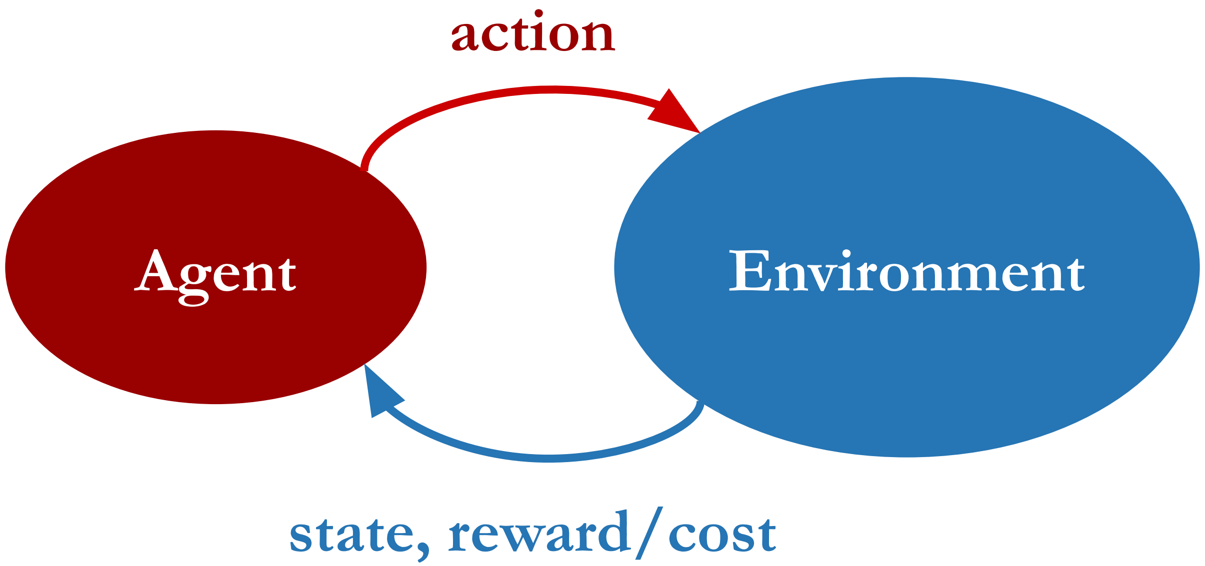
One Subgoal at a Time: Zero-Shot Generalization to Arbitrary Linear Temporal Logic Requirements in Multi-Task Reinforcement Learning

Zijian Guo[1], İlker Işık[2], H. M. Sabbir Ahmad[1], Wenchao Li[1,2]

[1]Division of Systems Engineering, Boston University
[2]Department of Electrical and Computer Engineering, Boston University
{zjguo, iilker, sabbir92, wenchao}@bu.edu

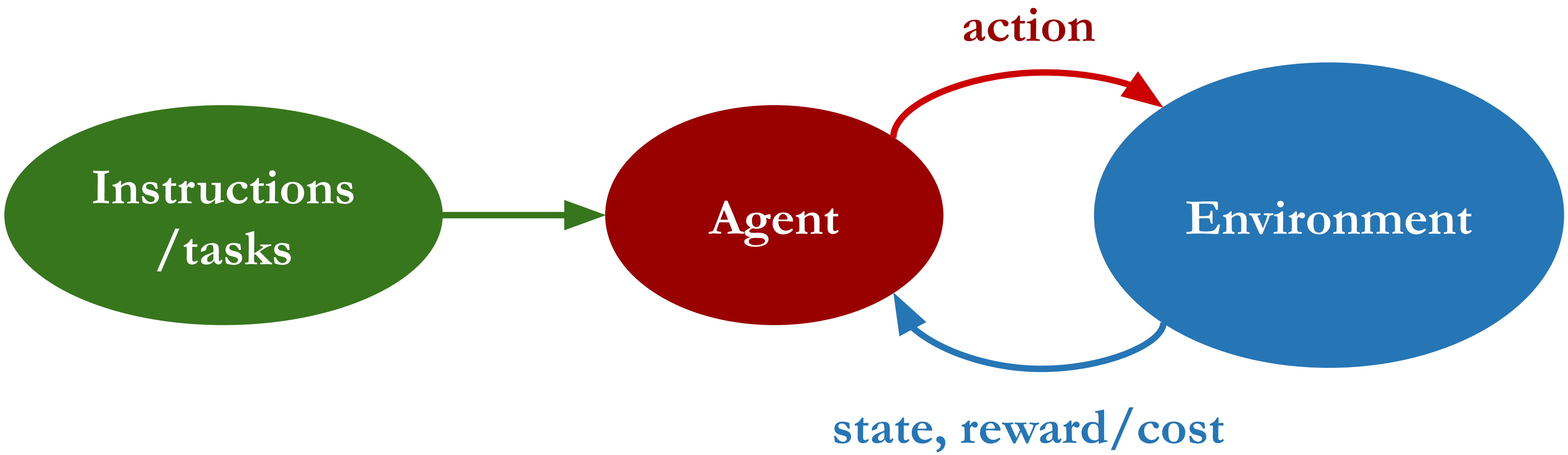
Reinforcement Learning



Reinforcement Learning

How to follow diverse, complex, and even unseen instructions/tasks?

- long-horizon goals, logical dependencies, safety constraints



BOSTON UNIVERSITY

Background: Linear Temporal Logic

- **Formal language** to specify system's behaviors

Background: Linear Temporal Logic

- **Formal language** to specify system's behaviors
- **Syntax** of LTL

$$\varphi := \boldsymbol{a} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \mathsf{F}\ \varphi \mid \mathsf{G}\ \varphi \mid \varphi_1 \mathsf{U}\ \varphi_2$$

  - Atomic propositions: $AP, \boldsymbol{a} \in AP$
  - Boolean ($\neg$, $\wedge$, $\vee$) and temporal (F, G, U) operators.

Background: Linear Temporal Logic

- **Formal language** to specify system's behaviors
- **Syntax** of LTL

$$\varphi := \boldsymbol{a} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \mathsf{F}\ \varphi \mid \mathsf{G}\ \varphi \mid \varphi_1 \mathsf{U}\ \varphi_2$$

  - Atomic propositions: $AP, \boldsymbol{a} \in AP$
  - Boolean ($\neg$, $\wedge$, $\vee$) and temporal (F, G, U) operators.
- **Tasks can be expressed over high-level environment features**

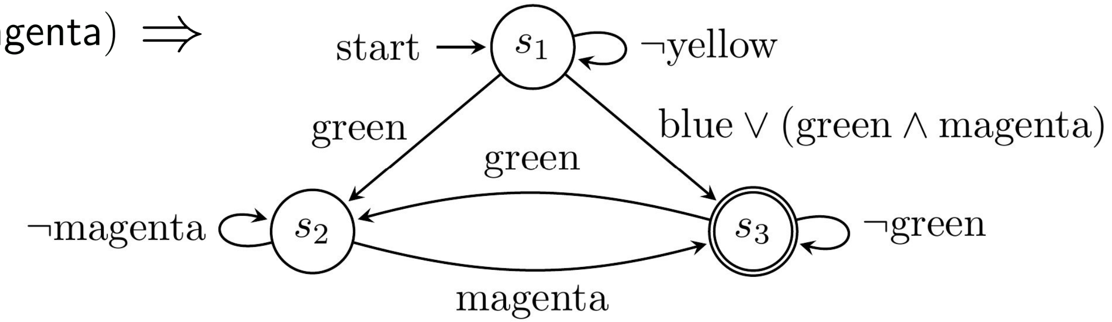
$$\varphi := (\neg\mathsf{yellow}\ \mathsf{U}\ (\mathsf{green} \vee \mathsf{blue})) \wedge \mathsf{G}(\mathsf{green} \Rightarrow \mathsf{F}\,\mathsf{magenta})$$

(avoid yellow until reaching green or blue; whenever green is visited, magenta must eventually follow.)

Background: Linear Temporal Logic

- **Büchi automata (BA):** for any LTL formula, it can be converted to an equivalent BA, which can be represented as **directed state-transition graphs**.

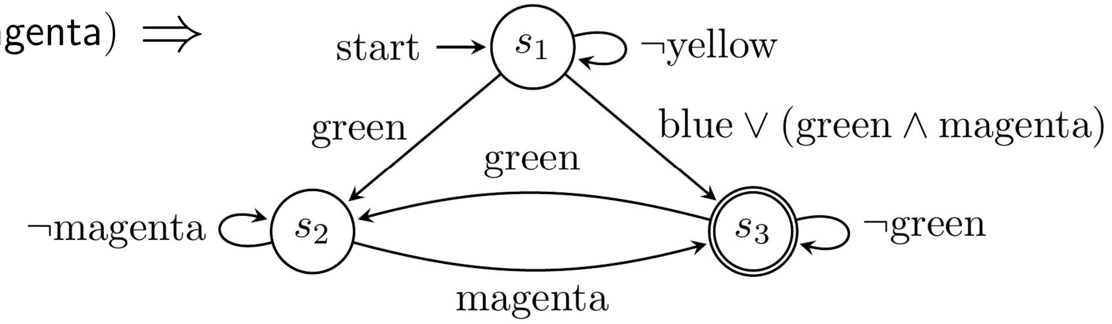
$\varphi := (\neg\mathsf{yellow}\ \mathsf{U}\ (\mathsf{green} \vee \mathsf{blue})) \wedge \mathsf{G}(\mathsf{green} \Rightarrow \mathsf{F}\,\mathsf{magenta}) \implies$



Background: Linear Temporal Logic

- **Büchi automata (BA):** for any LTL formula, it can be converted to an equivalent BA, which can be represented as **directed state-transition graphs**.

$\varphi := (\neg\mathsf{yellow}\ \mathsf{U}\ (\mathsf{green} \vee \mathsf{blue})) \wedge \mathsf{G}(\mathsf{green} \Rightarrow \mathsf{F}\,\mathsf{magenta}) \implies$



- **Reach-Avoid Subgoal Construction:** depth-first search (DFS) to enumerate all possible paths and extract reach-avoid subgoals (α^+, A^-)

$$p_1 = \{(\alpha^+ = \{\mathsf{green}\},\ A^- = \{\mathsf{yellow}\}), (\alpha^+ = \{\mathsf{magenta}\},\ A^- = \emptyset)\}$$
$$p_2 = \{(\alpha^+ = \{\mathsf{blue}\},\ A^- = \{\mathsf{yellow}\})\}$$

Challenges of Generalization

- Satisfying an LTL formula = completing a sequence of reach-avoid subgoals

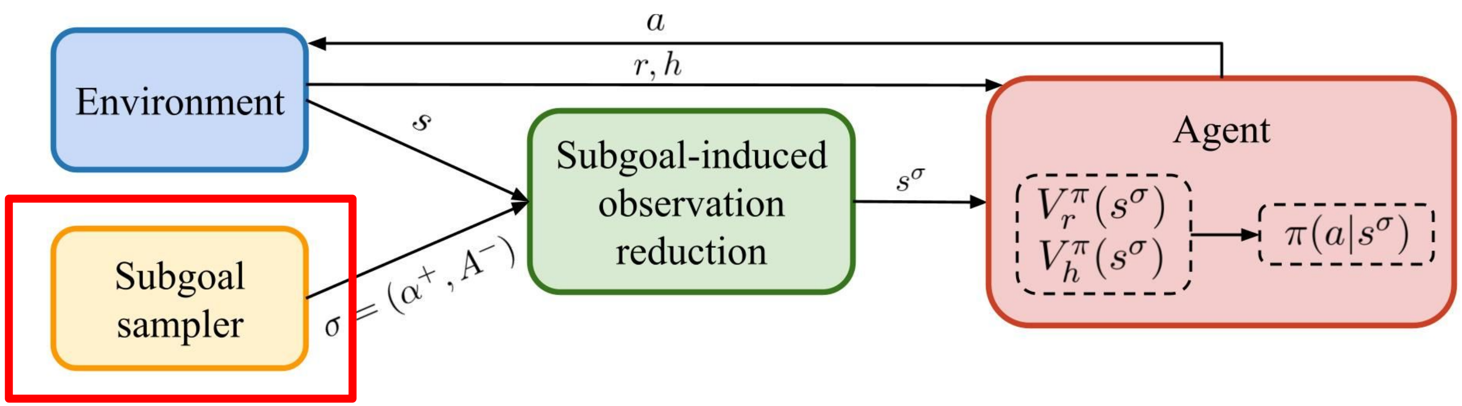
Challenges of Generalization

- Satisfying an LTL formula = completing a sequence of reach-avoid subgoals
- Existing methods:
    - Structure of the automaton/entire subgoal sequence
    - Policy conditioned on those representations
    - **Limitation: Out-of-distribution (OOD) issue** of new LTL formulas at test time

Challenges of Generalization

- Satisfying an LTL formula = completing a sequence of reach-avoid subgoals
- Existing methods:
    - Structure of the automaton/entire subgoal sequence
    - Policy conditioned on those representations
    - **Limitation: Out-of-distribution (OOD) issue** of new LTL formulas at test time
- In contrast, we address this problem by solving **one subgoal at a time**

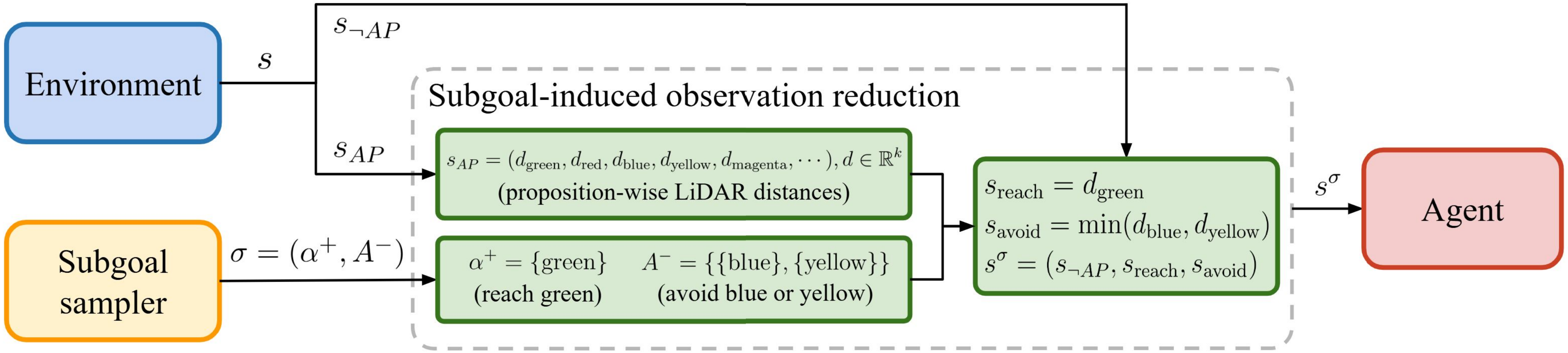
GenZ-LTL: Training



- **Subgoal sampling:** all possible subgoals $\xi = \{(\alpha^+, A^-)_i\}_{i=1}^M$
  - Enumerate each assignment $\boldsymbol{a} \in 2^{AP}$ as a candidate α^+
  - For each such α^+, we filter out the remaining assignments that conflict with it. We then enumerate all possible combinations of the filtered assignments to form A^-

GenZ-LTL: Training

- **Subgoal-Induced observation reduction**
  - Note that $\boldsymbol{a} \in 2^{AP}$, so the total number of subgoal grows exponentially
  - Idea: focus only on subgoal-relevant observations to reduce sample complexity



GenZ-LTL: Training

- **Policy learning with reachability constraints**

$$\pi_{k+1} = \arg\max_{\pi} \mathbb{E}_{\sigma\sim\mathrm{Unif}(\xi), s\sim d^{\pi_k}, a\sim\pi_k}\left[\frac{\pi}{\pi_k} A_r^{\pi_k}(s^\sigma, a)\right]$$

$$\text{s.t.} \quad \mathbb{E}_{\sigma\sim\mathrm{Unif}(\xi), s\sim d^{\pi_k}}\left[\mathcal{D}_{KL}(\pi, \pi_k)\right] \leq \epsilon$$

$$\mathbb{E}_{\sigma\sim\mathrm{Unif}(\xi), s\sim d^{\pi_k}, a\sim\pi_k}\left[(1-\gamma)J_h(\pi_k) + \frac{\pi}{\pi_k} A_h^{\pi_k}(s^\sigma, a)\right] \leq 0$$

where $h : \mathcal{S} \mapsto \mathbb{R}$ is the constraint violation function

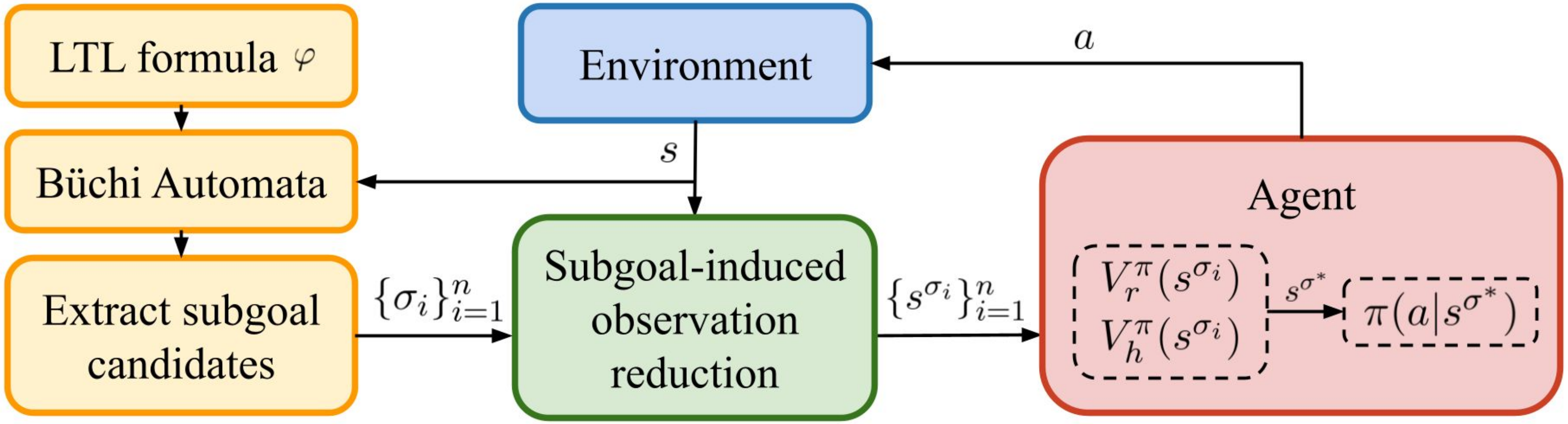
$$A_h^\pi(s^\sigma, a) := Q_h^\pi(s^\sigma, a) - V_h^\pi(s^\sigma)$$

$$Q_h^\pi(s^\sigma, a) := \max_{t\in\mathbb{N}} h(s_t^\sigma), s_0 = s^\sigma, a_0 = a, a_t \sim \pi$$

$$J_h(\pi) := \max_{t\in\mathbb{N}} h(\cdot), a \sim \pi$$

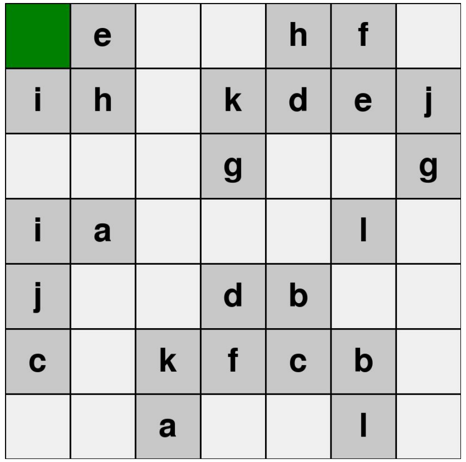
GenZ-LTL: Testing

- Given a target LTL specification, we construct the corresponding BA and identify candidate subgoals based on the current automaton state. The subgoal to be executed is selected as $\sigma^* = \arg\max_\sigma V_r(s^\sigma) - \lambda(s^\sigma)V_h(s^\sigma)$



Experimental Settings

- **Environments**
  - LetterWorld: 7×7 grid world
  - ZoneEnv: high-dimensional env with lidar observations
  - Randomized environments



Experimental Settings

- **LTL specifications**

	LetterWorld		ZoneEnv	
Finite-horizon	φ_1	$\mathsf{F}\,(a \wedge (\neg b\ \mathsf{U}\ c)) \wedge \mathsf{F}\,d$	φ_9	$(\mathsf{F}\,b) \wedge (\neg b\ \mathsf{U}\ (g \wedge \mathsf{F}\,y))$
	φ_2	$(\mathsf{F}\,d) \wedge (\neg f\ \mathsf{U}\ (d \wedge \mathsf{F}\,b))$	φ_{10}	$\neg(m \vee y)\ \mathsf{U}\ (b \wedge \mathsf{F}\,g)$
	φ_3	$\neg a\ \mathsf{U}\ (b \wedge (\neg c\ \mathsf{U}\ (d \wedge (\neg e\ \mathsf{U}\ f))))$	φ_{11}	$\neg g\ \mathsf{U}\ ((b \vee m) \wedge (\neg g\ \mathsf{U}\ y))$
	φ_4	$(a \vee b \vee c \vee d \Rightarrow \mathsf{F}\,(e \wedge \mathsf{F}\,(f \wedge \mathsf{F}\,g)))\ \mathsf{U}\ (h \wedge \mathsf{F}\,i)$	φ_{12}	$(g \vee b \Rightarrow (\neg y\ \mathsf{U}\ m))\ \mathsf{U}\ y$
	φ_5	$\mathsf{F}\,(d \wedge (\neg(a \vee b)\ \mathsf{U}\ (b \wedge (\neg e\ \mathsf{U}\ c)))) \wedge \mathsf{F}\,(\neg(f \vee g \vee h)\ \mathsf{U}\ a)$	φ_{13}	$\mathsf{F}\,(g \wedge (\neg(b \vee y)\ \mathsf{U}\ (y \wedge (\neg m\ \mathsf{U}\ b)))) \wedge \mathsf{F}\,(\neg g\ \mathsf{U}\ y)$
	φ_6	$\mathsf{F}\,((k \wedge ((\neg b \vee c)\ \mathsf{U}\ f)) \wedge (\neg(a \vee e \vee h)\ \mathsf{U}\ g)) \wedge \mathsf{F}\,d$	φ_{14}	$\mathsf{F}\,((b \vee g) \wedge (\neg y\ \mathsf{U}\ (b \wedge (\neg(g \vee m)\ \mathsf{U}\ m)))) \wedge \mathsf{F}\,(y \wedge (\neg b\ \mathsf{U}\ g))$
	φ_7	$\neg(j \vee b \vee d)\ \mathsf{U}\ (a \wedge (\neg c\ \mathsf{U}\ (f \wedge \mathsf{F}\,(g \wedge (\neg d\ \mathsf{U}\ e)))))$	φ_{15}	$\neg(m \vee y)\ \mathsf{U}\ (b \wedge (\neg g\ \mathsf{U}\ (y \wedge \mathsf{F}\,(g \wedge (\neg b\ \mathsf{U}\ m)))))$
	φ_8	$\neg(f \vee g)\ \mathsf{U}\ (a \wedge (\neg b\ \mathsf{U}\ c) \wedge \mathsf{F}\,(d \wedge (\neg e\ \mathsf{U}\ f)))$	φ_{16}	$\mathsf{F}\,(b \wedge (\neg y\ \mathsf{U}\ (g \wedge \mathsf{F}\,(y \wedge (\neg(m \vee g)\ \mathsf{U}\ b)))))$
Infinite-horizon	ψ_1	$\mathsf{G}\,\mathsf{F}\,(e \wedge (\neg a\ \mathsf{U}\ f)) \wedge \mathsf{G}\,\neg(c \vee d)$	ψ_4	$\mathsf{G}\,\mathsf{F}\,b \wedge \mathsf{G}\,\mathsf{F}\,g \wedge \mathsf{G}\,\neg(y \vee m)$
	ψ_2	$\mathsf{G}\,\mathsf{F}\,a \wedge \mathsf{G}\,\mathsf{F}\,b \wedge \mathsf{G}\,\mathsf{F}\,c \wedge \mathsf{G}\,\neg(e \vee f \vee i)$	ψ_5	$\mathsf{G}\,\mathsf{F}\,b \wedge \mathsf{G}\,\mathsf{F}\,y \wedge \mathsf{G}\,\mathsf{F}\,g \wedge \mathsf{G}\,\neg m$
	ψ_3	$\mathsf{G}\,\mathsf{F}\,c \wedge \mathsf{G}\,\mathsf{F}\,a \wedge \mathsf{G}\,\mathsf{F}\,(e \wedge (\neg f\ \mathsf{U}\ g)) \wedge \mathsf{G}\,\mathsf{F}\,k \wedge \mathsf{G}\,\neg(i \vee j)$	ψ_6	$\mathsf{F}\,\mathsf{G}\,y \wedge \mathsf{G}\,\neg(g \vee b \vee m)$

- **All LTL specifications are unseen at test time for our method**

Main Results

- **GenZ-LTL achieves higher success and lower violation rates, while learning more efficient policies**

		LTL2Action			GCRL-LTL			DeepLTL			RAD-embeddings			GenZ-LTL(ours)		
		$\eta_s \uparrow$	$\eta_v \downarrow$	$\mu \downarrow$	$\eta_s \uparrow$	$\eta_v \downarrow$	$\mu \downarrow$	$\eta_s \uparrow$	$\eta_v \downarrow$	$\mu \downarrow$	$\eta_s \uparrow$	$\eta_v \downarrow$	$\mu \downarrow$	$\eta_s \uparrow$	$\eta_v \downarrow$	$\mu \downarrow$
Letter	φ_{1-4}	$0.62_{\pm 0.16}$	$0.07_{\pm 0.09}$	$26.64_{\pm 5.87}$	$0.87_{\pm 0.11}$	$0.03_{\pm 0.05}$	$16.05_{\pm 6.13}$	$0.87_{\pm 0.04}$	$\mathbf{0.00_{\pm 0.01}}$	$7.51_{\pm 1.21}$	$0.90_{\pm 0.06}$	$0.04_{\pm 0.04}$	$17.79_{\pm 3.37}$	$\mathbf{0.98_{\pm 0.02}}$	$\mathbf{0.00_{\pm 0.00}}$	$\mathbf{7.22_{\pm 1.18}}$
	φ_{5-8}	$0.24_{\pm 0.12}$	$0.20_{\pm 0.25}$	$36.43_{\pm 6.08}$	$0.65_{\pm 0.08}$	$0.11_{\pm 0.06}$	$18.71_{\pm 2.54}$	$0.76_{\pm 0.05}$	$0.01_{\pm 0.02}$	$10.62_{\pm 1.47}$	$0.82_{\pm 0.10}$	$0.07_{\pm 0.08}$	$24.29_{\pm 3.77}$	$\mathbf{0.95_{\pm 0.03}}$	$\mathbf{0.00_{\pm 0.00}}$	$\mathbf{9.82_{\pm 1.42}}$
Zone	φ_{9-12}	$0.57_{\pm 0.37}$	$0.17_{\pm 0.21}$	$331.21_{\pm 165.88}$	$0.88_{\pm 0.04}$	$0.05_{\pm 0.02}$	$305.28_{\pm 123.33}$	$0.91_{\pm 0.05}$	$0.04_{\pm 0.03}$	$\mathbf{220.39_{\pm 78.77}}$	$0.94_{\pm 0.05}$	$0.04_{\pm 0.05}$	$269.53_{\pm 129.95}$	$\mathbf{0.99_{\pm 0.01}}$	$\mathbf{0.01_{\pm 0.01}}$	$254.69_{\pm 89.18}$
	φ_{13-16}	$0.12_{\pm 0.18}$	$0.17_{\pm 0.28}$	$886.39_{\pm 288.60}$	$0.70_{\pm 0.07}$	$0.09_{\pm 0.04}$	$606.42_{\pm 26.76}$	$0.87_{\pm 0.08}$	$0.03_{\pm 0.06}$	$505.17_{\pm 54.03}$	$0.70_{\pm 0.15}$	$\mathbf{0.00_{\pm 0.01}}$	$647.39_{\pm 40.74}$	$\mathbf{0.98_{\pm 0.02}}$	$0.01_{\pm 0.01}$	$\mathbf{408.71_{\pm 27.47}}$

Finite-horizon tasks

Methods	Metrics	LetterWorld			ZoneEnv		
		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
GenZ-LTL(ours)	$\mu_{acc} \uparrow$	$\mathbf{208.95_{\pm 14.39}}$	$\mathbf{102.12_{\pm 7.02}}$	$\mathbf{55.17_{\pm 1.08}}$	$\mathbf{55.16_{\pm 4.23}}$	$\mathbf{32.75_{\pm 1.20}}$	$\mathbf{8135.67_{\pm 1489.99}}$
	$\eta_v \downarrow$	$\mathbf{0.00_{\pm 0.00}}$	$\mathbf{0.00_{\pm 0.00}}$	$\mathbf{0.00_{\pm 0.00}}$	$\mathbf{0.07_{\pm 0.02}}$	$\mathbf{0.03_{\pm 0.01}}$	$\mathbf{0.03_{\pm 0.02}}$
DeepLTL	$\mu_{acc} \uparrow$	$142.56_{\pm 22.44}$	$48.28_{\pm 12.37}$	$19.21_{\pm 4.57}$	$30.03_{\pm 13.23}$	$15.73_{\pm 4.44}$	$7337.38_{\pm 2019.56}$
	$\eta_v \downarrow$	$0.04_{\pm 0.02}$	$0.09_{\pm 0.01}$	$0.09_{\pm 0.04}$	$0.39_{\pm 0.10}$	$0.38_{\pm 0.24}$	$0.13_{\pm 0.05}$
GCRL-LTL	$\mu_{acc} \uparrow$	$41.98_{\pm 15.80}$	$22.77_{\pm 9.50}$	$9.53_{\pm 2.28}$	$30.00_{\pm 3.72}$	$14.61_{\pm 1.62}$	$5584.34_{\pm 3180.15}$
	$\eta_v \downarrow$	$0.18_{\pm 0.08}$	$0.30_{\pm 0.08}$	$0.30_{\pm 0.18}$	$0.37_{\pm 0.08}$	$0.40_{\pm 0.08}$	$0.14_{\pm 0.01}$

Infinite-horizon tasks

Please scan the following QR codes for more details

Paper

Code

Project page