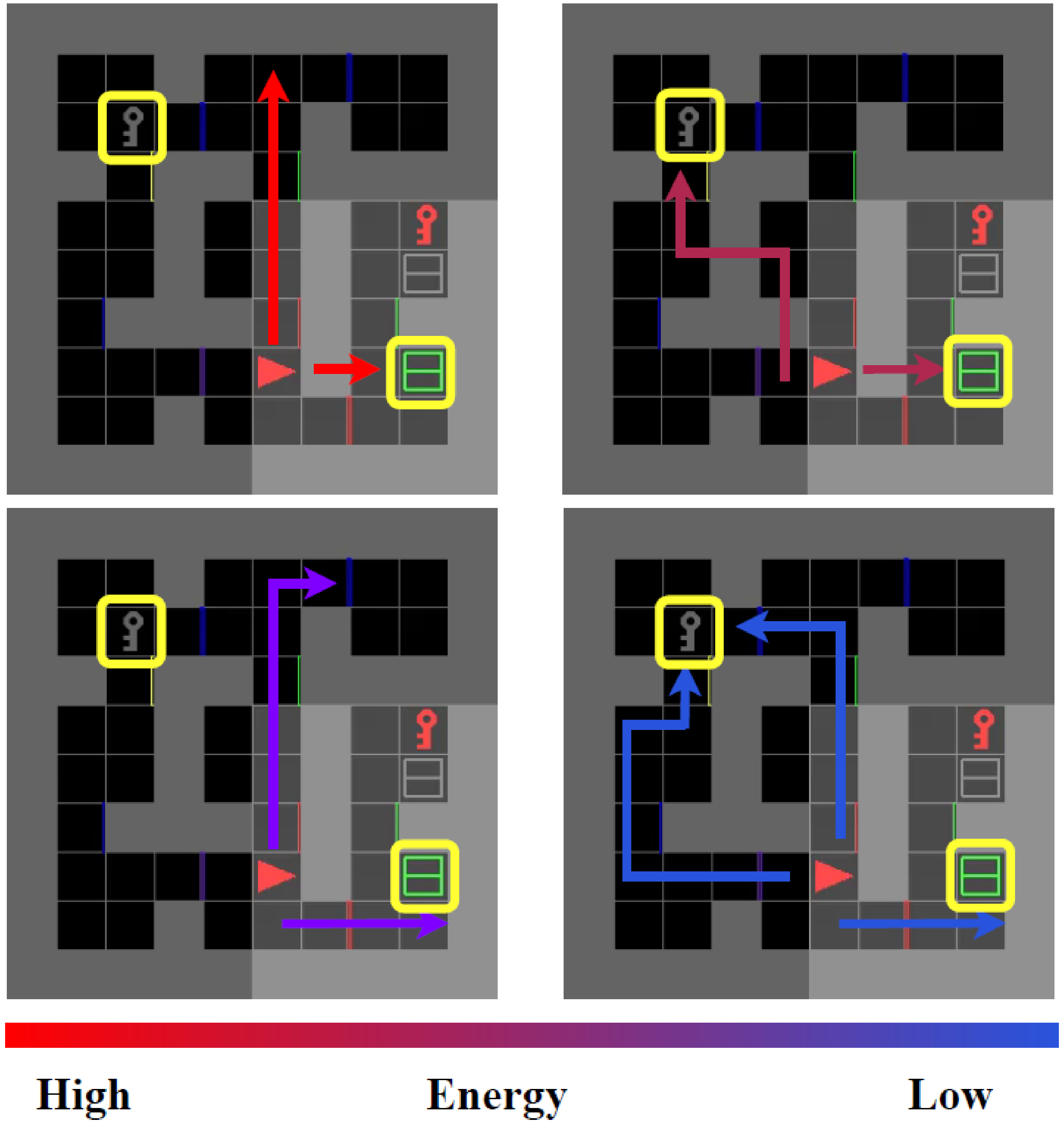
2B Unconditional Rollouts

2C Skill Composition / Multi-Modality

2D Intermediate Representations for Trajectory Optimization

2E Can we learn all in a sample-efficient manner?

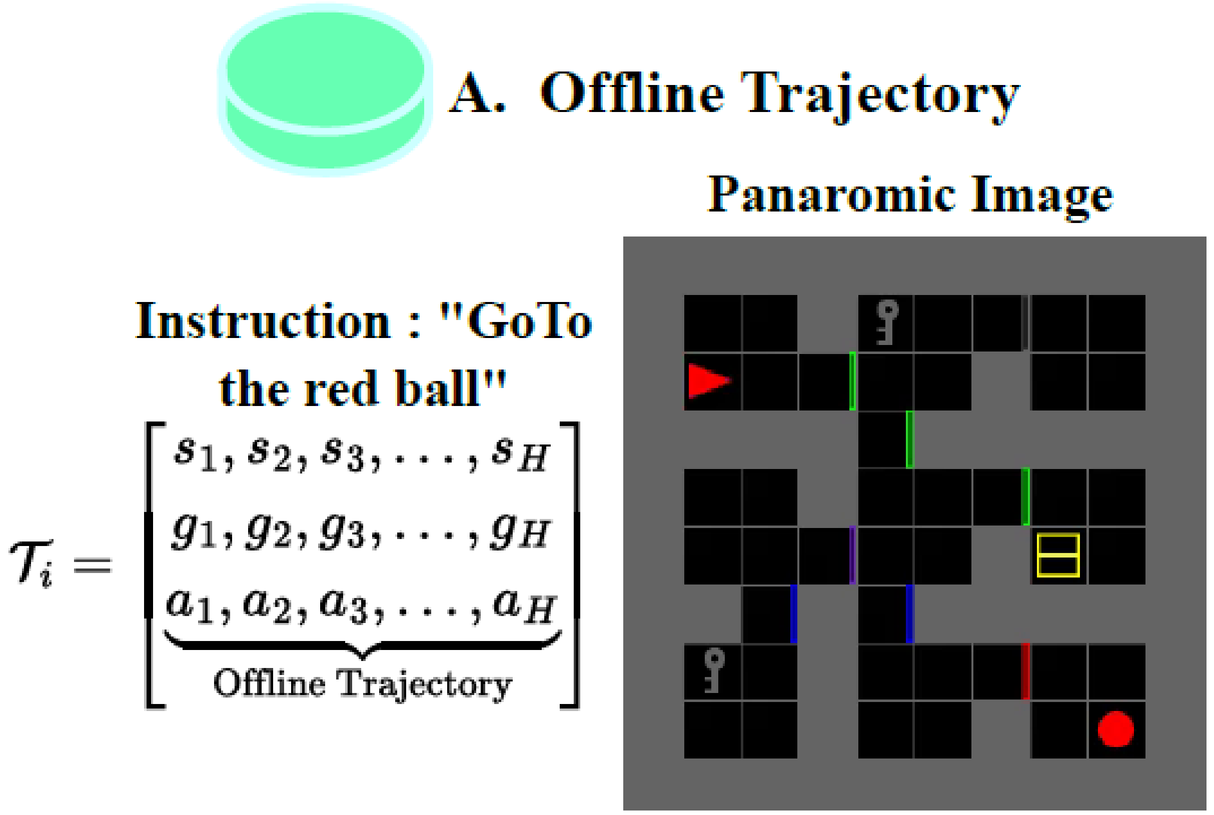
Task: "GoTo Green Box and Gray Key"



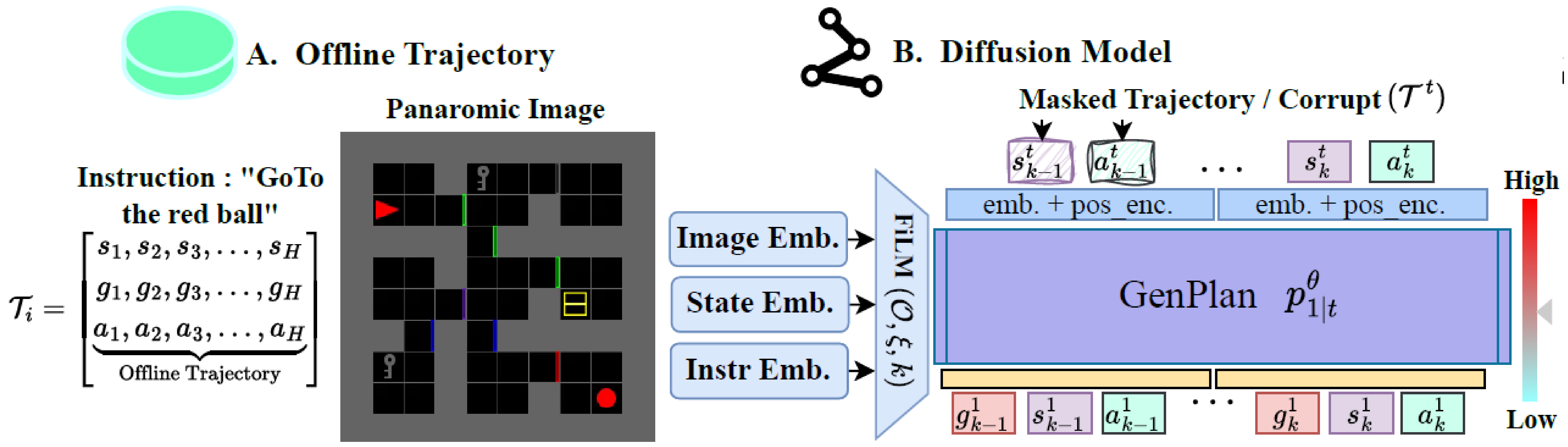
*Note. None of the plans are actually executed until the final minimum energy trajectory is obtained, this is only an illustration of the process

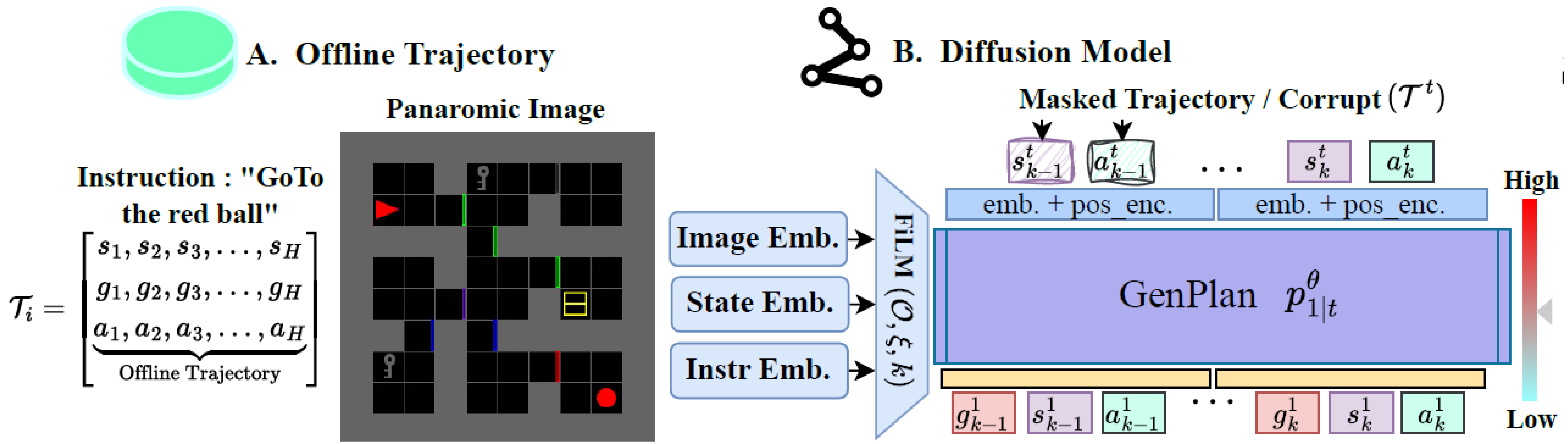
03 GENPLAN - METHOD

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING



A. We collect offline demonstration from the environment (# 500)

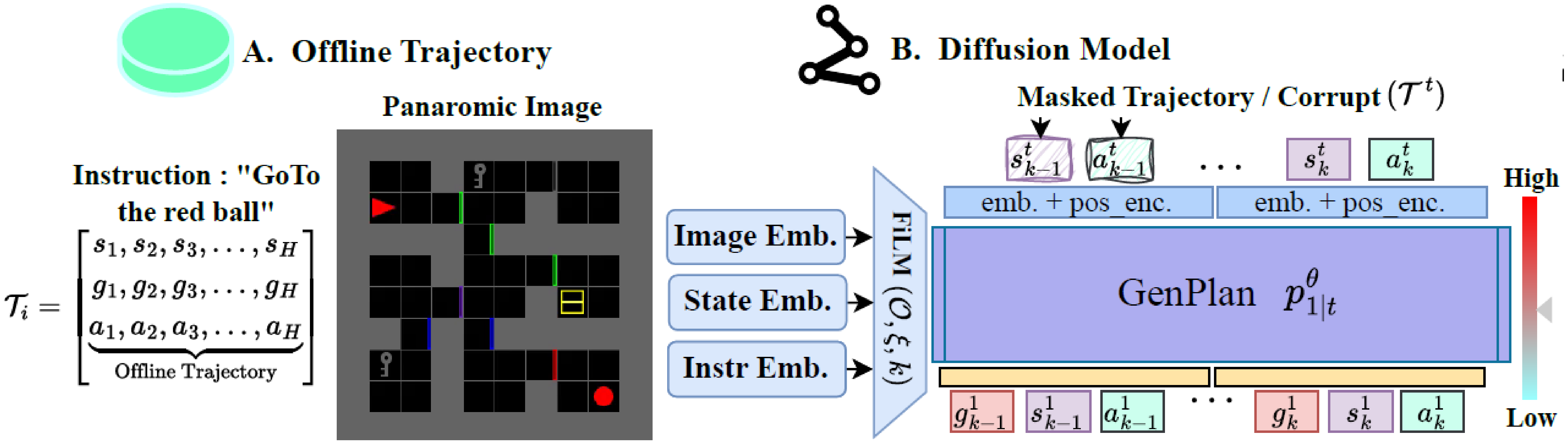

A. Offline Trajectory
Panaromic Image
Instruction : "GoTo the red ball"
$\mathcal{T}_i = \begin{bmatrix} s_1, s_2, s_3, \ldots, s_H \\ g_1, g_2, g_3, \ldots, g_H \\ \underbrace{a_1, a_2, a_3, \ldots, a_H}_{\text{Offline Trajectory}} \end{bmatrix}$
B. Diffusion Model
Masked Trajectory / Corrupt ($\mathcal{T}^t$)
s^t_{k-1} a^t_{k-1} . . . s^t_k a^t_k
emb. + pos_enc.
emb. + pos_enc.
Image Emb.
State Emb.
Instr Emb.
FiLM ($\mathcal{O}, \xi, k$)
GenPlan $p^\theta_{1|t}$
g^1_{k-1} s^1_{k-1} a^1_{k-1} . . . g^1_k s^1_k a^1_k
High
Low



$$\min_{\theta} \mathbb{E}_{\boldsymbol{a}^0 \sim p_0, \boldsymbol{o} \sim \mathcal{D}} \left[\sum_{k=1}^{H} -\log p^{\theta}_{1|t}(\boldsymbol{a}^1 | \boldsymbol{a}^0, \boldsymbol{o}) \right] \quad \text{s.t. } \mathbb{E}_{\boldsymbol{a}^0 \sim p_0, \boldsymbol{o} \sim \mathcal{D}} \left[\sum_{k=1}^{H} \mathcal{H}(p^{\theta}_{1|t}(\boldsymbol{a} | \boldsymbol{a}^0, \boldsymbol{o})) \right] \geq \beta$$

Energy Objective

$$\mathcal{E}(\boldsymbol{a}) \simeq \mathcal{E}(\boldsymbol{a}')$$



$$\min_\theta \mathbb{E}_{\boldsymbol{a}^0 \sim p_0, \boldsymbol{o} \sim \mathcal{D}} \left[\sum_{k=1}^{H} -\log p^\theta_{1|t}(\boldsymbol{a}^1 | \boldsymbol{a}^0, \boldsymbol{o}) \right] \quad \text{s.t. } \mathbb{E}_{\boldsymbol{a}^0 \sim p_0, \boldsymbol{o} \sim \mathcal{D}} \left[\sum_{k=1}^{H} \mathcal{H}(p^\theta_{1|t}(\boldsymbol{a} | \boldsymbol{a}^0, \boldsymbol{o})) \right] \geq \beta$$

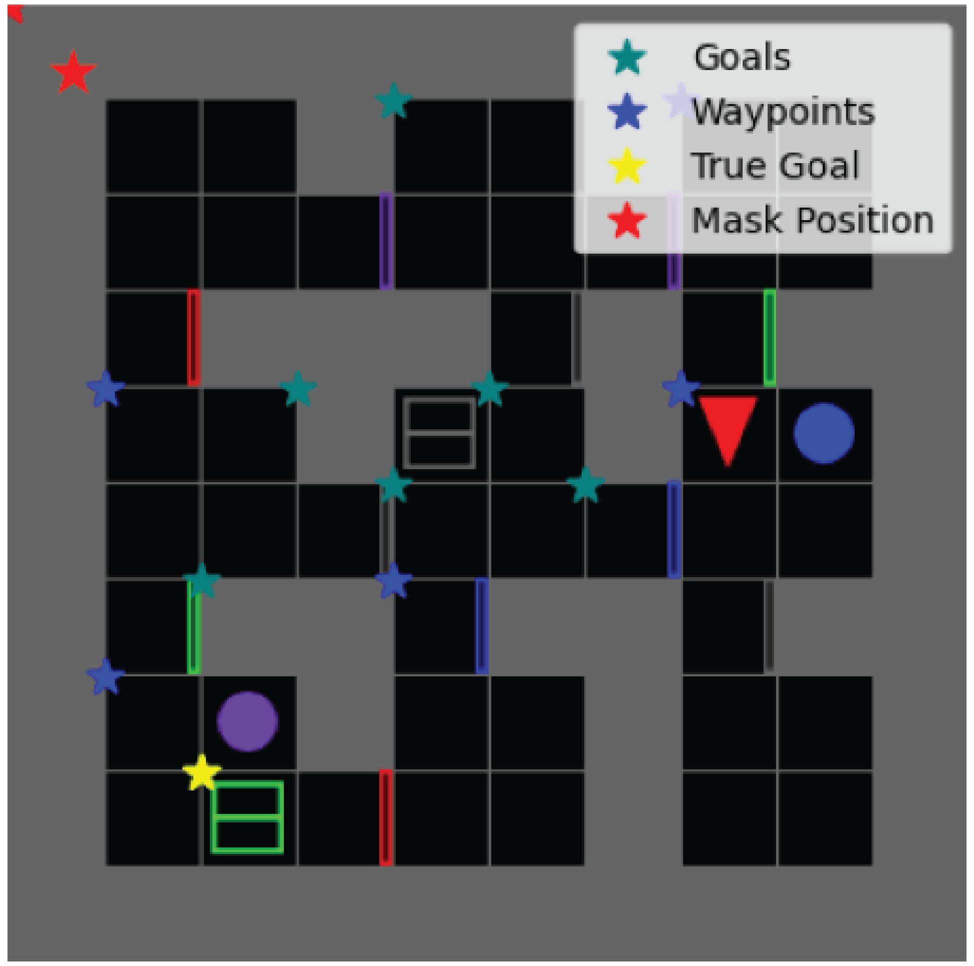
Entropy Lower-bound

Algorithm 2: GenPlan Sampling

1: **init** $\tau^0 \sim p_0$, choice of $R_t(\tau^t, \cdot|\tau^1)$, $\Delta t = \frac{1}{I_{max}}$, get $\boldsymbol{o}$

C. Rollout

Goal Generation



C. Rollout

Algorithm 2: GenPlan Sampling

init $\tau^0 \sim p_0$, choice of $R_t(\tau^t, \cdot | \tau^1)$, $\Delta t = \frac{1}{I_{max}}$, get $\boldsymbol{o}$

for $t \in \{0, \Delta t, 2\Delta t, \ldots, 1\}$ **do**

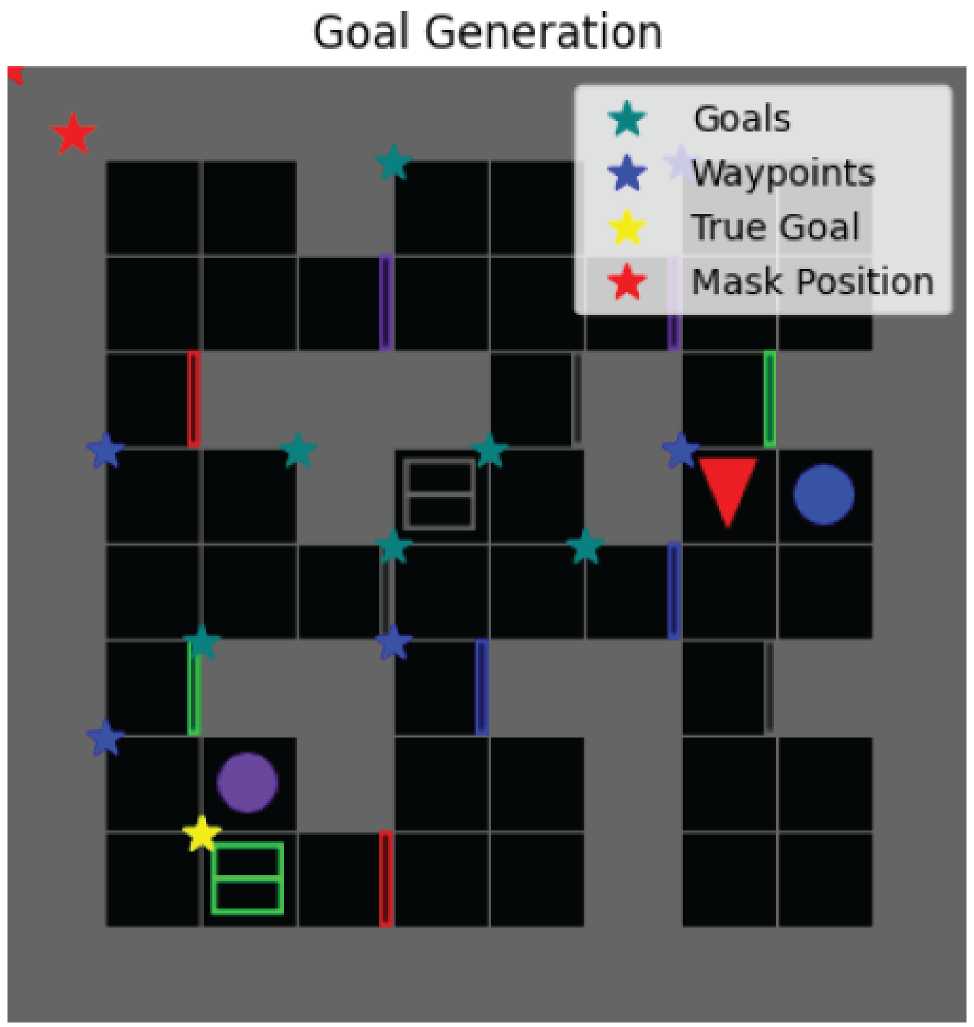
$\quad R_t^\theta(\tau^t, \cdot) \leftarrow \mathbb{E}_{p^\theta_{1|t}(\tau^1 | \tau^t, \boldsymbol{o})}\left[R_t(\tau^t, \cdot | \tau^1)\right]$

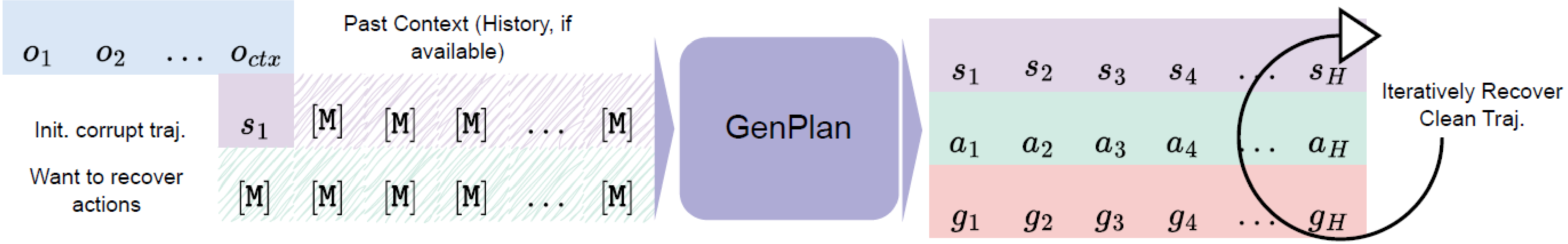
$\quad \tau^{t+\Delta t} \sim \mathcal{C}\left(\delta\{\tau^t, \tau^{t+\Delta t}\} + R_t^\theta(\tau^t, \tau^{t+\Delta t})\Delta t\right)$

$\quad t \leftarrow t + \Delta t$

end for

return $\boldsymbol{a}, \mathrm{s}, \boldsymbol{g}$ `// extract from` τ^1





04 RESULTS AND DISCUSSION

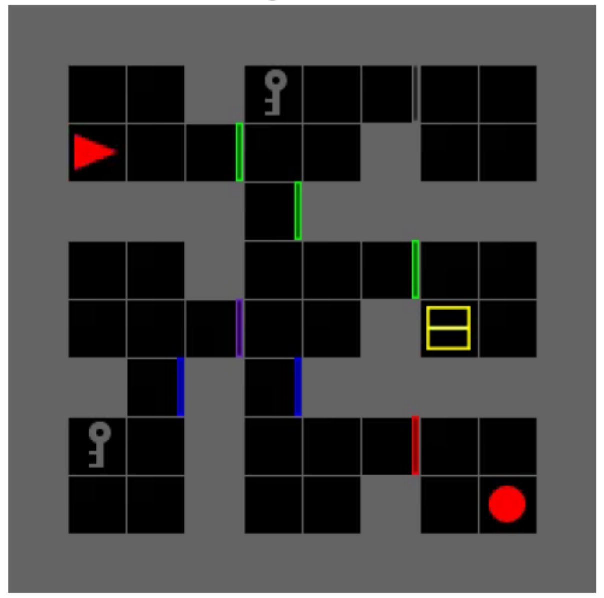
UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

04 RESULTS AND DISCUSSION

QUANTITATIVE RESULTS – SAME DIFFICULTY AS IN TRAINING (EASY)

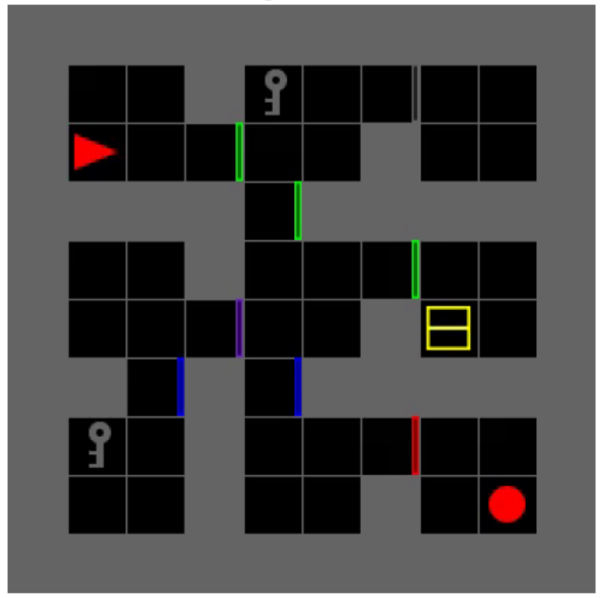
Env.	Uncond. Rollouts			Cond. Rollouts	
	GenPlan-U	GenPlan-M	LEAP⊖GC	LEAP	DT
Traj. Planning (TP)					
GoToObjMazeS4G1	52.4%	**62%**	44%	49.2%	46.8%
GoToObjMazeS4G2	38.8%	**39.6%**	20%	37.6%	35.2%
GoToObjMazeS7G1	**45.6%**	44.8%	12%	33.2%	40%
GoToObjMazeS7G2	**21.2%**	19.6%	3.6%	4%	13.6%
TP Mean (7.6 ↑)	39.5%	**41.5%**	19.9%	31%	33.9%

GoToObjMazeS4G1



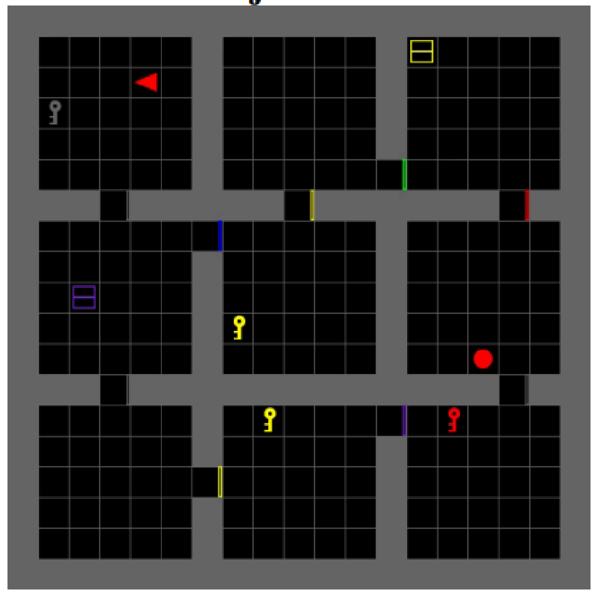
goto the yellow box

GoToObjMazeS4G2



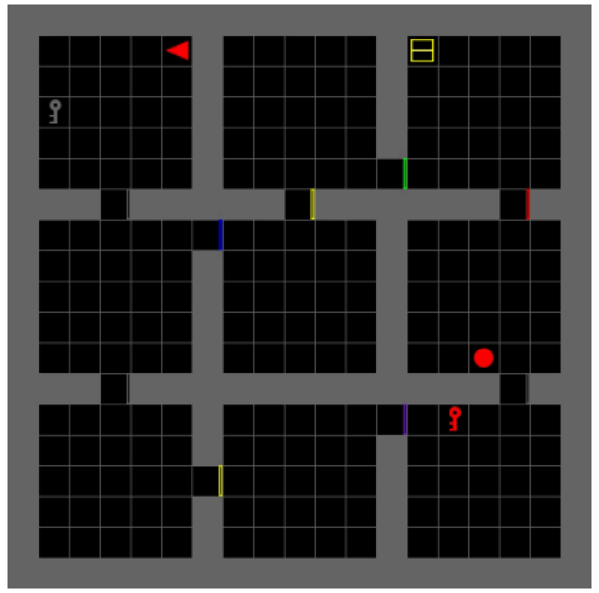
go to the yellow box and red

GoToObjMazeS7G1



go to the red key

GoToObjMazeS7G2



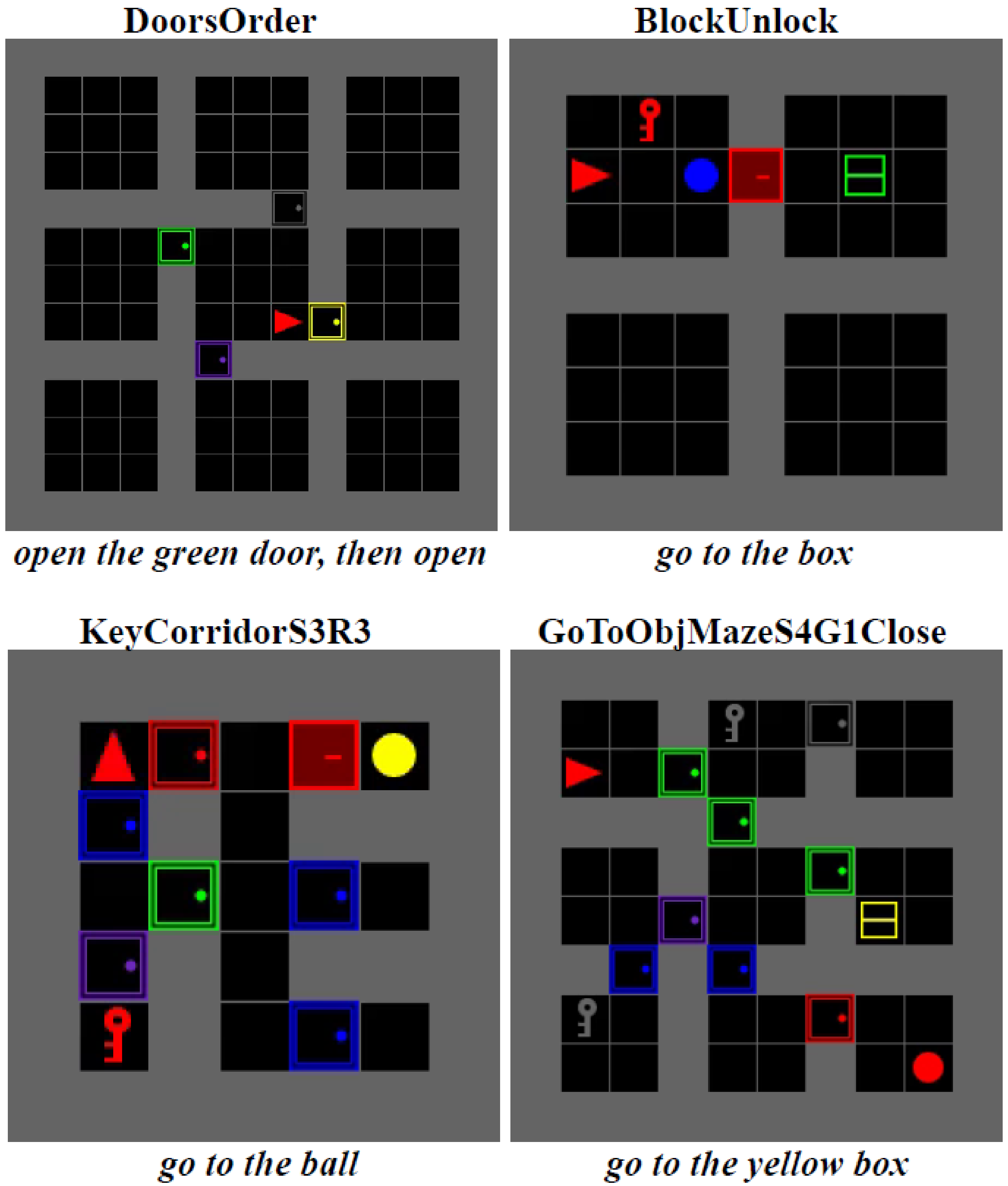
go to the red key and go to the yellow box

LEAP - Chen, H.; Du, Y.; Chen, Y.; Tenenbaum, J. B.; and Vela, P. A. Planning with Sequence Models through Iterative Energy Minimization. In ICLR 2023.
DT - Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I.Decision Transformer: Reinforcement Learning via Sequence Modeling. In Advances in Neural Information Processing Systems 2021, volume 34, 15084–15097.

04 RESULTS AND DISCUSSION

QUANTITATIVE RESULTS – INSTRUCTION COMPLETION

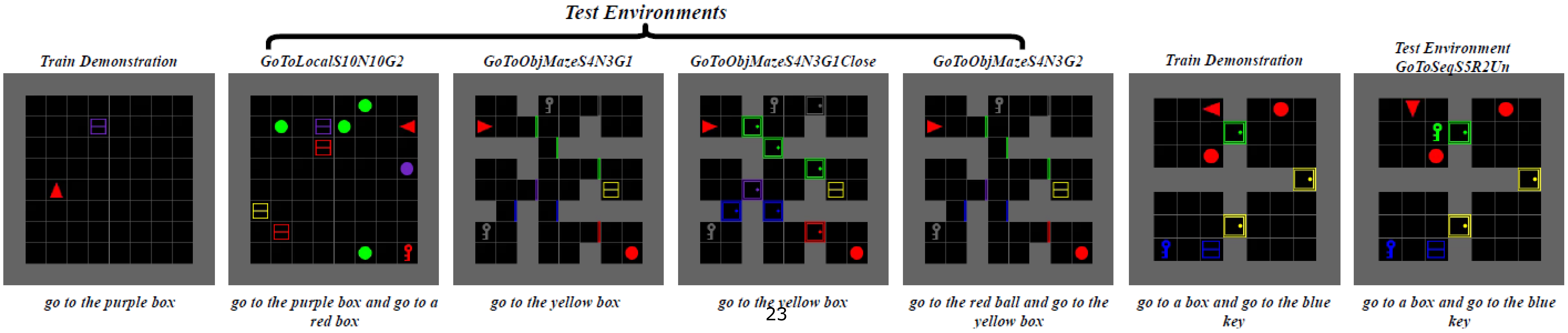
Env.	Uncond. Rollouts			Cond. Rollouts	
	GP-U	GP-M	LEAP⊖GC	LEAP	DT
Instr. Completion (IC)					
MazeClose	42.8%	**48.4%**	18%	38.8%	40%
DoorsOrder	**40.8%**	35.2%	11.2%	36.4%	**40.8%**
BlockUn	13.2%	**16%**	0%	0.8%	0%
KeyCorS3R3	11.6%	**17.6%**	0%	0.4%	3.6%
IC (8.2 ↑)	27.1%	**29.3%**	7.25%	19.1%	21.1%



04 RESULTS AND DISCUSSION

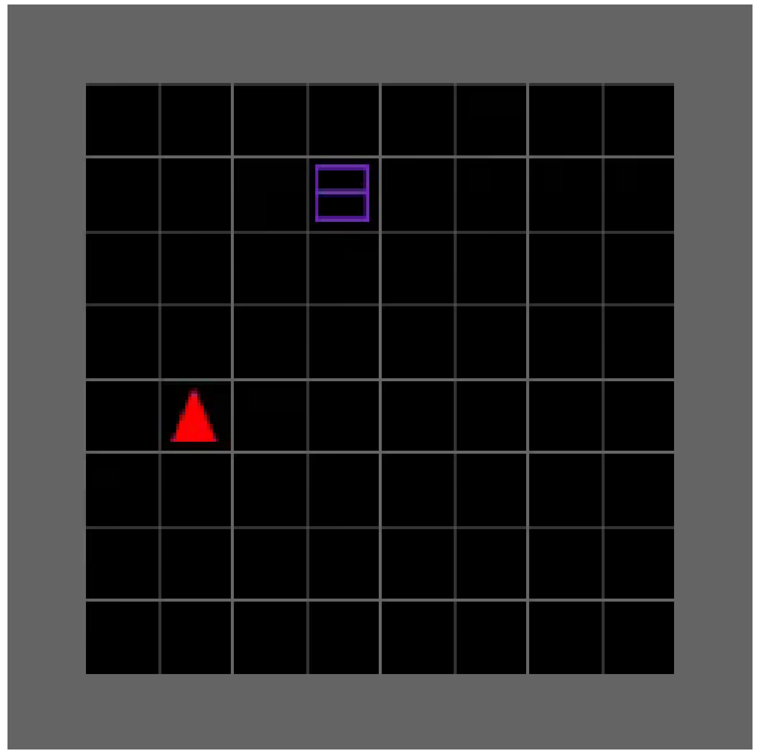
QUANTITATIVE RESULTS – ADAPTIVE PLANNING

	Unconditional Rollouts				Conditional Rollouts	
Environment	**GenPlan-U**	**GenPlan-M**	**LEAP$\ominus$GC**	**LEAP$\oplus\mathcal{H}$**	**LEAP**	**DT**
			Adaptive Planning (AP)			
GoToLocalS10N10G2	82.4%	**88%**	76%	69.2%	78%	25.6%
GoToObjMazeS4N3G1	56%	**62%**	44.8%	52%	48%	24%
GoToObjMazeClose	31.2%	**34.8%**	10%	16.4%	10%	8.8%
GoToObjMazeS4G2	28.8%	**34.8%**	14%	21.2%	18.4%	3.6%
GoToSeqS5R2Un	35.6%	**42%**	29.2%	30.8%	38%	29.2%
AP Mean (13.84 ↑)	46.8%	**52.32%**	34.8%	37.92%	38.48%	18.24%



STATE COVERAGE – ADAPTIVE PLANNING

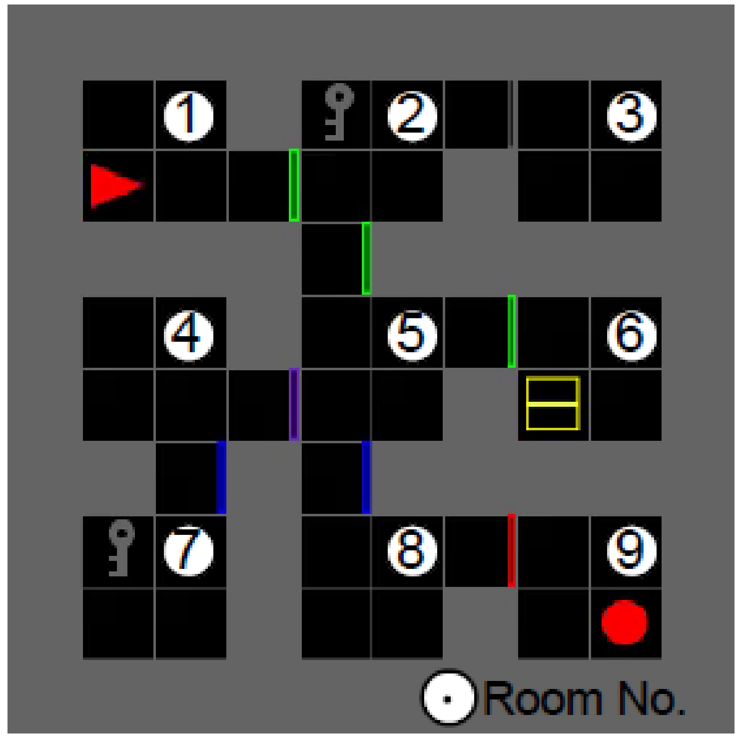
Train Demonstration



go to the purple box

We evaluate the performance in harder tasks

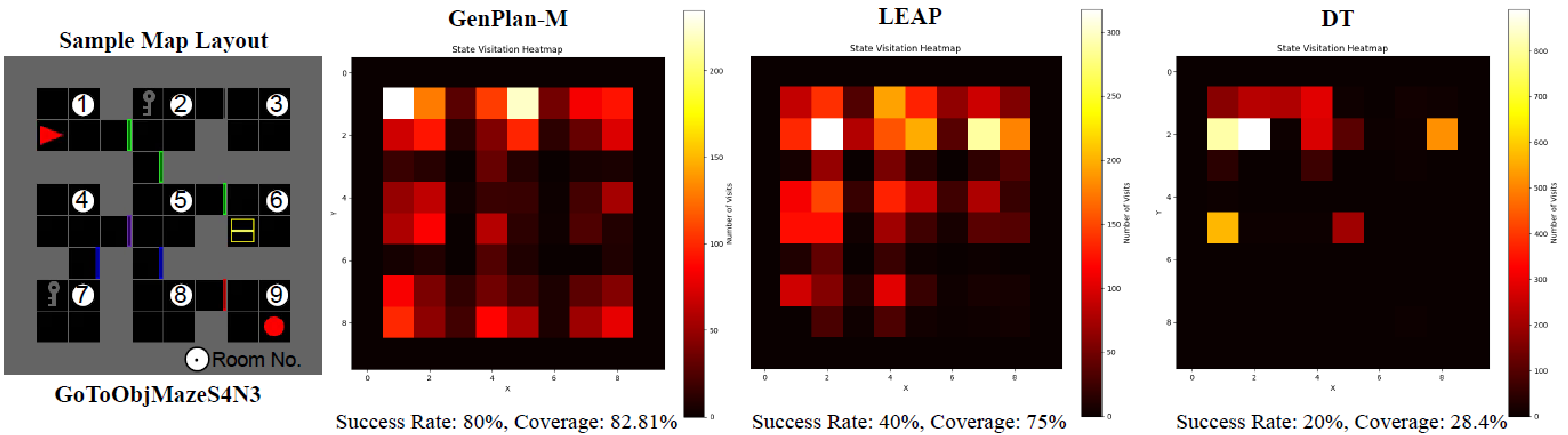
Sample Map Layout



GoToObjMazeS4N3

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

STATE COVERAGE – ADAPTIVE PLANNING

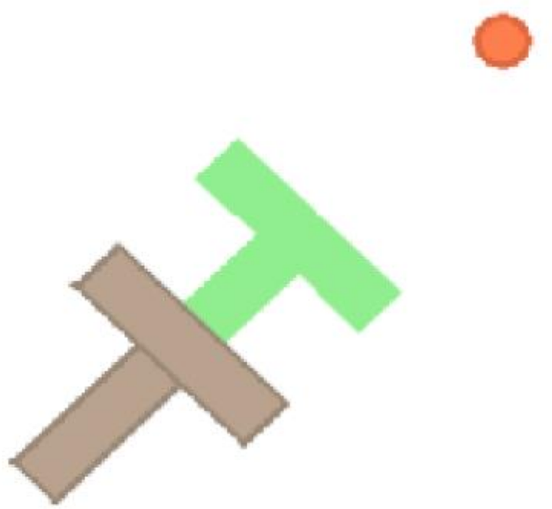


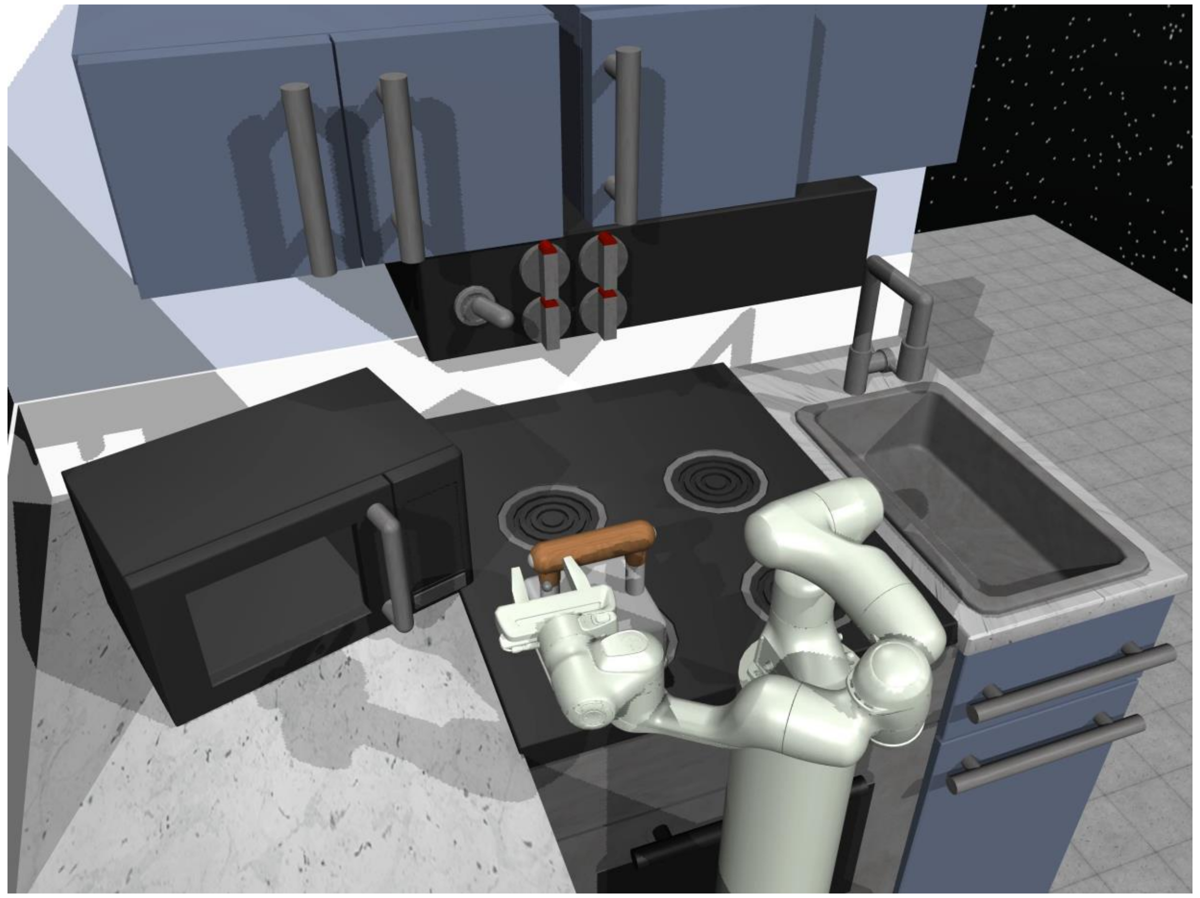
LEAP - Chen, H.; Du, Y.; Chen, Y.; Tenenbaum, J. B.; and Vela, P. A. Planning with Sequence Models through Iterative Energy Minimization. In ICLR 2023.
DT - Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I.Decision Transformer: Reinforcement Learning via Sequence Modeling. In Advances in Neural Information Processing Systems 2021, volume 34, 15084–15097.

04 RESULTS AND DISCUSSION

ADAPTATION TO CONTINUOUS TASKS

Env	Metric	GenPlan-M	VQ-BeT	DP-C	DP-T
PushT	Final Coverage	0.73	0.7	0.73	**0.74**
	Max Coverage	0.77	0.73	**0.86**	0.83
Kitchen	# Tasks	3.40	**3.66**	2.62	3.44





05 CONCLUSION

TAKEAWAYS

- We propose GenPlan, an energy-flow-based planner that learns annealed energy landscapes and uses DFM sampling to iteratively recover plans.
- Through simulation studies, we demonstrate how joint energy-based denoising improves performance in complex and long-horizon tasks.

FUTURE WORK

- In real-time scenarios, the inherent distribution tend to evolve and is dynamic. To address it, we plan to extend *GenPlan* with an online fine-tuning stage via hindsight experience replay **[1]**.
- The energy model as a denoising planner **[2]** can be extended to sample from pretrained masking model to improve sampling quality.

[1] Q. Zheng, A. Zhang, and A. Grover, "Online Decision Transformer," Jul. 13, 2022, *arXiv*: arXiv:2202.05607. doi: 10.48550/arXiv.2202.05607.
[2] S. Liu *et al.*, "Think While You Generate: Discrete Diffusion with Planned Denoising," Oct. 08, 2024, *arXiv*: arXiv:2410.06264. doi: 10.48550/arXiv.2410.06264.

Thank You for Listening!