因果强化学习最新进展分享

陈雄辉
南京大学

About Me

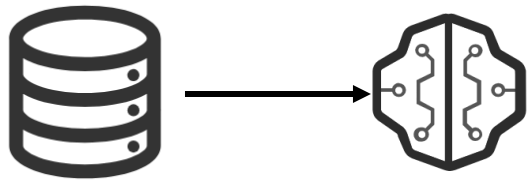
- 陈雄辉，南京大学20级硕转博博士生
- 南京大学LAMDA组俞扬老师指导下从事研究工作
- 研究关注：解决强化学习在交互成本敏感的真实应用场景中的挑战。目前的研究重点是离线强化学习、sim2real迁移、可泛化的真实世界环境模型学习。最近在探索基于大语言模型的决策和大型决策模型等相关课题。
- 10+篇论文发表在NeurIPS，ICML，ICLR，TPAMI等顶会上
- 关注学术成果转化：强化学习产品化落地：互联网企业（滴滴，美团），化工企业（施耐德），军工企业等
- 个人主页：https://xionghuichen.github.io/

目录

Reinforcement Learning

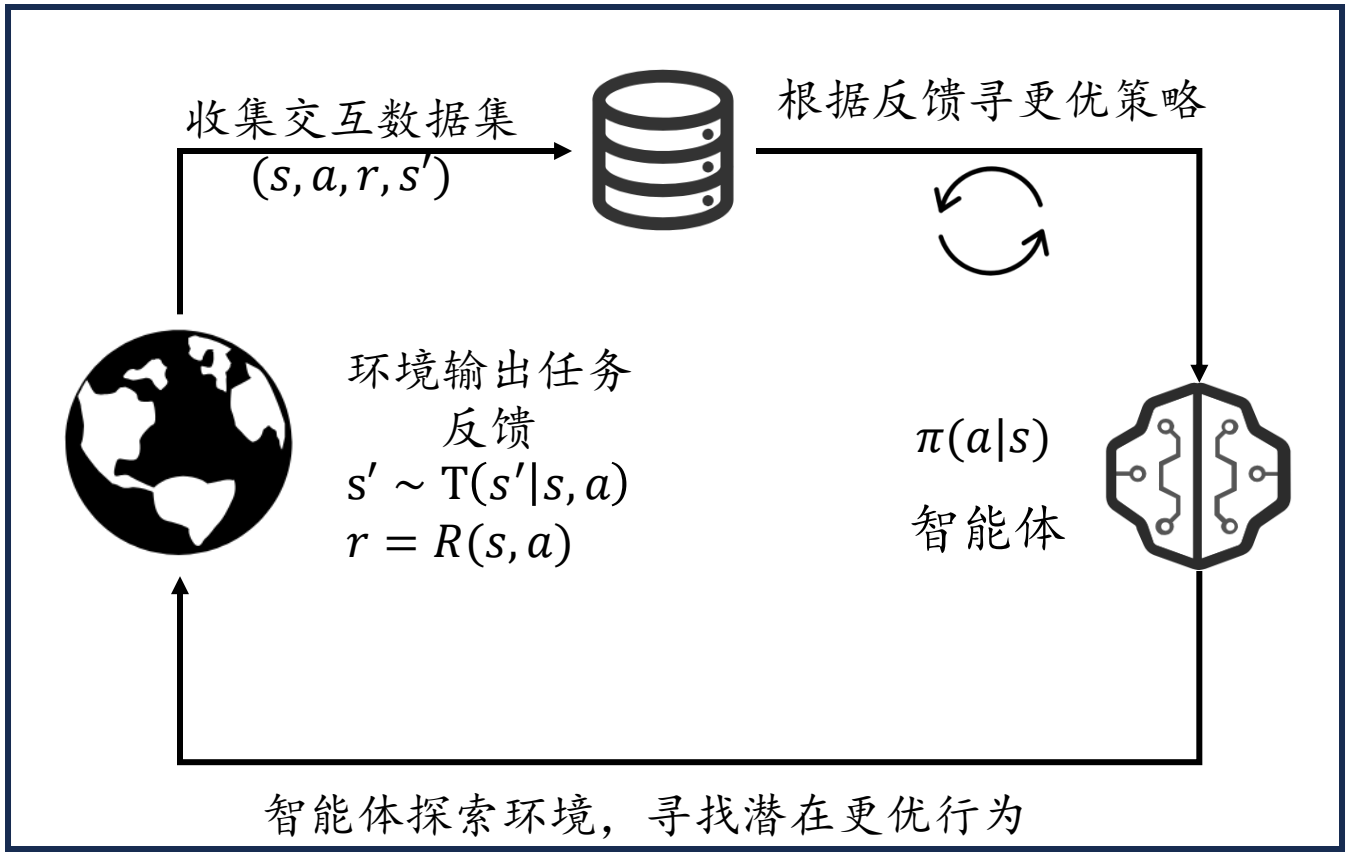
监督学习：

找到一个最优的模型，预测数据集中的标记



强化学习：

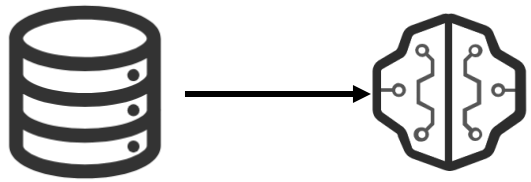
1. 搜索：在环境中进行各种试错，寻找对于完成指定任务最优的，每一个状态s下要做的动作a^*
2. 记忆/预测：将找到的最优动作用策略模型记忆下来



Reinforcement Learning

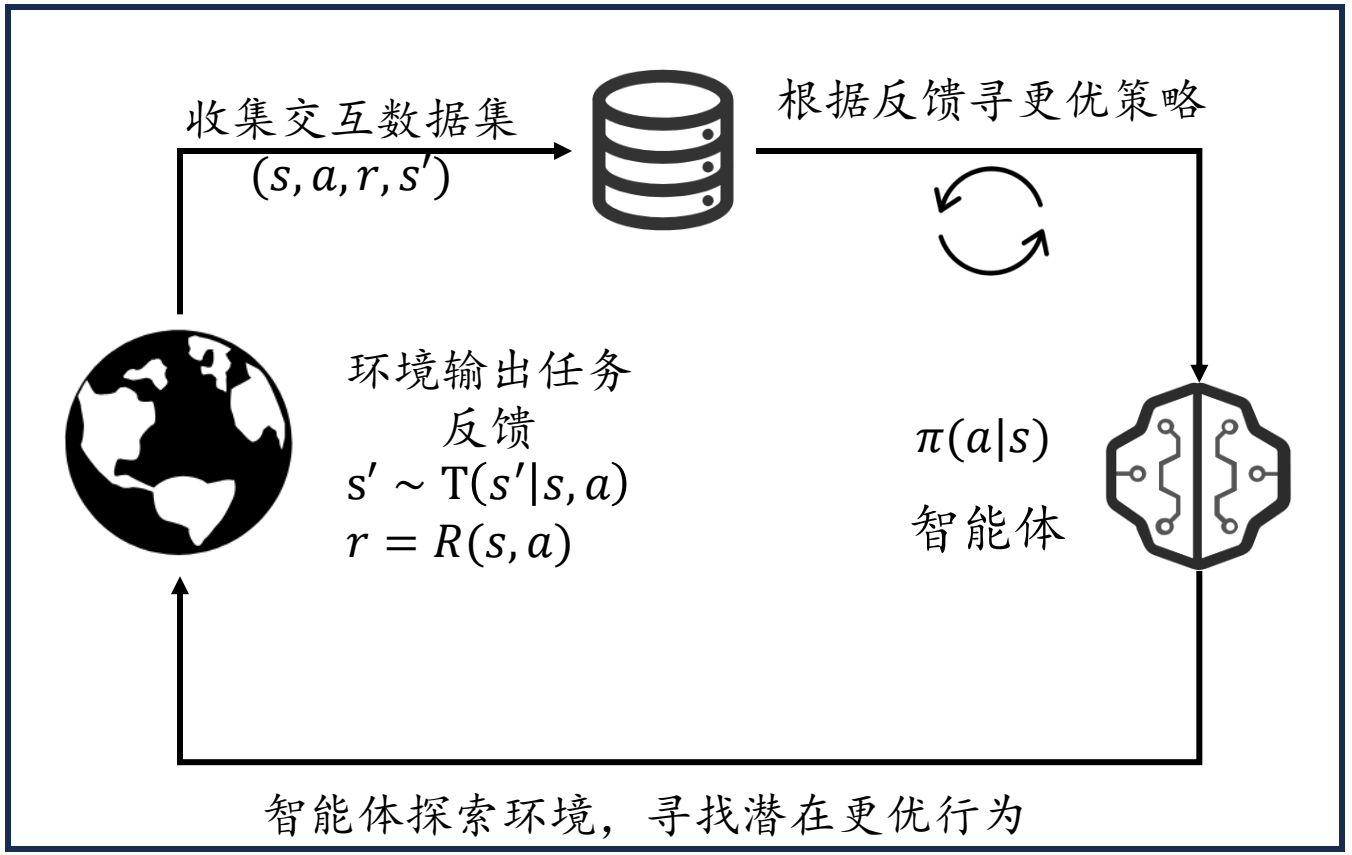
监督学习：

找到一个最优的模型，预测数据集中的标记



强化学习：

1. 搜索：在环境中进行各种试错，寻找对于完成指定任务最优的，每一个状态s下要做的动作a^*
2. 记忆/预测：将找到的最优动作用策略模型记忆下来



之所以需要强化，是因为在每一个状态下的s 最优标记a^*是事先未知的

Reinforcement Learning in trial-and-error cost sensitive applicaitons

强化学习：

1. 搜索：在环境中进行各种试错，寻找对于完成指定任务最优的，每一个状态s下要做的动作a^*
2. 记忆/预测：将找到的最优动作用策略模型记忆下来

AlphaGo

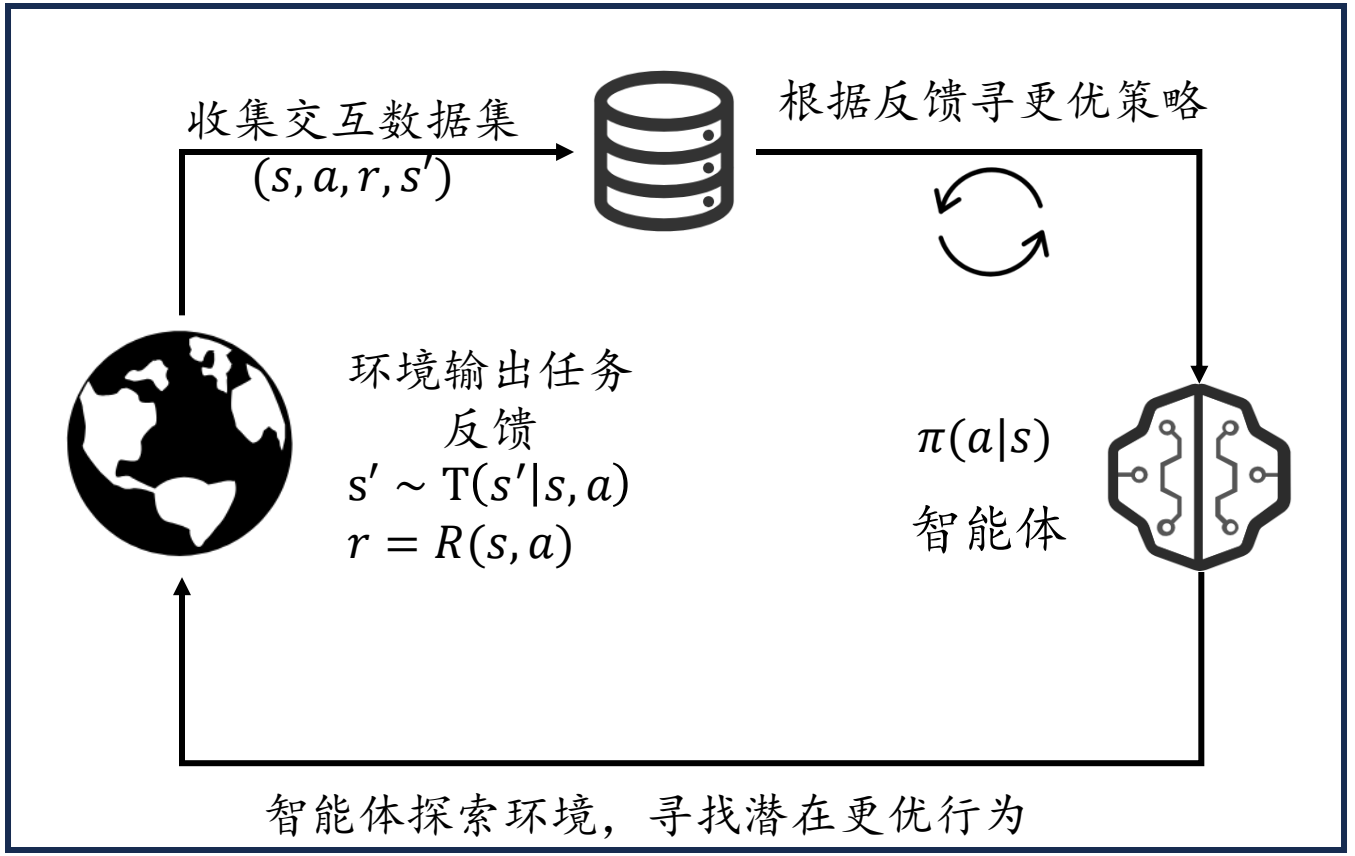
OpenAI Five

√

机器人

线上定价系统

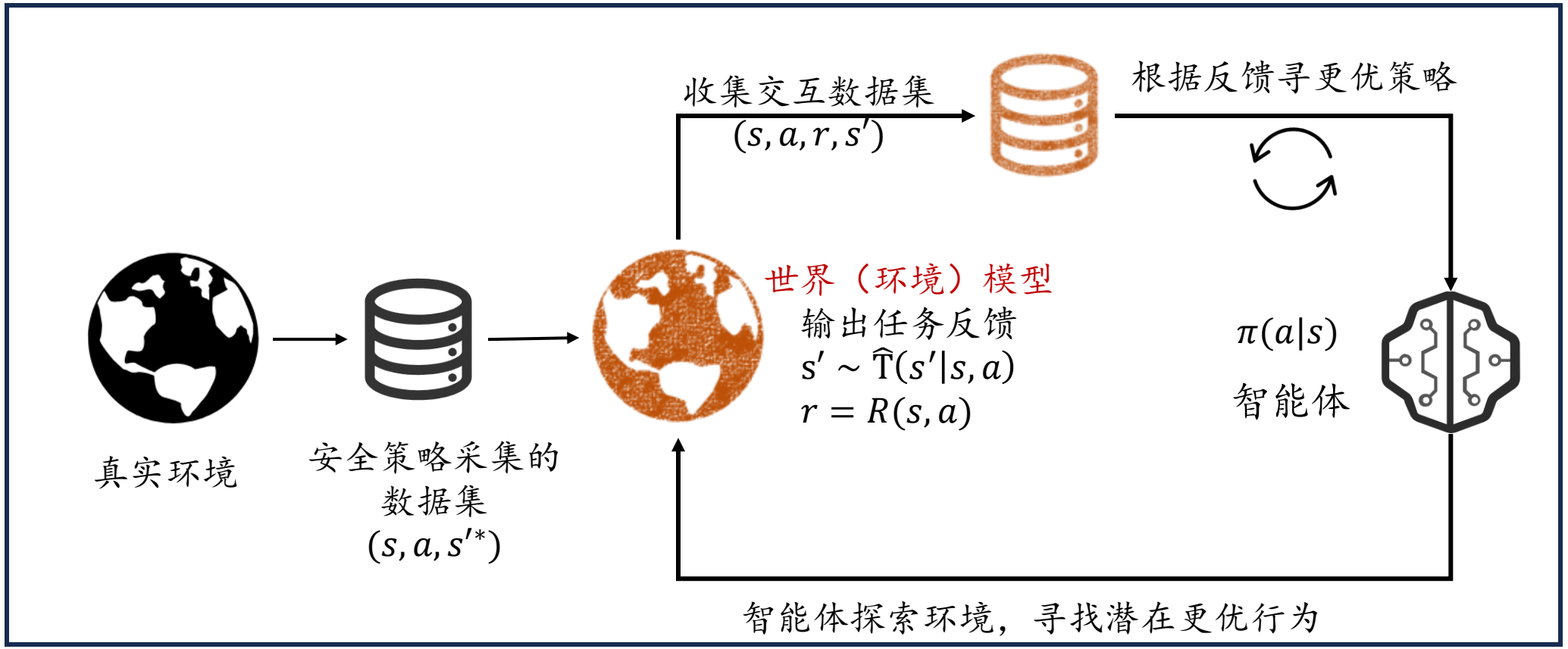
x



在很多应用场景中，“试错”是一种获取反馈昂贵的行为，会产生不可控的生产风险。

Q:有没有可能，尽可能少试错，甚至不试错，就完成策略求解的闭环？

World Model in Reinforcement Learning



Q: 有没有可能，尽可能少试错，甚至不试错，就完成策略求解的闭环？

A: 如果我们能够从真实环境中获得世界模型（环境模型），通过和模型交互进行试错，就可以减少/避免其在真实环境试错带来的成本和风险

The Role of World Model in Reinforcement Learning

1. 世界模型的作用是为策略优化提供交互的目标和获取反馈数据的数据源；
2. 为了找到更优策略，我们需要询问世界模型，之前没见过的动作，会有什么样的反馈。本质上是回答一个“*what if* I take another action?”的问题；
3. 这些动作，可能在离线数据集中也没见过。

The Role of World Model in Reinforcement Learning

1. 世界模型的作用是为策略优化提供交互的目标和获取反馈数据的数据源；
2. 为了找到更优策略，我们需要询问世界模型，之前没见过的动作，会有什么样的反馈。本质上是回答一个“*what if* I take another action?”的问题；
3. 这些动作，可能在离线数据集中也没见过。

Q:监督学习是否足以获得好的环境模型？

The Role of World Model in Reinforcement Learning

1. 世界模型的作用是为策略优化提供交互的目标和获取反馈数据的数据源；
2. 为了找到更优策略，我们需要询问世界模型，之前没见过的动作，会有什么样的反馈。本质上是回答一个“*what if* I take another action?”的问题；
3. 这些动作，可能在离线数据集中也没见过。

Q:监督学习是否足以获得好的环境模型？

1. 可以，但是训练效率低，因为有些状态预测的错误并不影响最后策略求解的性能。
2. 不可以，因为相关性不等于因果性。存在一些干扰的变量会影响泛化能力
3. 不可以，因为离线数据生成时的内在选择偏好会让环境模型误判因果关系，导致其回答“what if”问题时会出现灾难性的失败

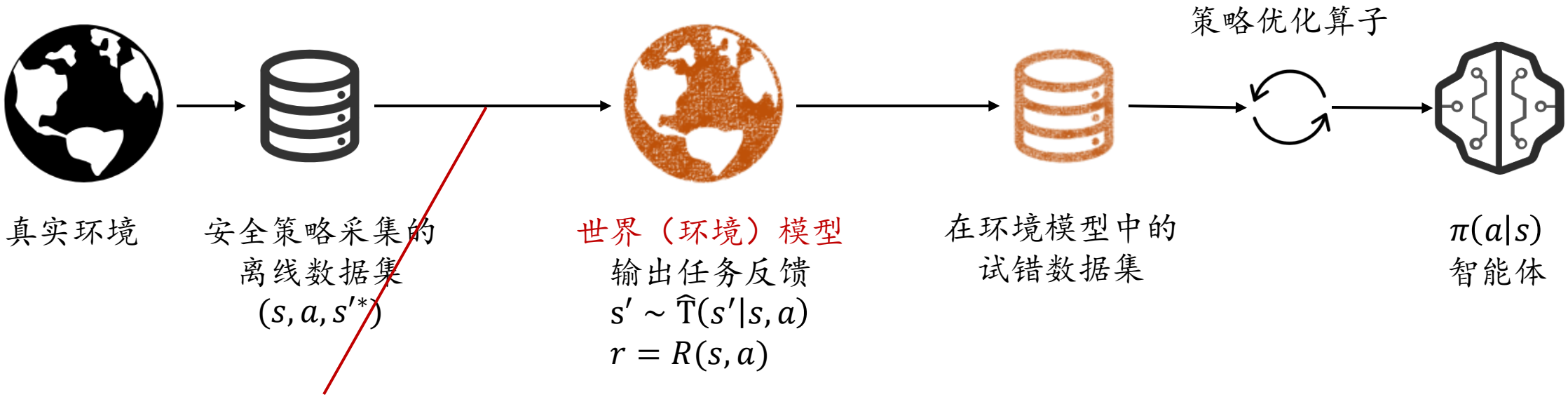
The Role of World Model in Reinforcement Learning

Q: 监督学习是否足以获得好的环境模型？

- 可以，但是训练效率低，因为有些状态预测的错误并不影响最后策略求解的性能。
  - ✓ **基于因果表征学习的强化学习方法**
- 不可以，因为相关性不等于因果性。存在一些干扰的变量会影响泛化能力
  - ✓ **基于因果结构发现的离线强化学习方法**
- 不可以，因为离线数据生成时的内在选择偏好会让环境模型误判因果关系，导致其回答“what if”问题时会出现灾难性的失败
  - ✓ **基于动作因果效用估计的离线强化学习方法**

目录

The Role of World Model in Reinforcement Learning



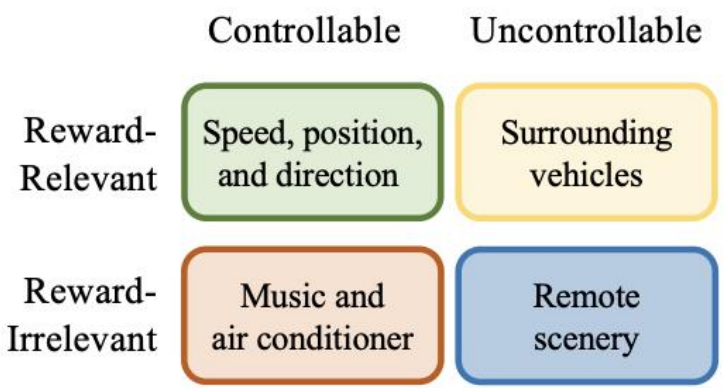
Q: 监督学习是否足以获得好的环境模型？

- 可以，但是训练效率低，因为有些状态预测的错误并不影响最后策略求解的性能。
  - ✓ **基于因果表征学习的强化学习方法**

环境模型中的四种成分

给定环境，决定策略性能的两个要素：（1）你做的动作；（2）你获得的奖励

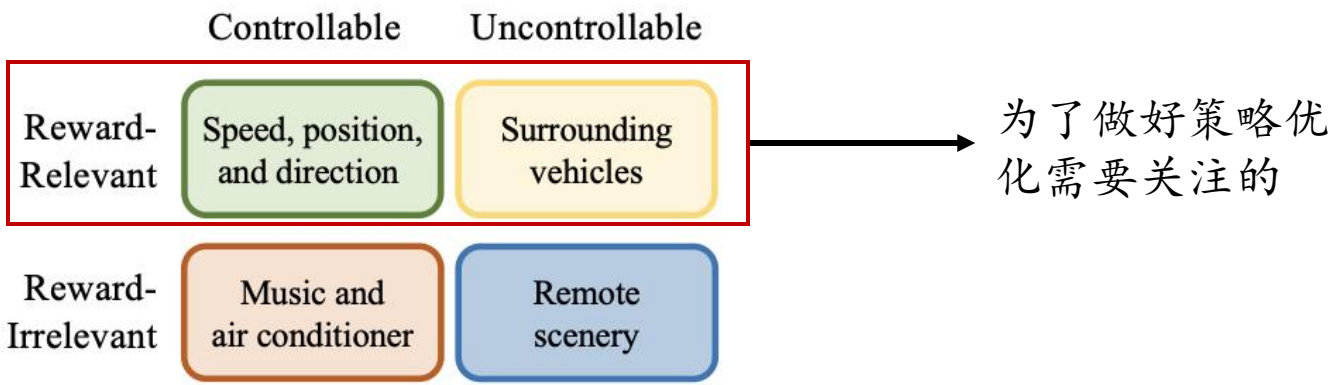
以此为基准，我们可以将环境的状态分成四种成分



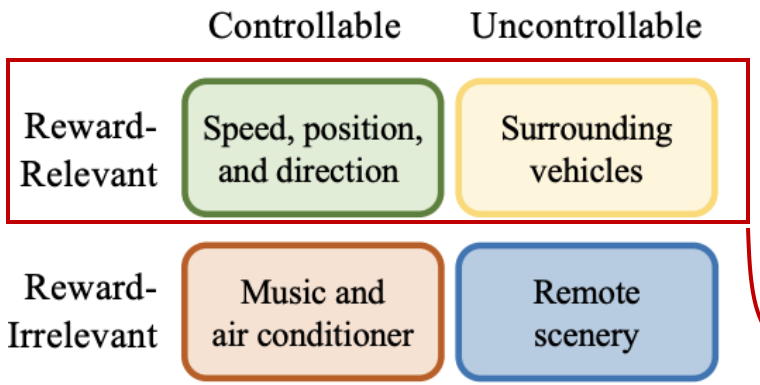
环境模型中的四种成分

给定环境，决定策略性能的两个要素：（1）你做的动作；（2）你获得的奖励

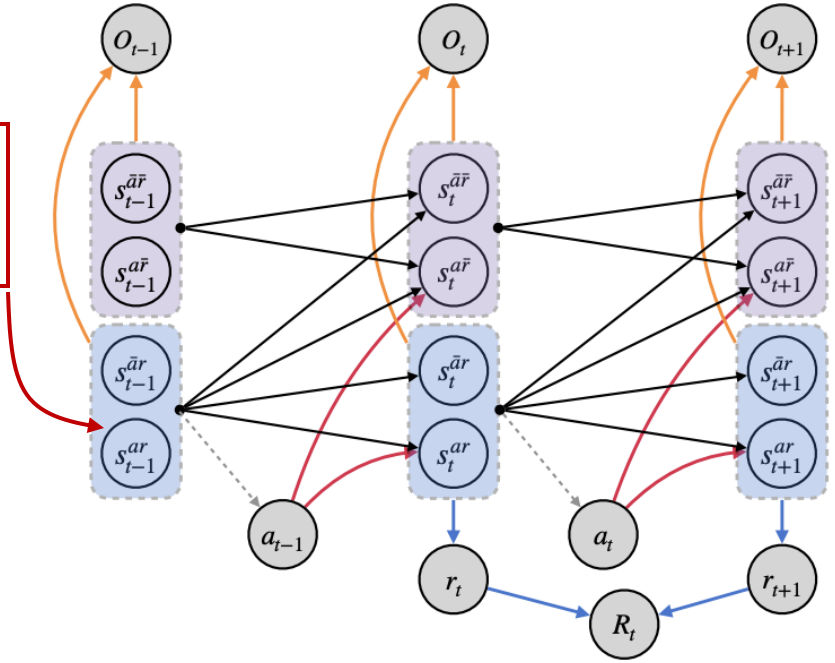
以此为基准，我们可以将环境的状态分成四种成分



环境模型中的四种成分：world model 建模



将环境的状态分成四种成分



得到对应的因果图

Disentangled Transition Model:

$$\begin{cases} p_{\gamma_1}(s_t^{ar} \mid s_{t-1}^{r}, a_{t-1}) \\ p_{\gamma_2}(s_t^{\bar{a}r} \mid s_{t-1}^{r}) \\ p_{\gamma_3}(s_t^{a\bar{r}} \mid \mathbf{s_{t-1}}, a_{t-1}) \\ p_{\gamma_4}(s_t^{\bar{a}\bar{r}} \mid \mathbf{s_{t-1}}) \end{cases}$$

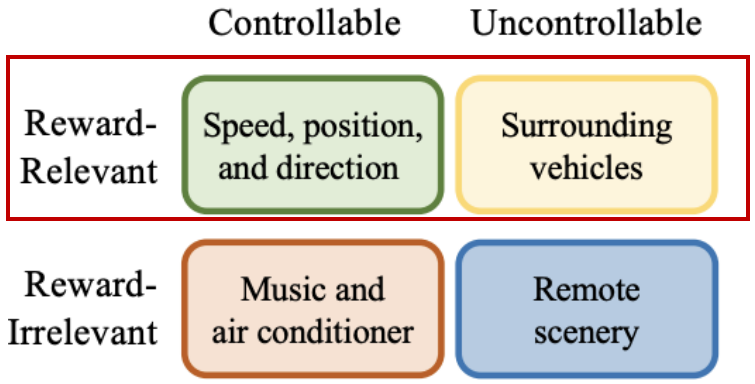
Disentangled Representation Model:

$$\begin{cases} q_{\phi_1}(s_t^{ar} \mid o_t, s_{t-1}^{r}, a_{t-1}) \\ q_{\phi_2}(s_t^{\bar{a}r} \mid o_t, s_{t-1}^{r}) \\ q_{\phi_3}(s_t^{a\bar{r}} \mid o_t, \mathbf{s_{t-1}}, a_{t-1}) \\ q_{\phi_4}(s_t^{\bar{a}\bar{r}} \mid o_t, \mathbf{s_{t-1}}) \end{cases}$$

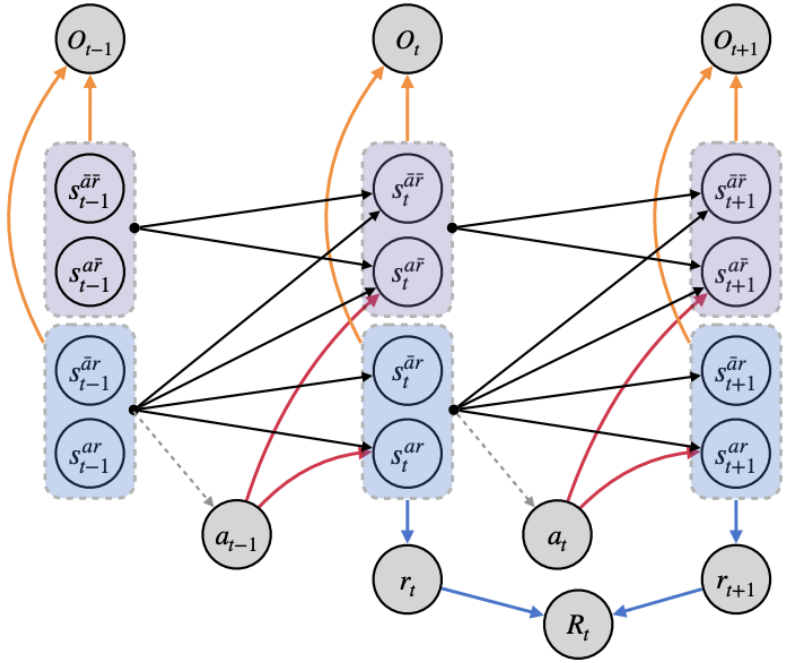
根据图连接关系获得多个子模型的连接关系

- Type 1: s_t^{ar} has an incident edge from a_{t-1}, and there is a directed path from s_t^{ar} to r_t.
- Type 2: $s_t^{\bar{a}r}$ has no incident edge from a_{t-1}, and there is a directed path from $s_t^{\bar{a}r}$ to r_t.
- Type 3: $s_t^{a\bar{r}}$ has an incident edge from a_{t-1}, and there is no directed path from $s_t^{a\bar{r}}$ to r_t.
- Type 4: $s_t^{\bar{a}\bar{r}}$ has no incident edge from a_{t-1}, and there is no directed path from $s_t^{\bar{a}\bar{r}}$ to r_t.

IFactor： World Models with Identifiable Factorization



将环境的状态分成四种成分



得到对应的因果图

Disentangled Transition Model:

$$\begin{cases} p_{\gamma_1}(s_t^{ar} \mid s_{t-1}^{r}, a_{t-1}) \\ p_{\gamma_2}(s_t^{\bar{a}r} \mid s_{t-1}^{r}) \\ p_{\gamma_3}(s_t^{a\bar{r}} \mid \mathbf{s_{t-1}}, a_{t-1}) \\ p_{\gamma_4}(s_t^{\bar{a}\bar{r}} \mid \mathbf{s_{t-1}}) \end{cases}$$

Disentangled Representation Model:

$$\begin{cases} q_{\phi_1}(s_t^{ar} \mid o_t, s_{t-1}^{r}, a_{t-1}) \\ q_{\phi_2}(s_t^{\bar{a}r} \mid o_t, s_{t-1}^{r}) \\ q_{\phi_3}(s_t^{a\bar{r}} \mid o_t, \mathbf{s_{t-1}}, a_{t-1}) \\ q_{\phi_4}(s_t^{\bar{a}\bar{r}} \mid o_t, \mathbf{s_{t-1}}) \end{cases}$$

根据图连接关系获得多个子模型的连接关系

$$\mathcal{J}_{\text{TOTAL}} = \mathrm{E}_{q_\phi}\left(\sum_t \left(\mathcal{J}_{\text{O}}^t + \mathcal{J}_{\text{R}}^t + \mathcal{J}_{\text{D}}^t + \mathcal{J}_{\text{RS}}^t + \mathcal{J}_{\text{AS}}^t\right)\right).$$

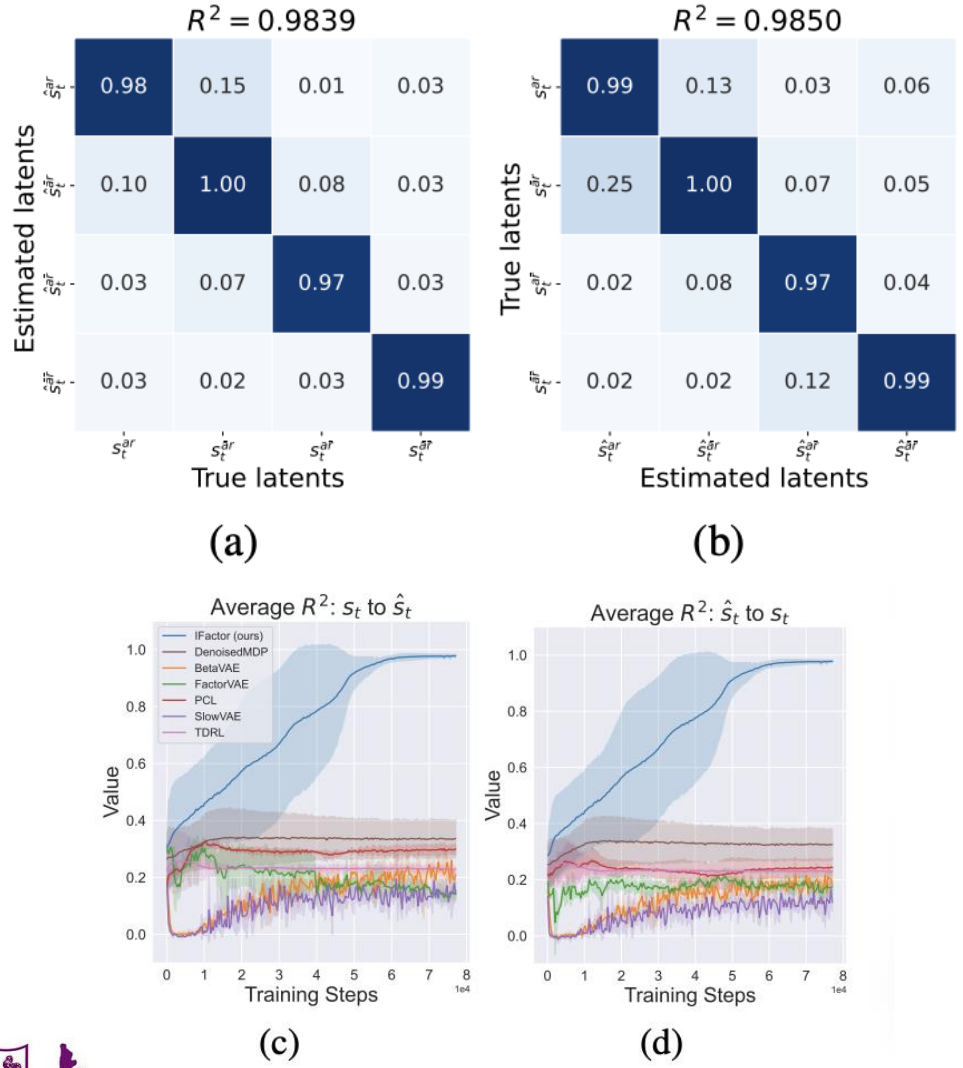
重构目标： $\mathcal{J}_{\text{O}}^t = \ln p_\theta(o_t \mid \mathbf{s_t}) \quad \mathcal{J}_{\text{R}}^t = \ln p_\theta(r_t \mid s_t^r)$

预测和推断一致性 $\mathcal{J}_{\text{D}}^t = -\beta_1 \cdot \mathrm{KL}\left(q_{\phi_1} \| p_{\gamma_1}\right) - \beta_2 \cdot \mathrm{KL}\left(q_{\phi_2} \| p_{\gamma_2}\right) - \beta_3 \cdot \mathrm{KL}\left(q_{\phi_3} \| p_{\gamma_3}\right) - \beta_4 \cdot \mathrm{KL}\left(q_{\phi_4} \| p_{\gamma_4}\right)$

解耦表征 $\mathcal{J}_{\text{RS}}^t = \lambda_1 \cdot \{I_{\alpha_1}(R_t; s_t^r, a_{t-1:t}, \mathbf{sg}(s_{t-1}^r)) - I_{\alpha_2}(R_t; s_t^{\bar{r}}, a_{t-1:t}, \mathbf{sg}(s_{t-1}^r))\}.$

$\mathcal{J}_{\text{AS}}^t = \lambda_2 \cdot \{I_{\alpha_3}(a_{t-1}; s_t^a, \mathbf{sg}(\mathbf{s_{t-1}})) - I_{\alpha_4}(a_{t-1}; s_t^{\bar{a}}, \mathbf{sg}(\mathbf{s_{t-1}}))\}.$

实验



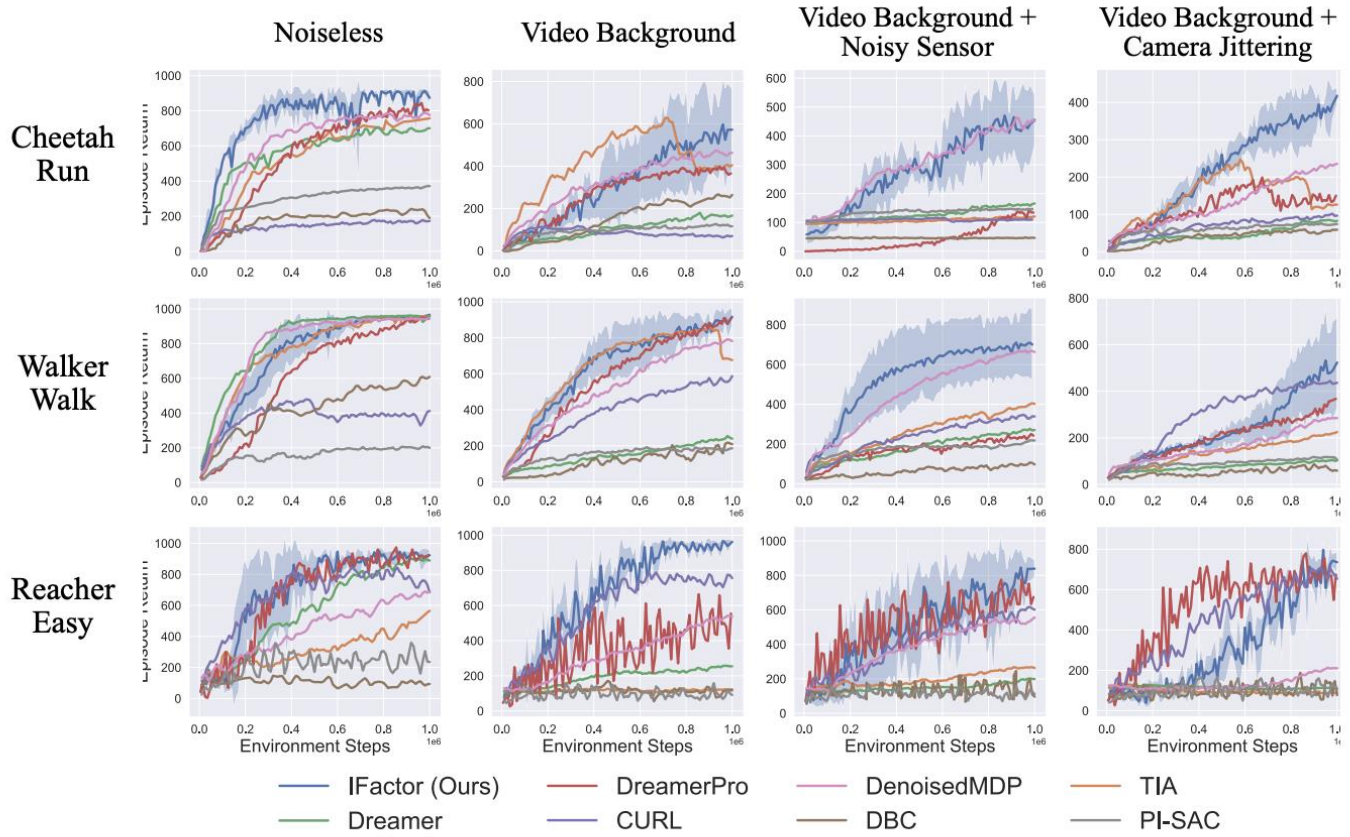


Figure 6: Policy optimization on variants of DMC with various distractors.

The Role of World Model in Reinforcement Learning

Q: 监督学习是否足以获得好的环境模型？

- 可以，但是训练效率低，因为有些状态预测的错误并不影响最后策略求解的性能。
  - ✓ 基于因果表征学习的强化学习方法
- 不可以，因为相关性不等于因果性。存在一些干扰的变量会影响泛化能力
  - ✓ **基于因果结构发现的离线强化学习方法**

Spurious variable影响模型泛化能力： a 2D-car example

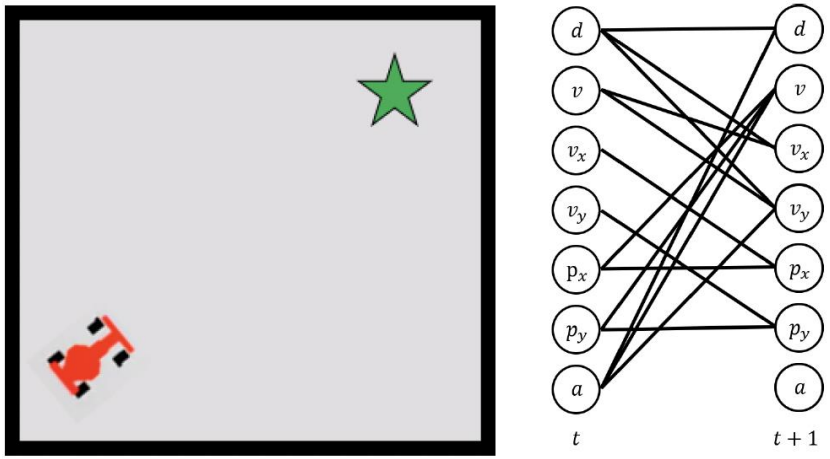


Fig. 11: The visualization of the state and the causal structure for the Car Driving benchmark. **Left:** the Toy Car Driving. The goal of the agent is to arrive at the star-shape destination. **Right:** The ground truth of the causal structure in Toy Car Driving. The state is vector-based and its value is continuous.

The state includes the direction d, the velocity v, the velocity on the x-axis v_x, the velocity on the y-axis v_y and the position (p_x, p_y). The action is the steering angle a.

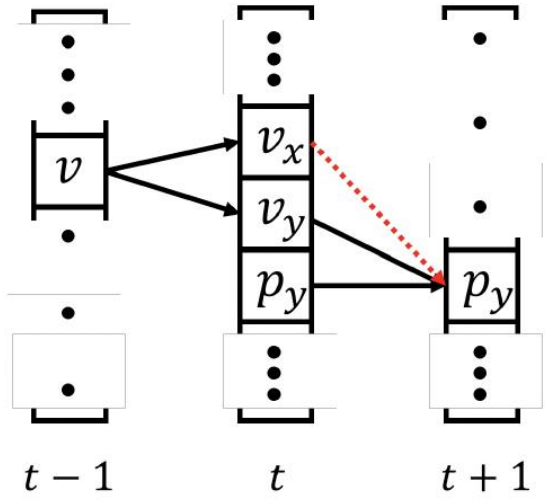


Fig. 4: The visualization of the example. The red dotted arrow presents that $(v_x)_t$ is a spurious variable for $(p_y)_{t+1}$.

Example: As shown in Figure 4, when the velocity v_{t-1} maintains stationary due to an imperfect sample policy, $(v_x)_t$ and $(v_y)_t$ have strong relatedness that $(v_x)_t^2 + (v_y)_t^2 = v_{t-1}^2$ and one can represent the other. Since we design that $(p_y)_{t+1} - (p_y)_t = (v_y)_t$, $(v_x)_t$ and $(p_y)_{t+1} - (p_y)_t$ also have strong relatedness, which leads to that $(v_x)_t$ becomes a spurious variable of $(p_y)_{t+1}$ given $(p_y)_t$, despite that $(v_x)_t$ is not the causal parent of y_{t+1}. By contrast, when the data is uniformly sampled with various velocities, this spuriousness will not exist.

Spurious variable影响模型泛化能力： 理论分析

Theorem 2 (RL Spurious Theorem). *Given an MDP with the state dimension n_s and the action dimension n_a, a data-collecting policy π_D, let M^* denote the true transition model, M_θ denote the learned model that M_θ^i predicts the i^{th} dimension with spurious variable sets spu_i and causal variables cau_i, i.e., $\hat{S}_{t+1,i} = M_\theta^i((\boldsymbol{S}_t, \boldsymbol{A}_t) \circ \omega_{cau_i \cup spu_i})$. Let $V_\pi^{M_\theta}$ denote the policy value of the policy π in model M_θ and correspondingly $V_\pi^{M^*}$. For an arbitrary bounded divergence policy π, i.e. $\max_S D_{KL}(\pi(\cdot|S), \pi_D(\cdot|S)) \leq \epsilon_\pi$, we have the policy evaluation error bound:*

$$|V_\pi^{M_\theta} - V_\pi^{M^*}| \leq \frac{2\sqrt{2}R_{max}}{(1-\gamma)^2}\sqrt{\epsilon_\pi} + \frac{R_{max}\gamma}{2(1-\gamma)^2} S_{max}[n_s\epsilon_c + (1+\gamma_{max})\lambda_{max} n_s(n_s+n_a)R_{spu}] \quad (6)$$

where

$$R_{spu} = \frac{\sum_{i=1}^{n_s} |spu_i|}{n_s(n_s+n_a)},$$

which represents the spurious variable density, that is, the ratio of spurious variables in all input variables.

Theorem 2 shows the relation between the policy evaluation error bound and the spurious variable density, which indicates that:

- When we use a non-causal model that all the spurious variables are input, R_{spu} reaches its maximum value $\bar{R}_{spu} < 1$. By contrast, in the optimal causal structure, R_{spu} reaches its minimum value of 0.
- The density of spurious variables R_{spu} and the correlation strength of spurious variables λ_{max} both influence the policy evaluation error bound. However, if we exclude all the spurious variables, i.e., $R_{spu} = 0$, the correlation strength of spurious variables will have no effect.

t+1的第i个变量有多少个t时刻的冗余父变量

FOCUS： oFfline mOdel-based reinforcement learning with CaUsal Structured World Models

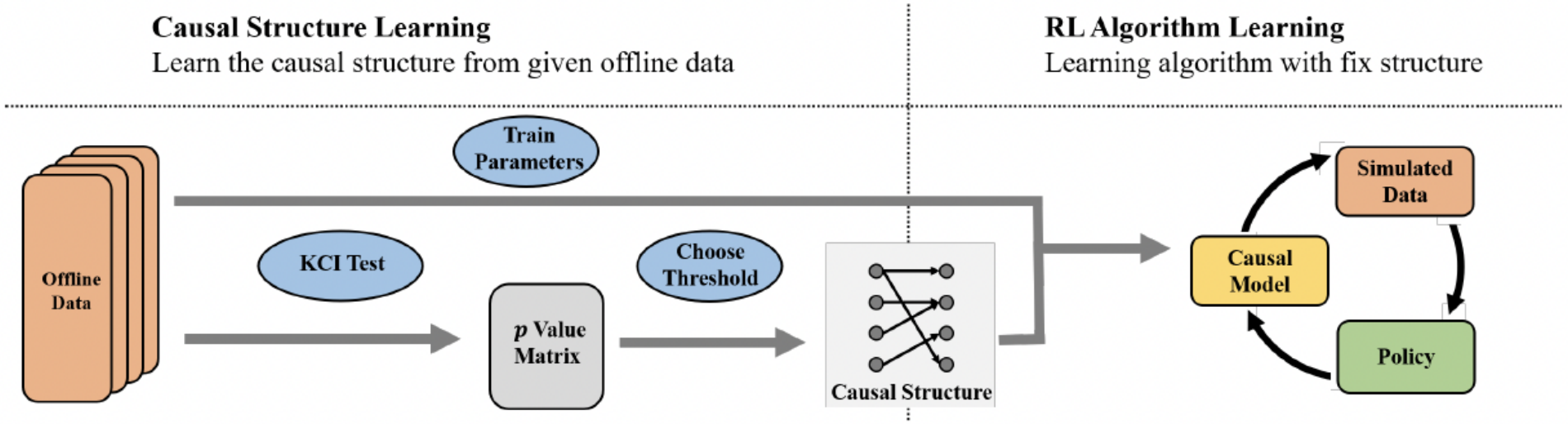


Fig. 1: The architecture of FOCUS. Given offline data, FOCUS learns a p value matrix by KCI test and then gets the causal structure by choosing a p threshold. After combining the learned causal structure with the neural network, FOCUS learns the policy through an offline MBRL algorithm.

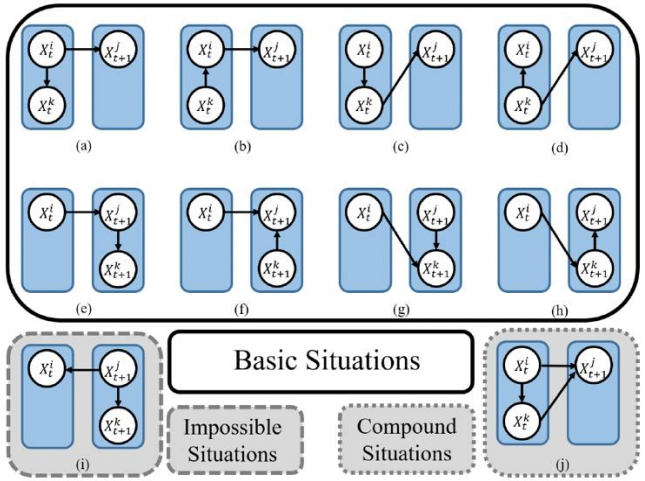


Fig. 3: The basic, impossible and compound situations of the causation between target variables and condition variables. In the basic situations, **Top Line:** (a)-(d) list the situations that the condition variable X^k is in the t time step. **Bottom Line:** Similarly, (e)-(h) list the situations that the condition variable X^k is in the $t+1$ time step.

Algorithm 2 Causal Structure Network $\mathcal{M}_{Causal}(\cdot)$

Input: state $\mathbf{s}_t \in \mathbb{R}^{n_s}$,
action $\mathbf{a}_t \in \mathbb{R}^{n_a}$,
causal structure mask matrix $G \in \{0,1\}^{(n_s,n_a)\times n_s}$,

Make $\mathcal{M}_i(\cdot;\theta_i)$ as the copy of the basic model $\mathcal{M}(\cdot;\theta)$, where $i = 1, \cdots, n_s$.
for $i = 1$ **to** n_s **do**
  Let $G_{\cdot,i}$ denote the i^{th} column of G
  Get the masked input $X = (\mathbf{s}_t, \mathbf{a}_t) \circ G_{\cdot,i}$
  Get prediction $\tilde{Y} = \mathcal{M}_i(X;\theta_i) \in \mathbb{R}^{n_s}$
  Let Y_i denote the i^{th} element of $\tilde{Y}$.
end for

Return $Y = (Y_i)_{i=1}^{n_s}$.

NANJING U

LAMDA Learning And Mining from DatA

Experiment

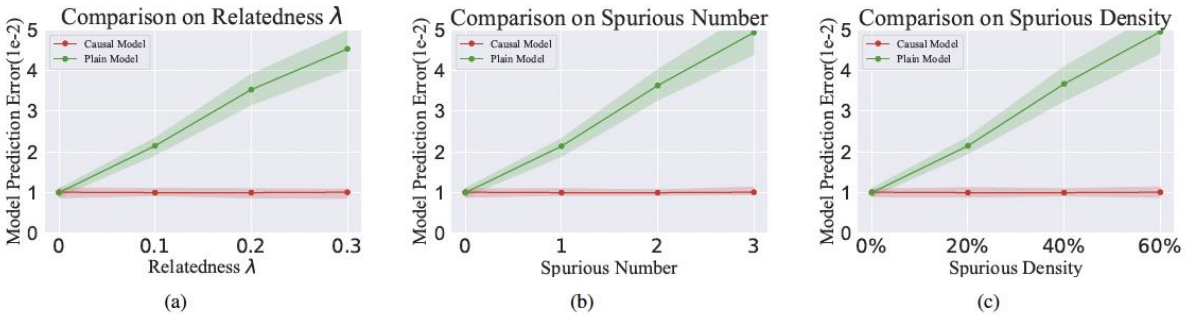


Fig. 6: Comparison of Causal World Models and Plain World Models to validate the spurious theorem. We evaluate the model prediction error on the relatedness, number and density of spurious variables in the offline dataset.

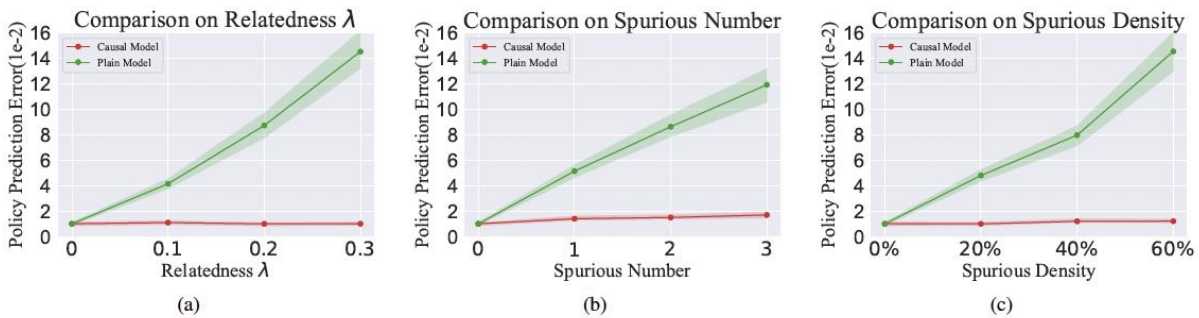


Fig. 7: Comparison of Causal World Models and Plain World Models to validate the RL spurious theorem. We evaluate the policy evaluation error on the relatedness, number and density of spurious variables in the offline dataset.

Table 1: The results on causal structure learning of our model and the baselines. Both the accuracy and the variance are calculated by five times experiments. *FOCUS (-KCI)* represents FOCUS with a linear independence test. *FOCUS (-CONDITION)* represents FOCUS with choosing all other variables as conditional variables.

INDEX	FOCUS	LNCM	FOCUS(-KCI)	FOCUS(-CONDITION)
ACCURACY	**0.993**	0.52	0.62	0.65
ROBUSTNESS	**0.001**	0.025	0.173	0.212
EFFICIENCY(SAMPLES)	$\mathbf{1 \times 10^6}$	1×10^7	1×10^6	1×10^6

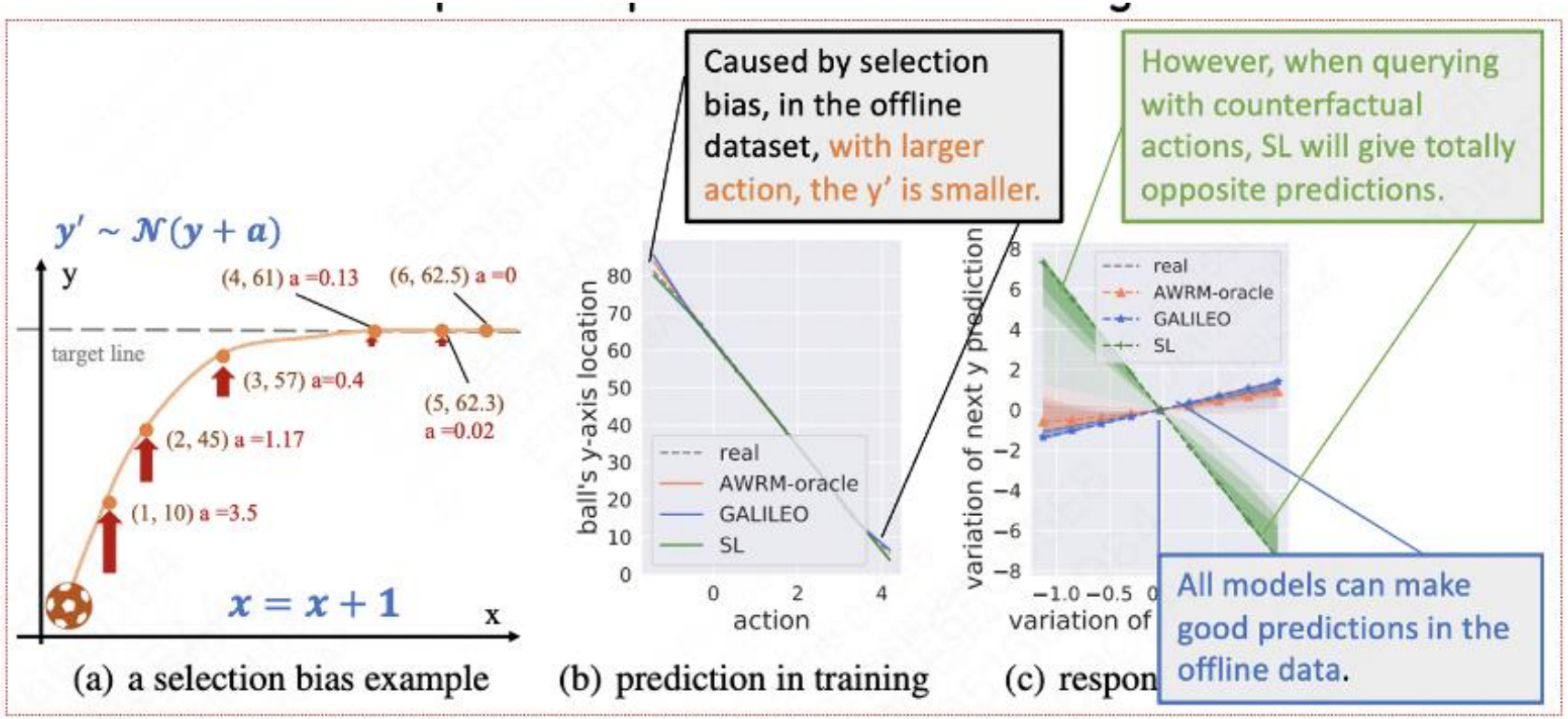
ENV	CAR DRIVING			MUJOCO(INVERTED PENDULUM)		
	RANDOM	MEDIUM	REPLAY	RANDOM	MEDIUM	REPLAY
FOCUS	**68.1 ± 20.9**	−58.9 ± 41.3	**86.2 ± 18.2**	**23.5 ± 17.9**	**24.9 ± 14.1**	**49.2 ± 19.0**
MOPO	−30.3 ± 49.9	−50.1 ± 34.2	46.2 ± 28.1	8.5 ± 6.2	2.5 ± 0.08	43.4 ± 7.7
LNCM	9.9 ± 42.5	**−5.4 ± 32.5**	11.4 ± 24.0	13.3 ± 0.9	3.1 ± 0.7	16.3 ± 6.4

The Role of World Model in Reinforcement Learning

Q: 监督学习是否足以获得好的环境模型？

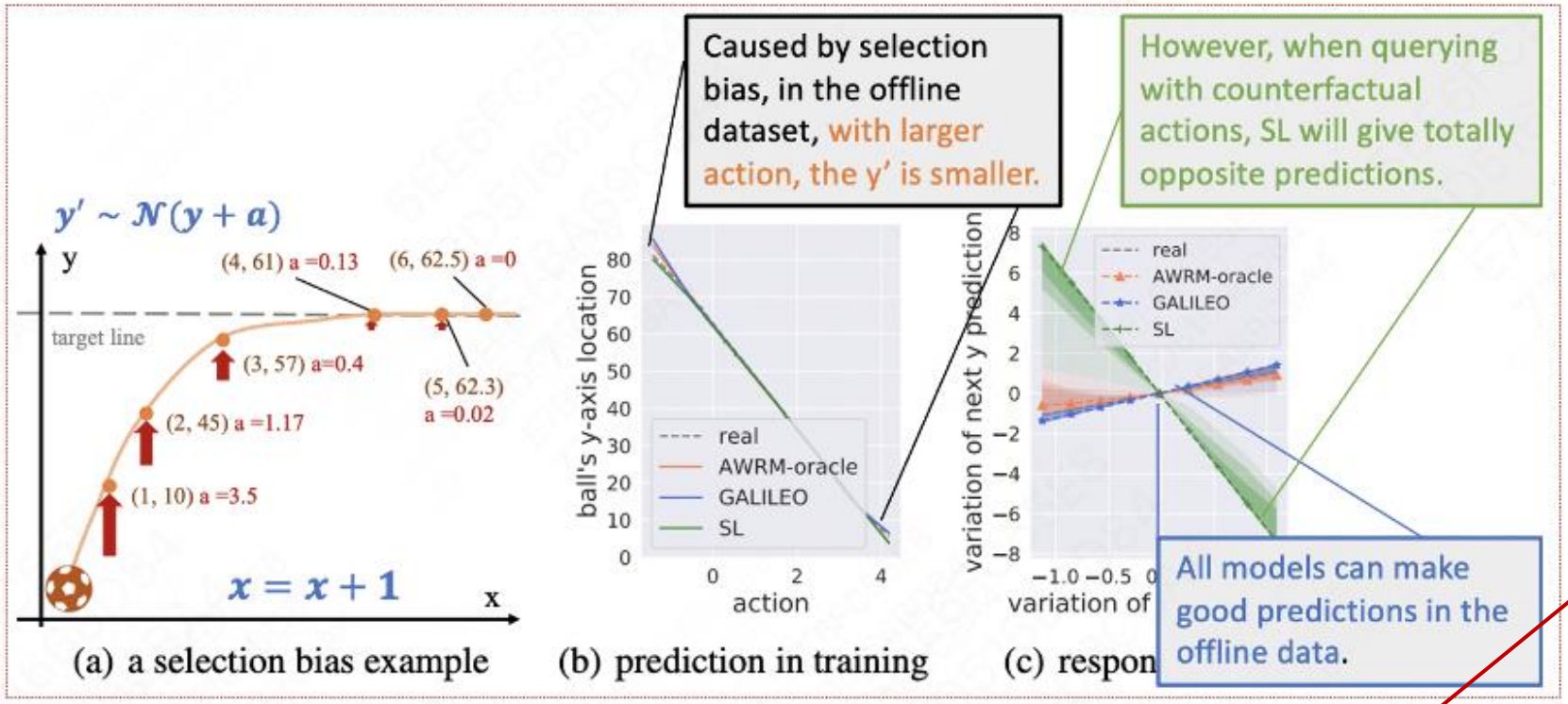
- 可以，但是训练效率低，因为有些状态预测的错误并不影响最后策略求解的性能。
  - ✓ 基于因果表征学习的强化学习方法
- 不可以，因为相关性不等于因果性。存在一些干扰的变量会影响泛化能力
  - ✓ 基于因果结构发现的离线强化学习方法
- 不可以，因为离线数据生成时的内在选择偏好会让环境模型误判因果关系，导致其回答“what if”问题时会出现灾难性的失败
  - ✓ **基于动作因果效用估计的离线强化学习方法**

选择偏差对环境模型预测的影响

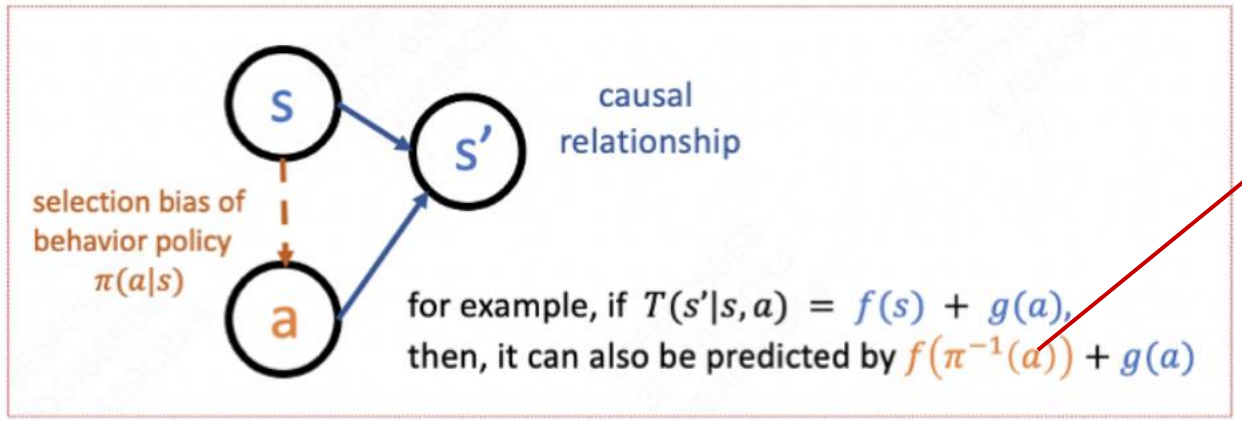


尽管训练数据集下的预测精度极高，但是此时在未见过的动作下做what if 询问的结果是完全错误的，这将严重误导策略优化

选择偏差对环境模型预测的影响：原因



当这部分的关系更容易被捕捉时，模型会倾向于偷懒用这种方式实现预测



according to $x_{t+1} = x_t + 1$ and $y_{t+1} \sim \mathcal{N}(y_t + a_t, 2)$. Here, a_t is chosen by a control policy $a_t \sim P_\phi(a|y_t) = \mathcal{N}((\phi - y_t)/15, 0.05)$ parameterized by ϕ, which tries to keep the ball near the line $y = \phi$. In Fig. 1(a), the behavior policy μ is $P_{62.5}$. Fig. 1(b) shows the collected training data and the learned models' prediction of the next position of y. Besides, the dataset superficially presents the relation that the corresponding next y will be smaller with a larger action. However, the truth is not because the larger a_t causes a smaller y_{t+1}, but the policy selects a small a_t when y_t is close to the target line. **Mistakenly exploiting the "association" will lead to local optima with serious factual errors**, e.g., believed that $y_{t+1} \propto P_\phi^{-1}(y_t|a) + a_t \propto \phi - 14a_t$, where P_ϕ^{-1} is the inverse function of P_ϕ. When we estimate the response curves by fixing y_t and reassigning action a_t with other actions $a_t + \Delta a$, where $\Delta a \in [-1, 1]$ is a variation of action value, we found that the model of SL indeed exploit the association and give opposite responses, while in AWRM and its practical implementation GALILEO, the predictions are closer to the ground truths ($y_{t+1} \propto y_t + a_t$). The result is in Fig. 1(c), where the darker a region is, the more samples are fallen in. AWRM injects data collected by adversarial policies for model learning to eliminate the unidentifiability between $y_{t+1} \propto P_\phi^{-1}(y_t|a) + a_t$ and $y_{t+1} \propto y_t + a_t$ in offline data.

解决方案：经验风险最小化 ->对抗重加权经验风险最小化（Adversarial Weighted empirical risk minimization）

经验风险最小化

$$\min_{T\in\mathcal{T}} \mathbb{E}_{s,a,s'\sim\rho^{\pi}_{T^\star}} [-\log T(s'|s,a)]$$

对抗重加权经验风险最小化

$$\min_{T\in\mathcal{T}} \max_{\beta\in\Pi} L(\rho^{\beta}_{T^\star}, T) = \min_{T\in\mathcal{T}} \max_{\beta\in\Pi} \mathbb{E}_{s,a,s'\sim\rho^{\mu}_{T^\star}} [\omega(s,a|\rho^{\beta}_{T^\star})\ell_T(s,a,s')],$$

$$\omega(s,a|\rho^{\beta}_{T^\star}) = \frac{\rho^{\beta}_{T^\star}(s,a)}{\rho^{\mu}_{T^\star}(s,a)}$$

Algorithm 2 AWRM with Oracle Counterfactual Datasets

Input:
Φ: policy space; N: total iterations
Process:

The idea of AWRM: Iteratively construct an adversarial policy that hinders the model's prediction, then update the model under this policy's distribution.

1: Generate counterfactual datasets $\{\mathcal{D}_{\pi_\phi}\}$ for all adversarial policies $\pi_\phi, \phi \in \Phi$
2: Initialize an environment model M_θ
3: **for** i = 1:N **do**
4: Select $\mathcal{D}_{\pi_\phi}$ with worst prediction errors through M_θ from $\{\mathcal{D}_{\pi_\phi}\}$
5: Optimize M_θ with standard supervised learning based on $\mathcal{D}_{\pi_\phi}$
6: **end for**

解决方案：经验风险最小化 ->对抗重加权经验风险最小化（Adversarial Weighted empirical risk minimization）

经验风险最小化

$$\min_{T\in\mathcal{T}} \mathbb{E}_{s,a,s'\sim\rho^{\pi}_{T\star}} [-\log T(s'|s,a)]$$

对抗重加权经验风险最小化

$$\min_{T\in\mathcal{T}} \max_{\beta\in\Pi} L(\rho^{\beta}_{T\star}, T) = \min_{T\in\mathcal{T}} \max_{\beta\in\Pi} \mathbb{E}_{s,a,s'\sim\rho^{\mu}_{T\star}} [\omega(s,a|\rho^{\beta}_{T\star})\ell_T(s,a,s')]$$

$$\omega(s,a|\rho^{\beta}_{T\star}) = \frac{\rho^{\beta}_{T\star}(s,a)}{\rho^{\mu}_{T\star}(s,a)}$$

Algorithm 2 AWRM with Oracle Counterfactual Datasets

Input:
Φ: policy space; N: total iterations
Process:

1: Generate counterfactual datasets $\{\mathcal{D}_{\pi_\phi}\}$ for all adversarial policies $\pi_\phi, \phi \in \Phi$
2: Initialize an environment model M_θ
3: **for** i = 1:N **do**
4: Select $\mathcal{D}_{\pi_\phi}$ with worst prediction errors through M_θ from $\{\mathcal{D}_{\pi_\phi}\}$
5: Optimize M_θ with standard supervised learning based on $\mathcal{D}_{\pi_\phi}$
6: **end for**

The idea of AWRM: Iteratively construct an adversarial policy that hinders the model's prediction, then update the model under this policy's distribution.

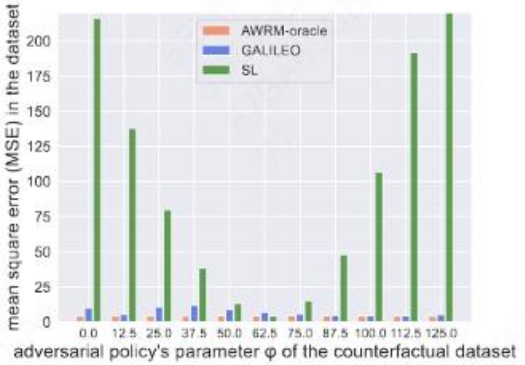


Figure 2: An illustration of the prediction error in counterfactual datasets. The error of SL is small only in training data (ϕ = 62.5) but becomes much larger in the dataset "far away from" the training data. AWRM-oracle selects the oracle worst counterfactual dataset for training for each iteration (pseudocode is in Alg. 2) which reaches small MSE in all datasets and gives correct response curves (Fig. 1(c)). GALILEO approximates the optimal adversarial counterfactual data distribution based on the training data and model. Although the MSE of GALILEO is a bit larger than SL in the training data, in the counterfactual datasets, the MSE is on the same scale as AWRM-oracle.

解决方案：经验风险最小化 ->对抗重加权经验风险最小化（Adversarial Weighted empirical risk minimization）

经验风险最小化

$$\min_{T\in\mathcal{T}} \mathbb{E}_{s,a,s'\sim\rho^{\pi}_{T^\star}} [-\log T(s'|s,a)]$$

对抗重加权经验风险最小化

$$\min_{T\in\mathcal{T}} \max_{\beta\in\Pi} L(\rho^{\beta}_{T^\star}, T) = \min_{T\in\mathcal{T}} \max_{\beta\in\Pi} \mathbb{E}_{s,a,s'\sim\rho^{\mu}_{T^\star}} [\omega(s,a|\rho^{\beta}_{T^\star})\ell_T(s,a,s')], \quad \omega(s,a|\rho^{\beta}_{T^\star}) = \frac{\rho^{\beta}_{T^\star}(s,a)}{\rho^{\mu}_{T^\star}(s,a)}$$

策略求解要应对大量不同策略进行评估，
因此需要对抗项，从而做到对策略空间的无偏估计
而不是对单一策略（比如随机策略）的去偏

序列决策要考虑一个动作的累计影响，
而不是单步的影响

（灵感来源）
因果效用估计中
基于逆倾向得分的
去偏方法

$$\min_{T\in\mathcal{T}} \mathbb{E}_{x,a,y\sim p^{\mu}_{T^*}} [\omega(x,a)\ell(T(y|x,a),y)], \quad \omega(x,a) := \frac{\beta(a|x)}{\mu(a|x)}$$

解决方案：经验风险最小化 ->对抗重加权经验风险最小化（Adversarial Weighted empirical risk minimization）

对抗重加权经验风险最小化

$$\min_{T\in\mathcal{T}}\max_{\beta\in\Pi} L(\rho^{\beta}_{T^\star}, T) = \min_{T\in\mathcal{T}}\max_{\beta\in\Pi} \mathbb{E}_{s,a,s'\sim\rho^{\mu}_{T^\star}}[\omega(s,a|\rho^{\beta}_{T^\star})\ell_T(s,a,s')],$$

$$\omega(s,a|\rho^{\beta}_{T^\star}) = \frac{\rho^{\beta}_{T^\star}(s,a)}{\rho^{\mu}_{T^\star}(s,a)}$$

Algorithm 2 AWRM with Oracle Counterfactual Datasets

Input:
Φ: policy space; N: total iterations
Process:

The idea of AWRM: Iteratively construct an adversarial policy that hinders the model's prediction, then update the model under this policy's distribution.

1: Generate counterfactual datasets $\{\mathcal{D}_{\pi_\phi}\}$ for all adversarial policies $\pi_\phi, \phi \in \Phi$
2: Initialize an environment model M_θ
3: **for** i = 1:N **do**
4: Select $\mathcal{D}_{\pi_\phi}$ with worst prediction errors through M_θ from $\{\mathcal{D}_{\pi_\phi}\}$
5: Optimize M_θ with standard supervised learning based on $\mathcal{D}_{\pi_\phi}$
6: **end for**

离线策略求解场景，
无法获得对抗数据分布

解决方案：经验风险最小化 ->对抗重加权经验风险最小化（Adversarial Weighted empirical risk minimization）

对抗重加权经验风险最小化

$$\min_{T\in\mathcal{T}}\max_{\beta\in\Pi} L(\rho^{\beta}_{T^\star}, T) = \min_{T\in\mathcal{T}}\max_{\beta\in\Pi} \mathbb{E}_{s,a,s'\sim\rho^{\mu}_{T^\star}}[\omega(s,a|\rho^{\beta}_{T^\star})\ell_T(s,a,s')], \quad \omega(s,a|\rho^{\beta}_{T^\star}) = \frac{\rho^{\beta}_{T^\star}(s,a)}{\rho^{\mu}_{T^\star}(s,a)}$$

$$\theta_{t+1} = \min_{\theta} \mathbb{E}_{\rho^{\mu}_{M^*}}\left[\frac{-1}{\alpha_0(x,a)}\log M_\theta(x'|x,a)\underbrace{\left(\underbrace{f\left(\frac{\rho^{\mu}_{M_{\theta_t}}(x,a,x')}{\rho^{\mu}_{M^*}(x,a,x')}\right)}_{\text{discrepancy}} - \underbrace{f\left(\frac{\rho^{\mu}_{M_{\theta_t}}(x,a)}{\rho^{\mu}_{M^*}(x,a)}\right)}_{\text{density-ratio baseline}} + \underbrace{H_{M^*}(x,a)}_{\text{stochasticity}}\right)}_{W(x,a,x')}\right],$$

1. 最优对抗分布是“模型学的越差权重越大”的分布
2. 对抗模仿学习进行模型学习，使用discriminator近似估计了最优对抗数据分布，当前模型生成的分布和真实环境的数据分布差异大的，基于discriminator的加权项会有更高的权重

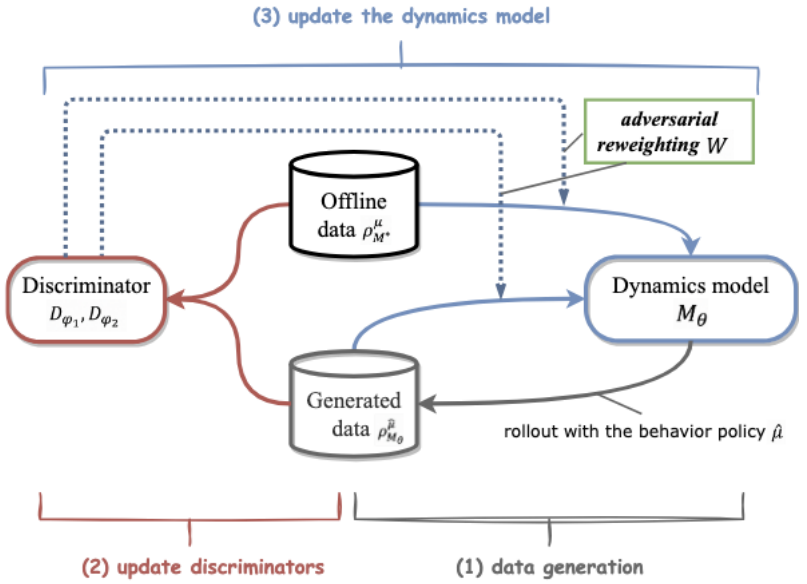


Figure 3: Illustration of the GALILEO workflow.

实验

$$\text{MISE} = \mathbb{E}\left[\int_{\mathcal{A}} (M^*(x'|x,a) - M(x'|x,a))^2 \, \mathrm{d}a\right]$$

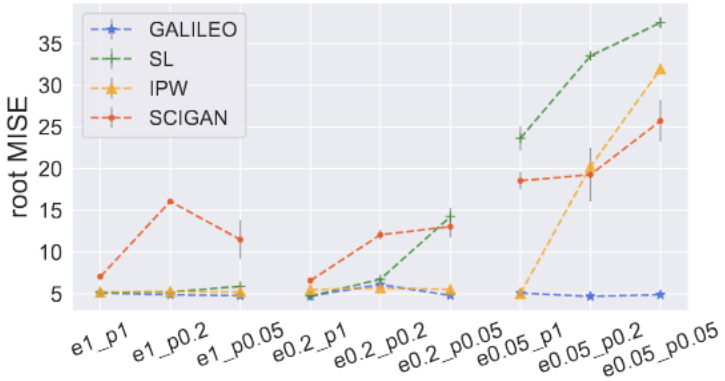
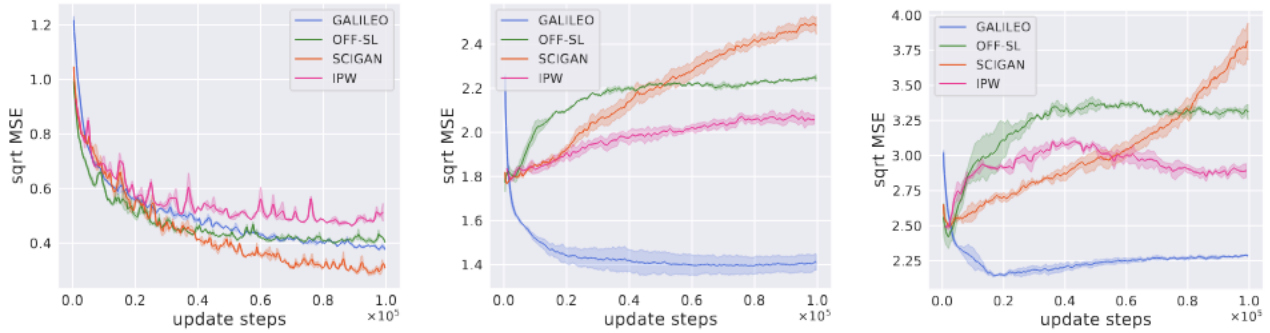


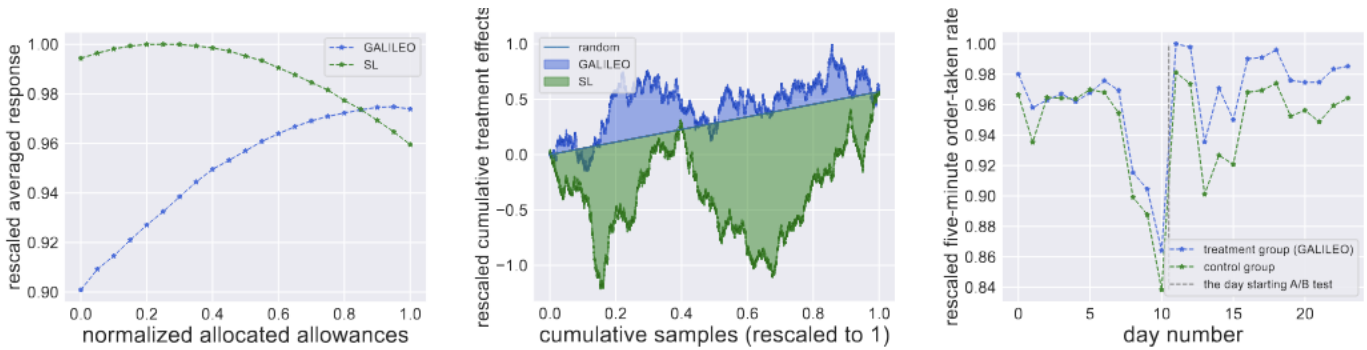
Table 1: Results of OPE on DOPE benchmark. We list the **averaged** performances on three tasks. The detailed results are in Appx. H.6. $\pm$ is the standard deviation among the tasks. We bold the best scores for each metric.

Algorithm	Norm. value gap	Rank corr.	Regret@1
GALILEO	**0.37 $\pm$ 0.24**	**0.44 $\pm$ 0.10**	**0.09 $\pm$ 0.02**
Best DICE	0.48 $\pm$ 0.19	0.15 $\pm$ 0.04	0.42 $\pm$ 0.28
VPM	0.71 $\pm$ 0.04	0.29 $\pm$ 0.15	0.17 $\pm$ 0.11
FQE (L2)	0.54 $\pm$ 0.09	-0.19 $\pm$ 0.10	0.34 $\pm$ 0.03
IS	0.67 $\pm$ 0.01	-0.40 $\pm$ 0.15	0.36 $\pm$ 0.27
Doubly Rubost	0.57 $\pm$ 0.07	-0.14 $\pm$ 0.17	0.33 $\pm$ 0.06



(a) medium (train) (b) medium-replay (test) (c) expert (test)

Figure 6: Illustration of learning curves of the halfcheetah tasks (full results are in Appx. H.5). The figures with titles ending in "(train)" means the dataset is used for training while the titles ending in "(test)" means the dataset is **just used for testing**. The X-axis records the steps of the environment model update, and the Y-axis is the prediction errors in the corresponding steps evaluated by the datasets. The solid curves are the mean reward and the shadow is the standard error of three seeds.



(a) response curves in City A (b) AUUC curves (c) A/B test in the City A

Figure 7: Parts of the results in BAT tasks. Fig. 7(a) demonstrate the averaged response curves of the SL and GALILEO model in City A. It is expected to be monotonically increasing through our prior knowledge. In Fig. 7(b) show the AUCC curves, where the model with larger areas above the "random" line makes better predictions in randomized-controlled-trials data [61].

总结

- 世界模型必须是考虑因果的
- 因果的机制的确能够让世界模型泛化得更好

挑战：

1. 其他的世界模型的问题，和因果上的解决方案
2. 规模可扩展的方法
3. 方法的假设实际不满足