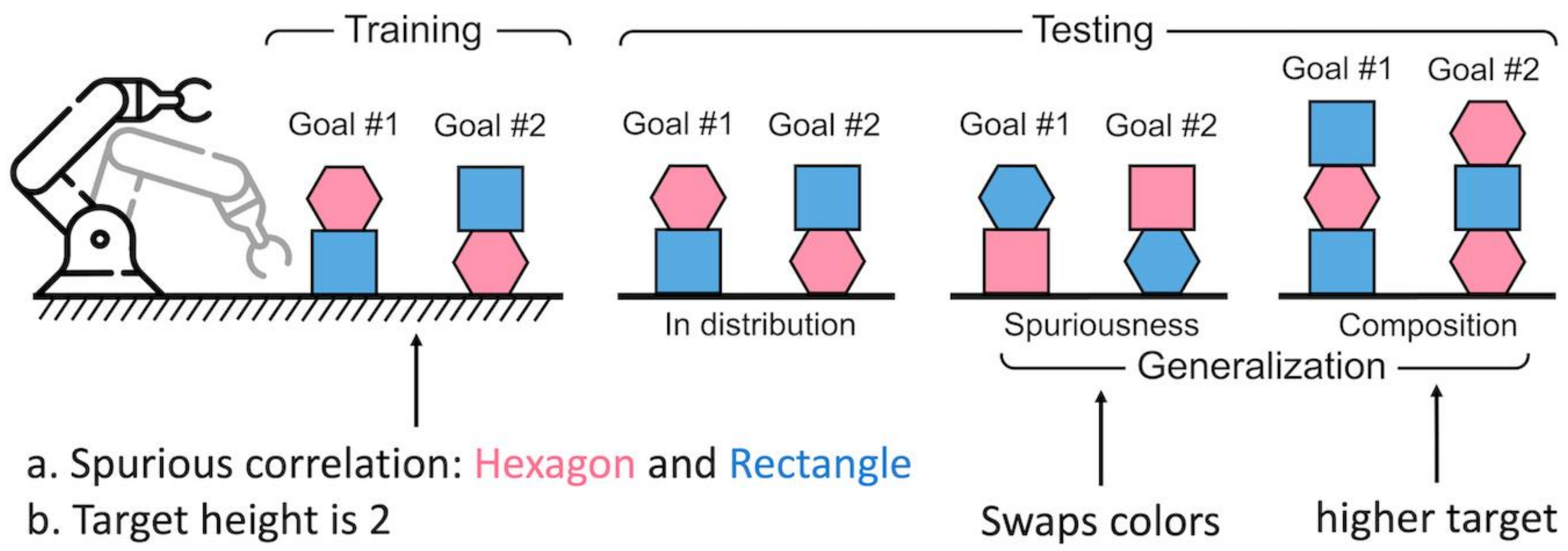
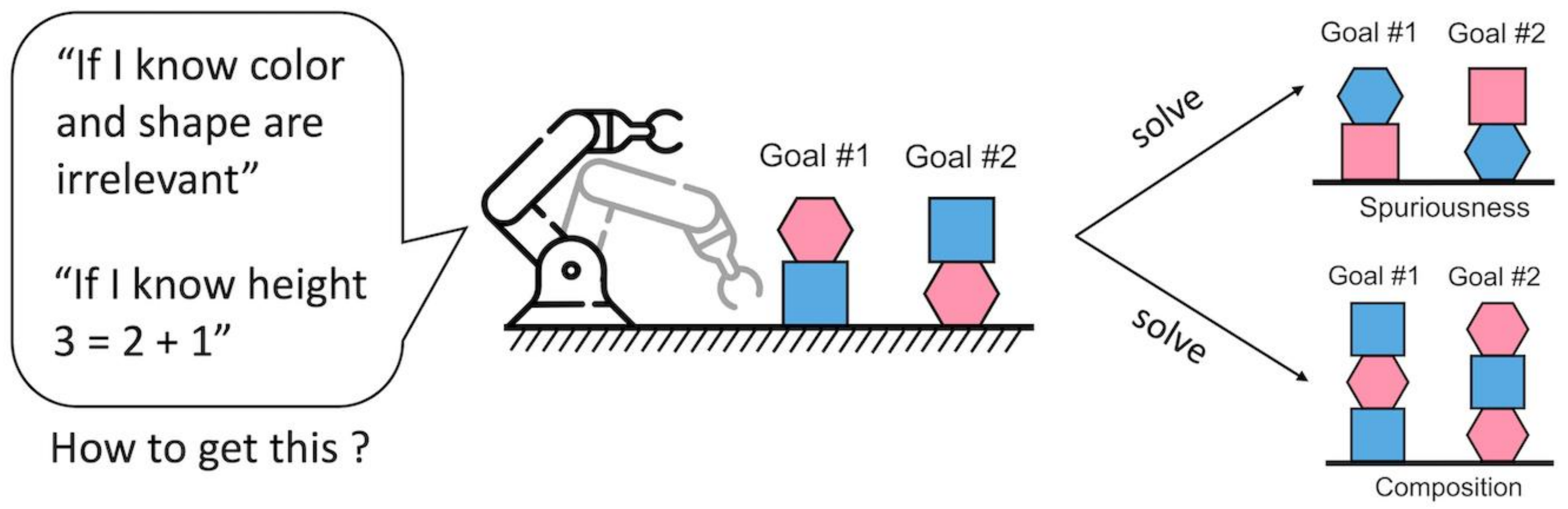
Generalizing Goal-Conditioned Reinforcement Learning with Variational Causal Reasoning (GRADER)

Wenhao Ding, Haohong Lin, Bo Li, Ding Zhao Contact: wenhaod@andrew.cmu.edu

Motivation





- We consider two generalization cases in goal-conditioned (GC) reinforcement learning framework: **spuriousness** and **composition**.
- We improve generalization by discovering explicit **causality**.

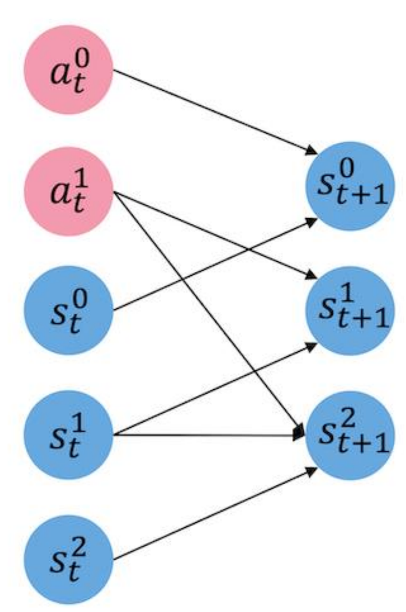
Representation of Causality

Assumption 1 (Space Factorization): The state space and action space can be factorized to disjoint components, e.g., objects and events.

Assumption 2 (Causal Sufficiency): All confounders are measured in the representation.

We use **Structural Causal Models** (SCM)

$$X_j := f_j(PA_j^{\mathcal{G}}, U_j)$$



- Parent of j: $PA_j \subset \{X_1, \ldots, X_d\} \backslash \{X_j\}$
- Random noise: $U = \{U_1, \ldots, U_d\}$

where $\mathcal{G}$ is Causal Graph, a Directed Acyclic Graphs

- Match transition model in MDP : edges point from t to t+1
- Nodes represent factorized actions a_t^i or states s_t^i
- Parents are the causes of children

Proposed Method (GRADER)

1. Formulating Goal-conditioned Reinforcement Learning

Traditional GCRL: "How to find actions to achieve the goal?"

Our formulation: "What are the actions if we achieved the goal?"

Trajectory: $\tau := \{s^0, a^0, \ldots, s^T\}$

Goal state: $s^* := \mathbb{1}(g = s^T)$

Prior of causal graph

$$\log p(\tau|s^*) = \log \int p(\tau|\mathcal{G}, s^*) p(\mathcal{G}|s^*) d\mathcal{G}$$

$$\geq \mathbb{E}_{q(\mathcal{G}|\tau)}[\log p(\tau|\mathcal{G}, s^*)] - \mathbb{D}_{\mathrm{KL}}[q(\mathcal{G}|\tau)||p(\mathcal{G})] \quad \text{(ELBO)}$$

2. Components of ELBO

Graph regularization

$$\log p(\tau|s^*) \geq \mathbb{E}_{q(\mathcal{G}|\tau)}[\log p(\tau|\mathcal{G}, s^*)] - \mathbb{D}_{\mathrm{KL}}[q(\mathcal{G}|\tau)||p(\mathcal{G})]$$

$$\underbrace{\log p(s^0)}_{\text{Constant}} + \underbrace{\sum_{t=0}^{T-1} \log p(s^{t+1}|s^t, a^t, \mathcal{G})}_{\text{Causal world model}} + \underbrace{\sum_{t=0}^{T-1} \log \pi(a^t|s^t, s^*, \mathcal{G})}_{\text{Causal policy model}} + \underbrace{\log p(g)}_{\text{Constant}}$$

3. Model Parametrization

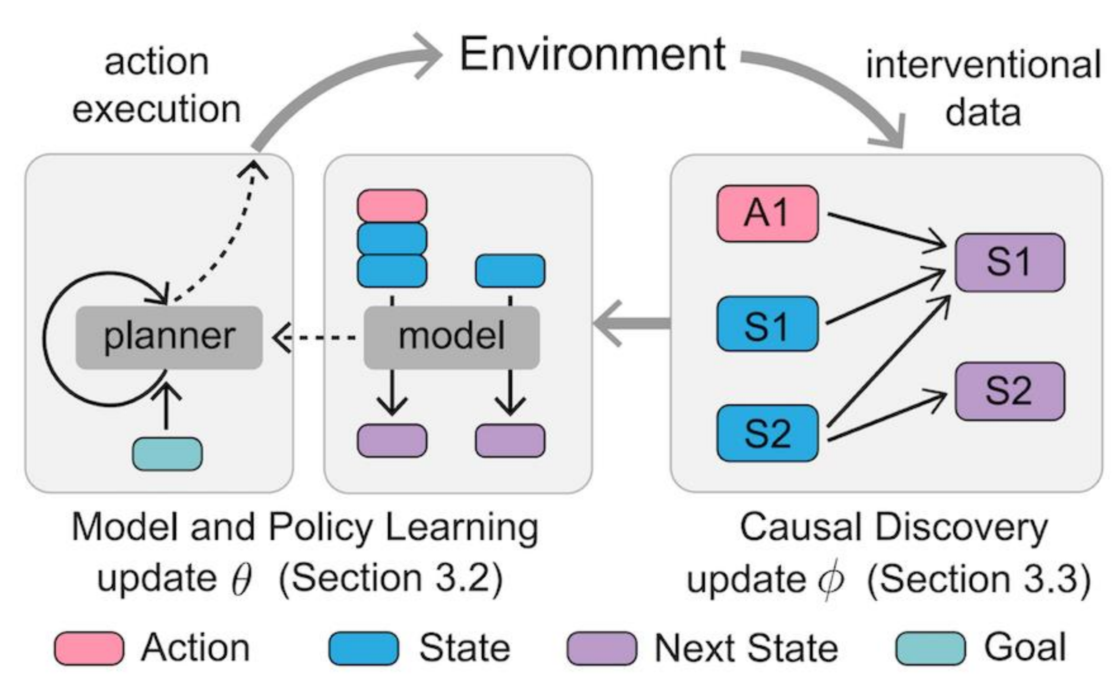
$$\mathcal{J}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathcal{G}|\tau)} \sum_{t=0}^{T-1} [\log p_\theta(s^{t+1}|s^t, a^t, \mathcal{G}) + \log \pi_\theta(a^t|s^t, s^*, \mathcal{G})] - \mathbb{D}_{\mathrm{KL}}[q_\phi(\mathcal{G}|\tau)||p(\mathcal{G})]$$

Parameters of model and policy (p_θ, π_θ)

Parameters of structural causal model (q_ϕ)

- Use two neural networks θ and ϕ to learn policy and causal model
- Iterative update them with convergence guarantee

4. Training iteration

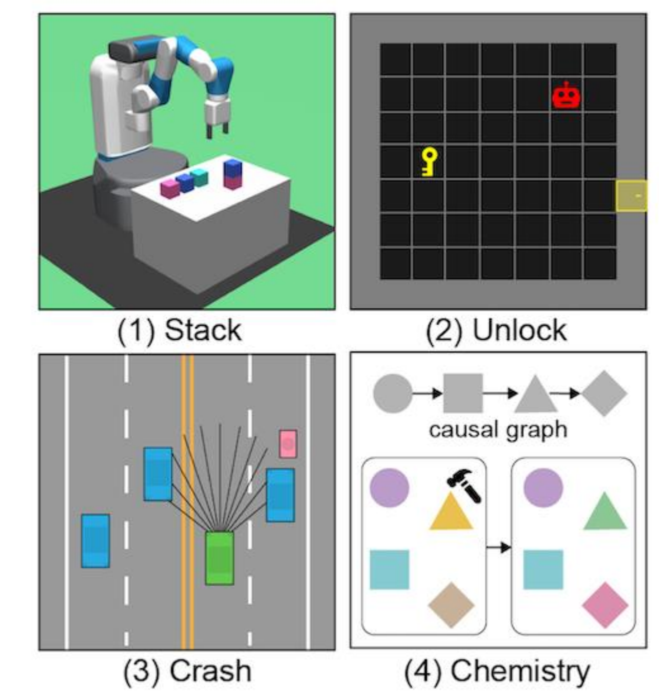


Algorithm 1: GRADER Training

Input: Trajectory buffer $\mathcal{B}_\tau$, Causal graph $\mathcal{G}$, Transition model f_θ, causal discovery threshold η

while θ *not converged* **do**
- // Policy from planning
- Sample a goal $g \sim p_{\text{train}}(g)$
- **while** $t < T$ **do**
  - $a^t \leftarrow \text{Planner}(f_\theta, s^t, g)$
  - $s^{t+1}, r^t \leftarrow \text{Env}(a^t, g)$
  - $\mathcal{B}_\tau \leftarrow \mathcal{B}_\tau \cup \{a^t, s^t, s^{t+1}\}$
- // Estimate causal graph
- **for** $i \leq M + N$ **do**
  - **for** $j \leq M$ **do**
    - Infer edge $e_{ij} \leftarrow q_\phi(\cdot|\mathcal{B}, \eta)$
- // Learn transition model
- Update $f_\theta(\mathcal{G})$ via (4) with $\mathcal{B}$

Environment and Causal Graph

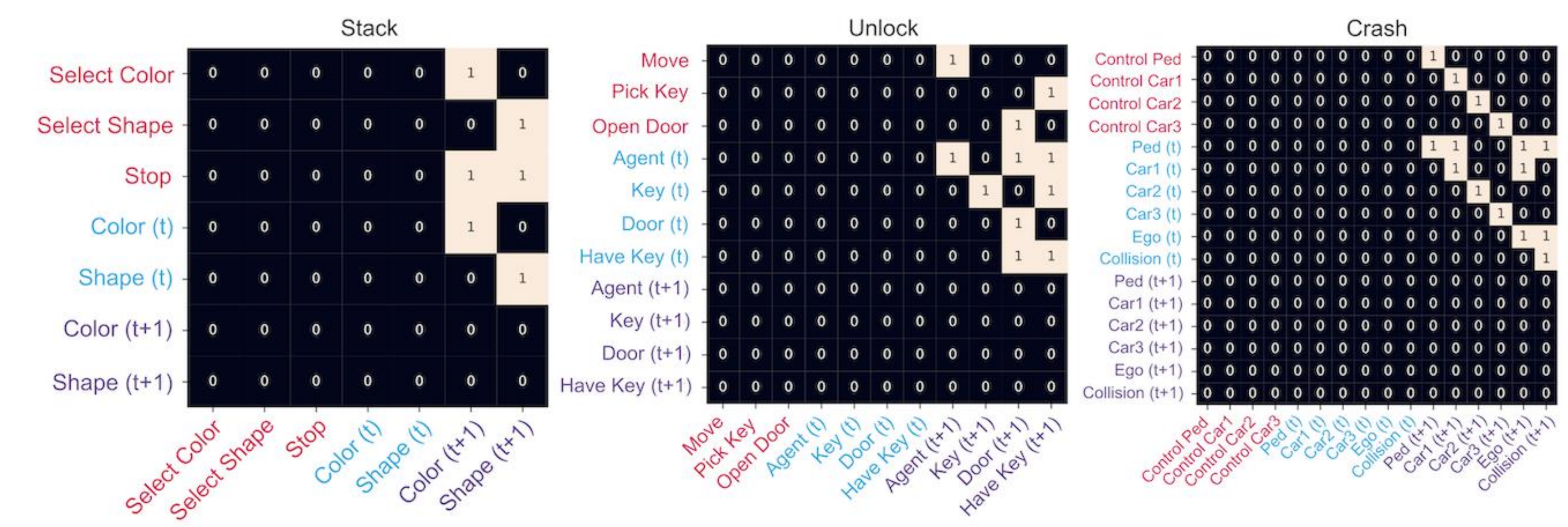


Stack: We design this manipulation task inspired by the CausalWorld, where the agent must stack objects to match specific shapes and colors

Unlock: We design a indoor house-holding task for the agent to collect a key to open doors.

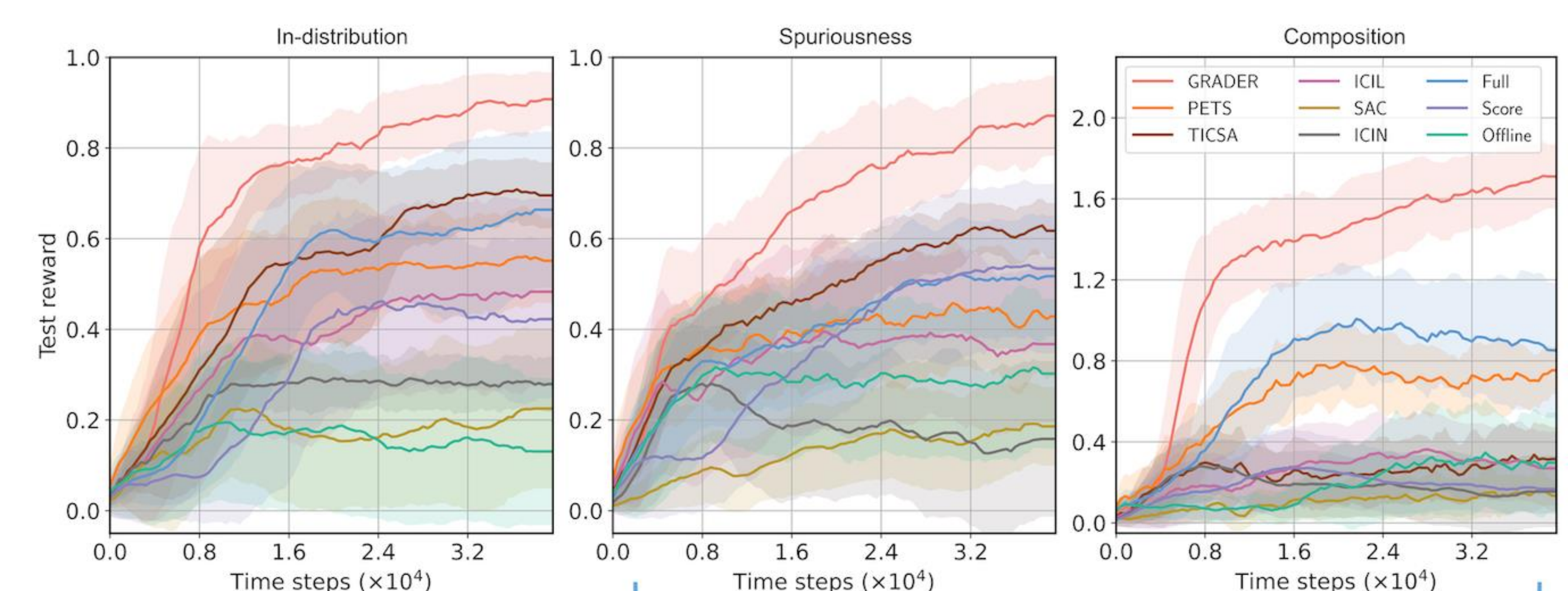
Crash: We design a crash scenario, where the goals are to create crashes between a pedestrian and different AVs.

Chemistry: An underlying causal graph controls the color-changing mechanism of all nodes

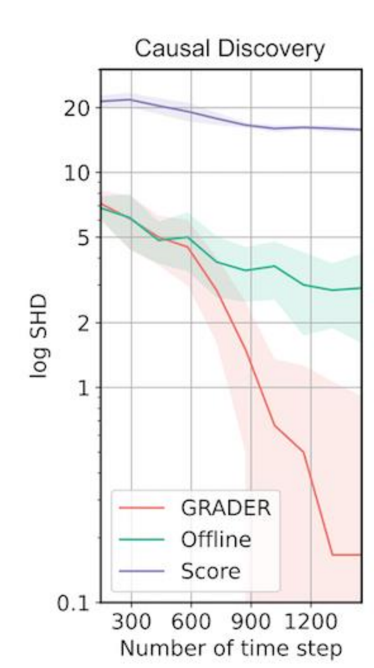


True causal graph of three environments. For the Chemistry environment, please check our paper for the 4 graphs used in our experiments

RL Generalization Improvement



Causality helps generalize to unseen scenarios



Our method is more efficient than score-based discovery method and offline discovery setting.

As the causal graph is closer to true graph, the task performance is better