

# PPO

## PPO×Family

### 第五讲：探索时序建模

主办



承办



协办



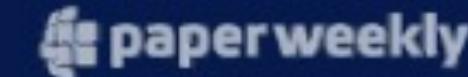
北京大学 人工智能  
研究院



浙江大学 上海高等研究院  
SHANGHAI INSTITUTE FOR ADVANCED STUDY  
ZHEJIANG UNIVERSITY



支持



# 向真实的决策问题前进（时序建模）



长期记忆

( 资源收集 , 科技发展 )

短期记忆

( 战场侦查 , 技能衔接 )

多智能体协作

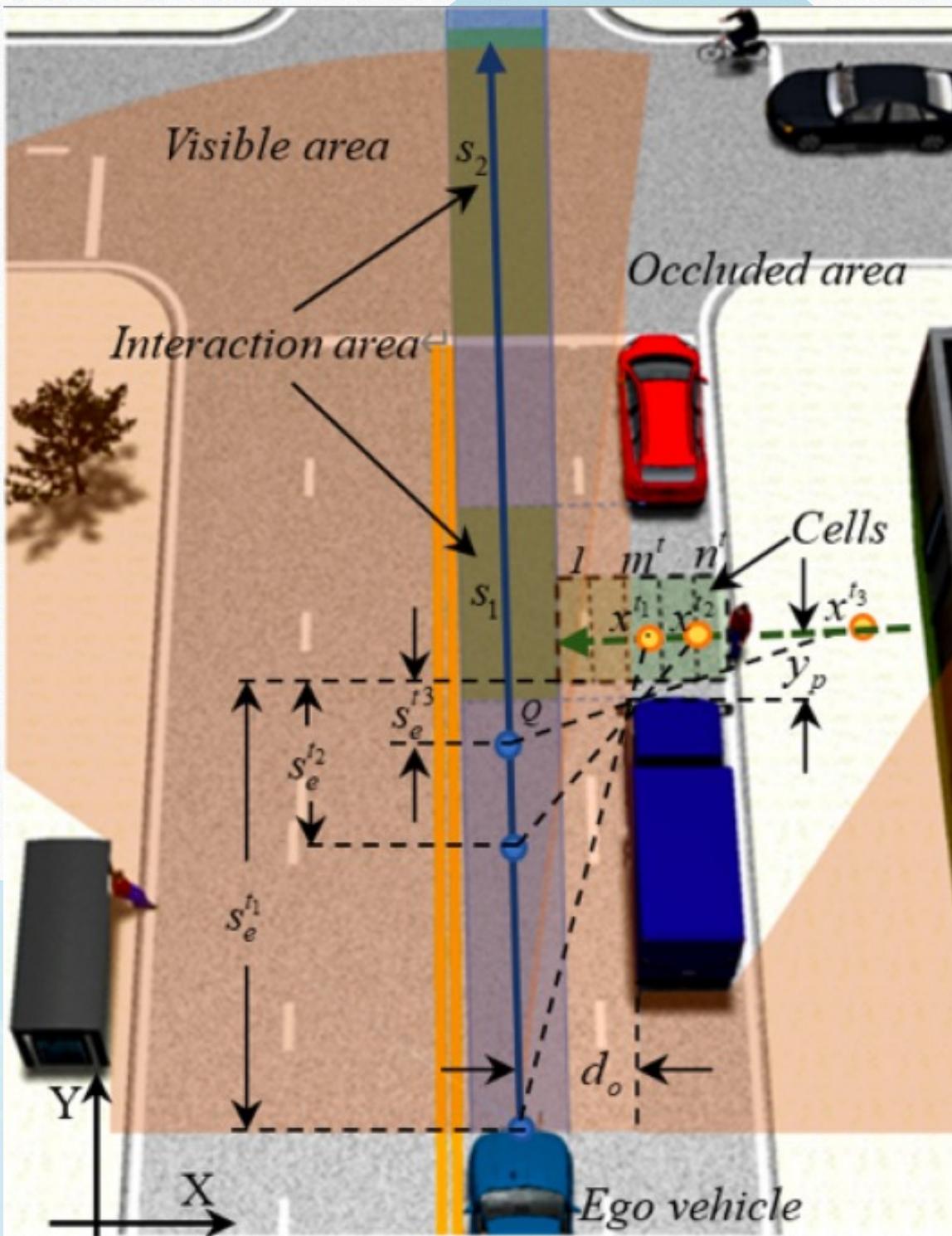
( 部队阵型 , 兵种配合 )

POMDP

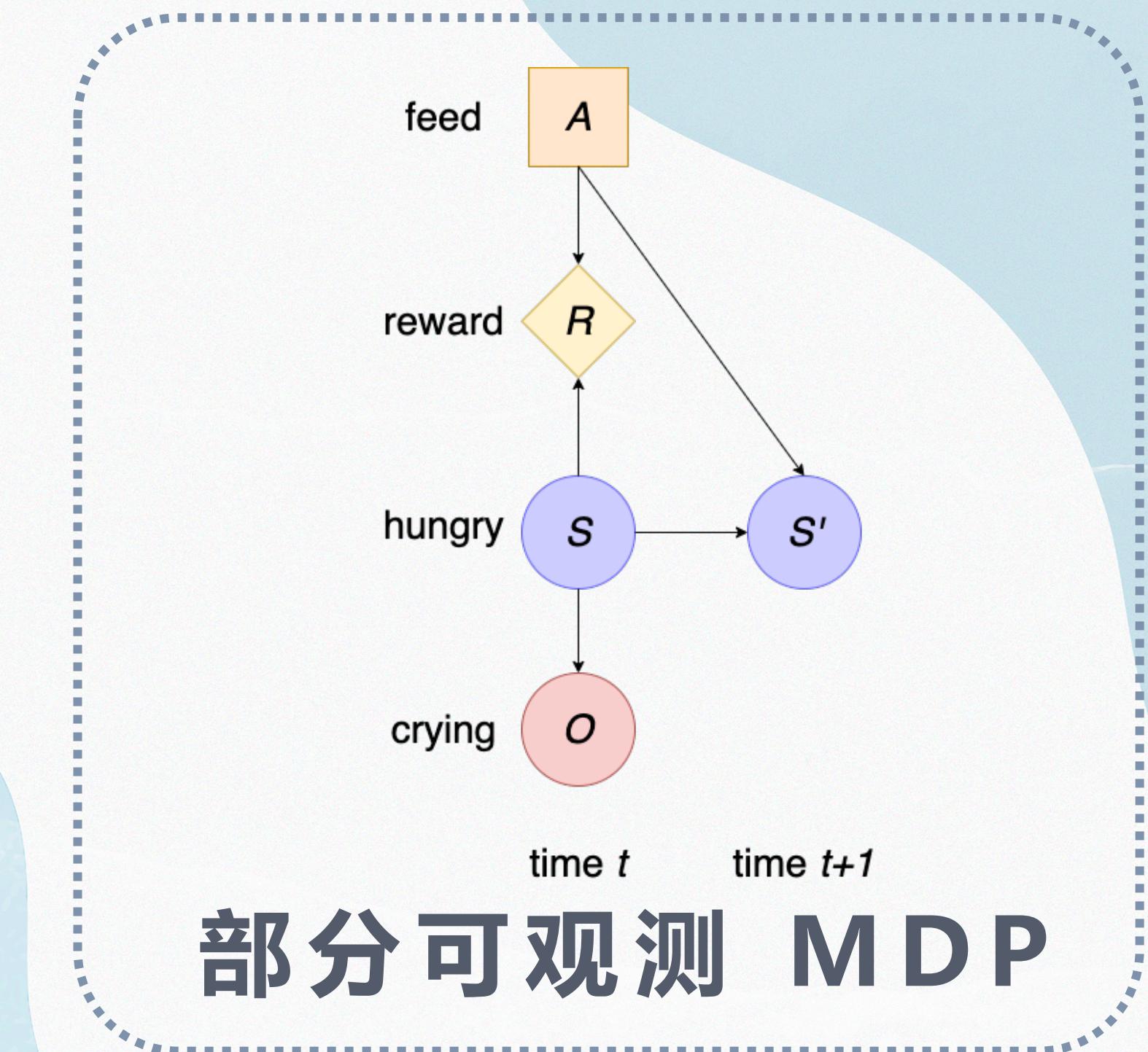
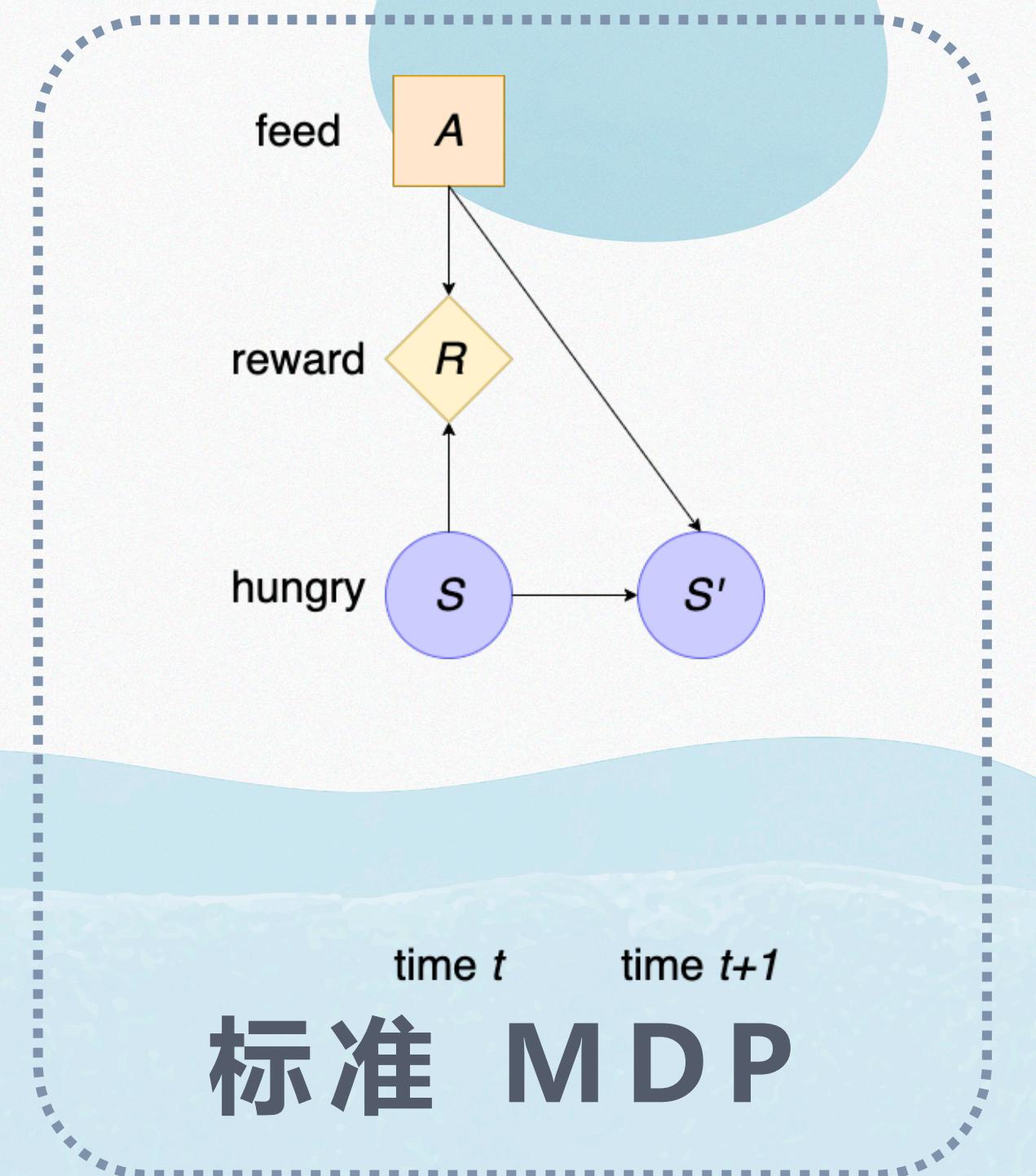
Overview

# 部分可观测

# 马尔可夫决策概述

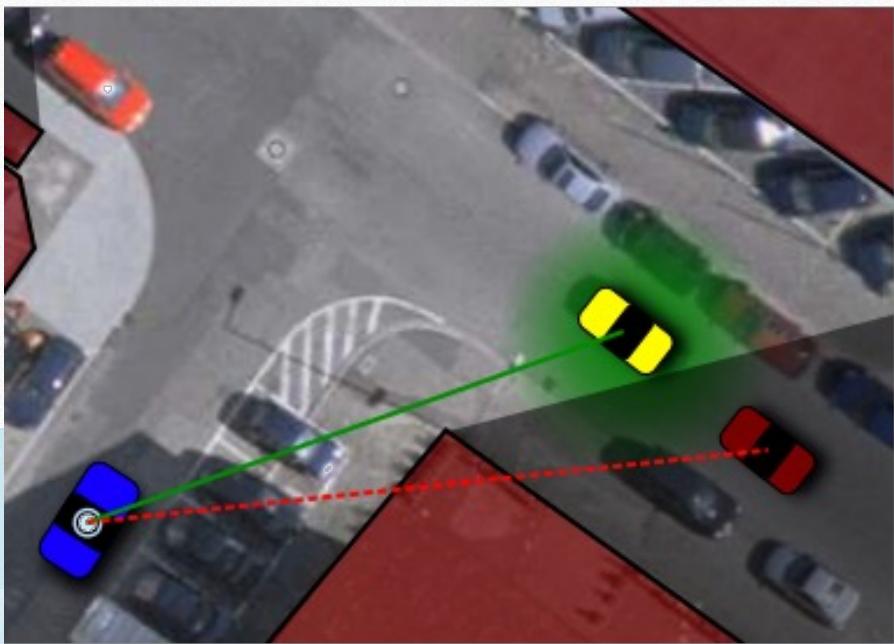


# 标准 MDP 的缺陷



# POMDP 无处不在

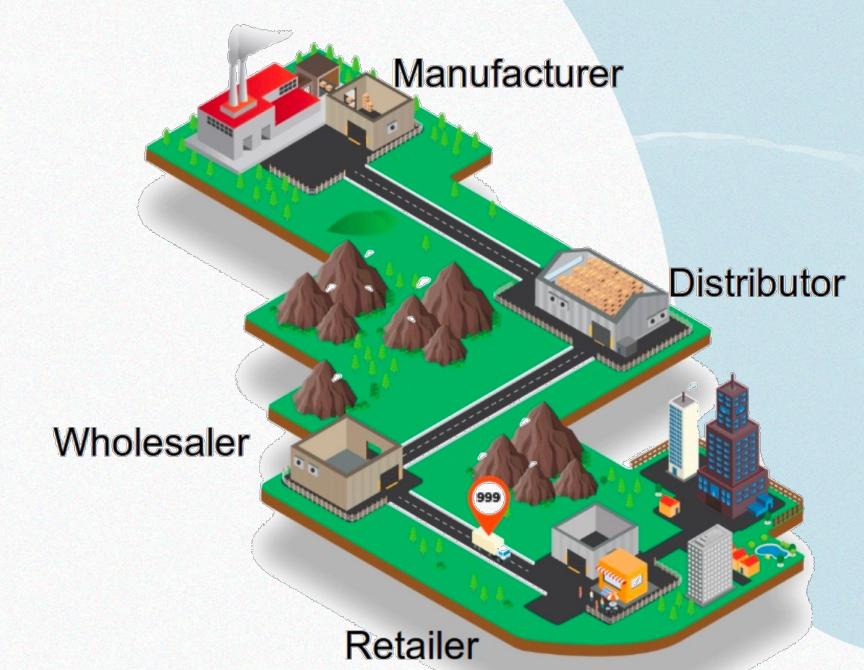
自动驾驶



机器人控制

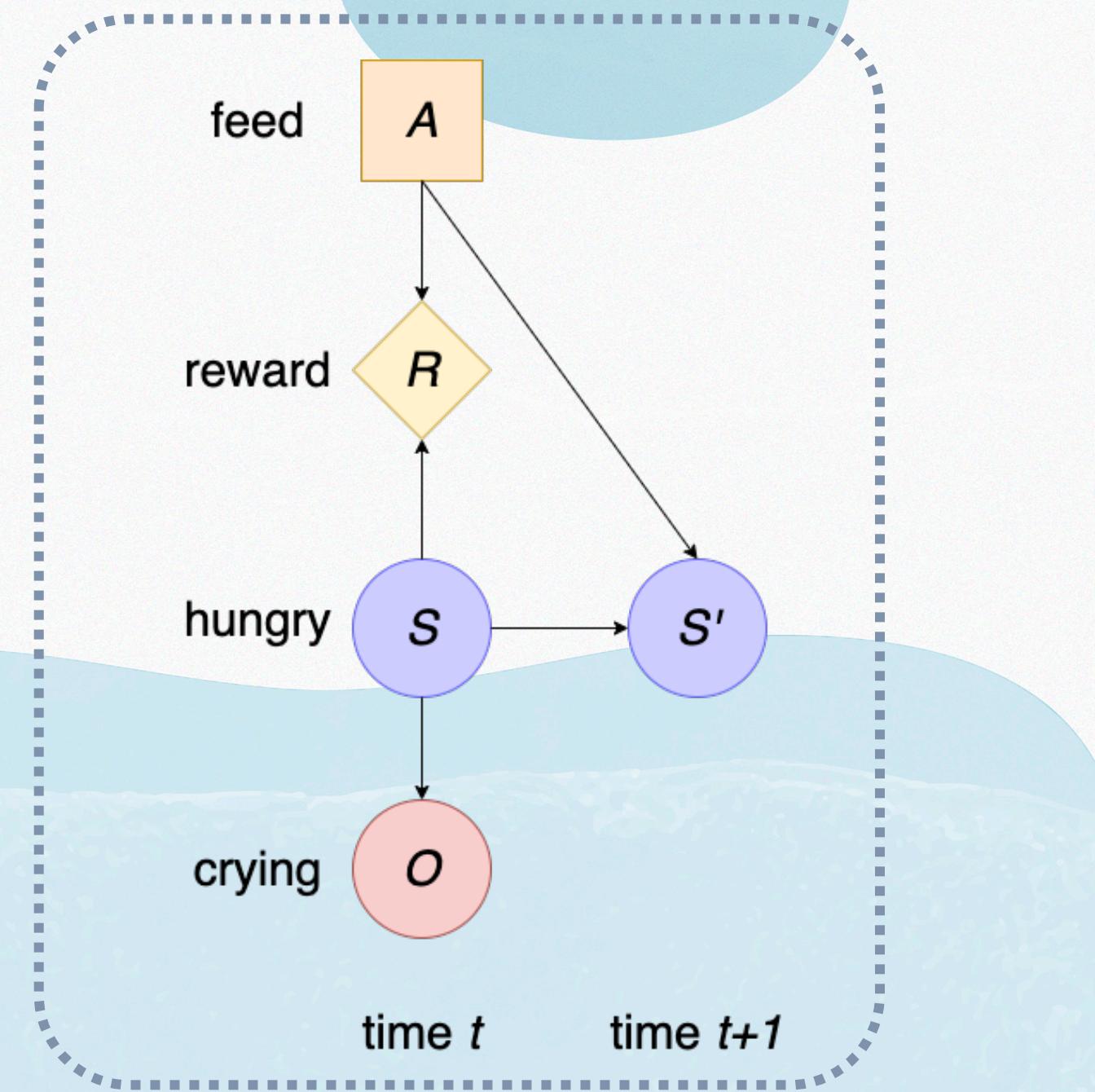


供应链优化

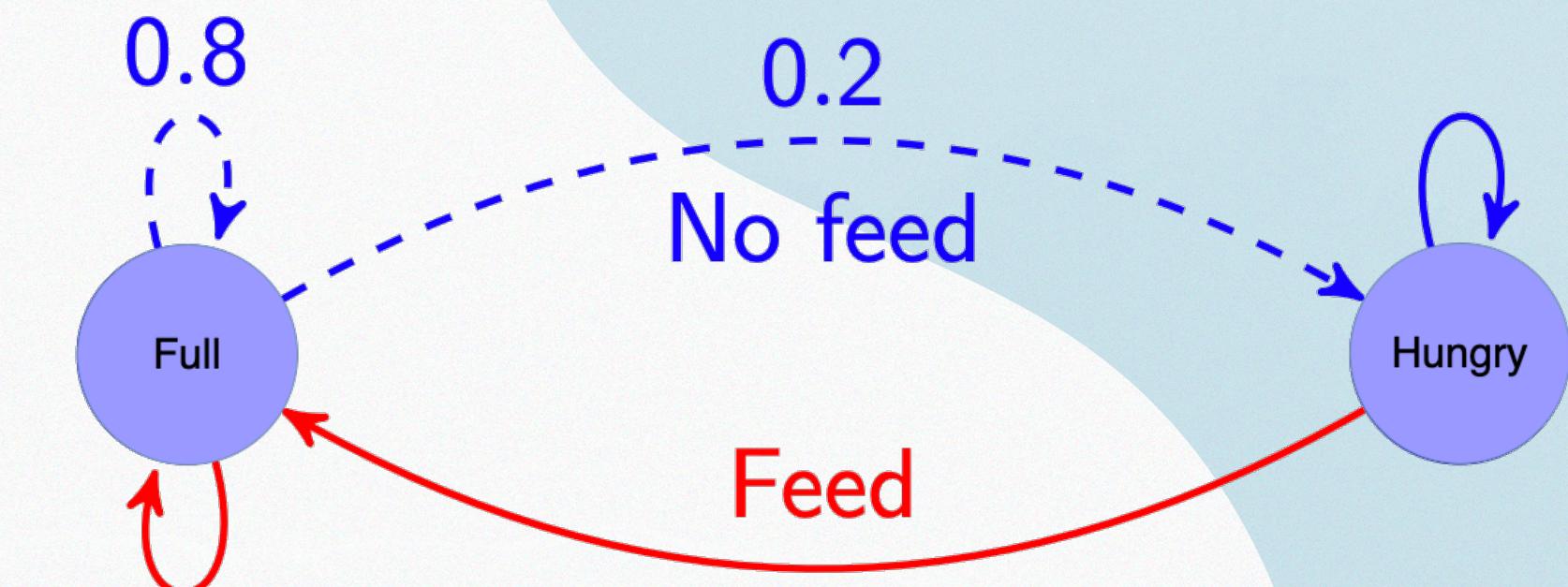


## Definition

# POMDP的定义



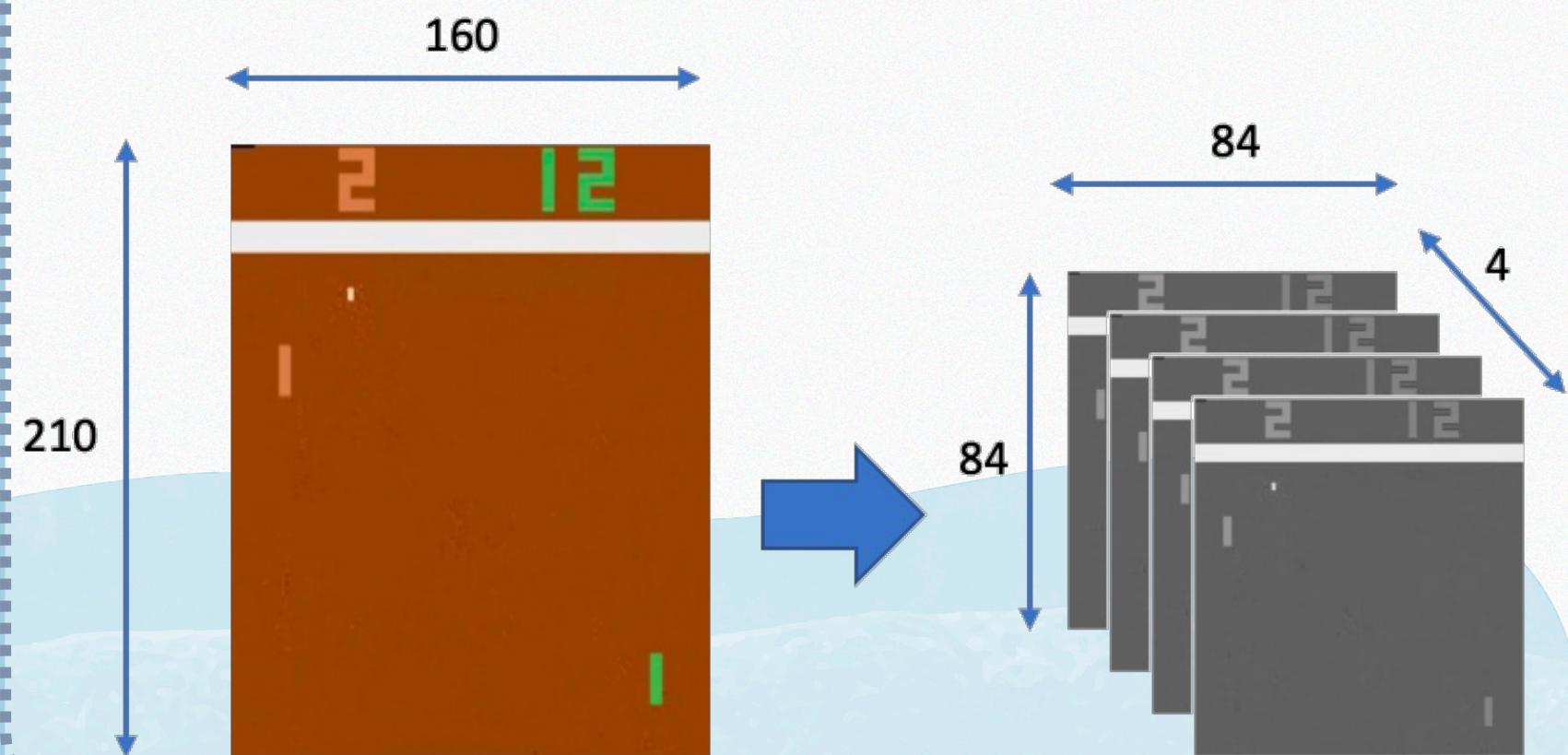
## Transition model



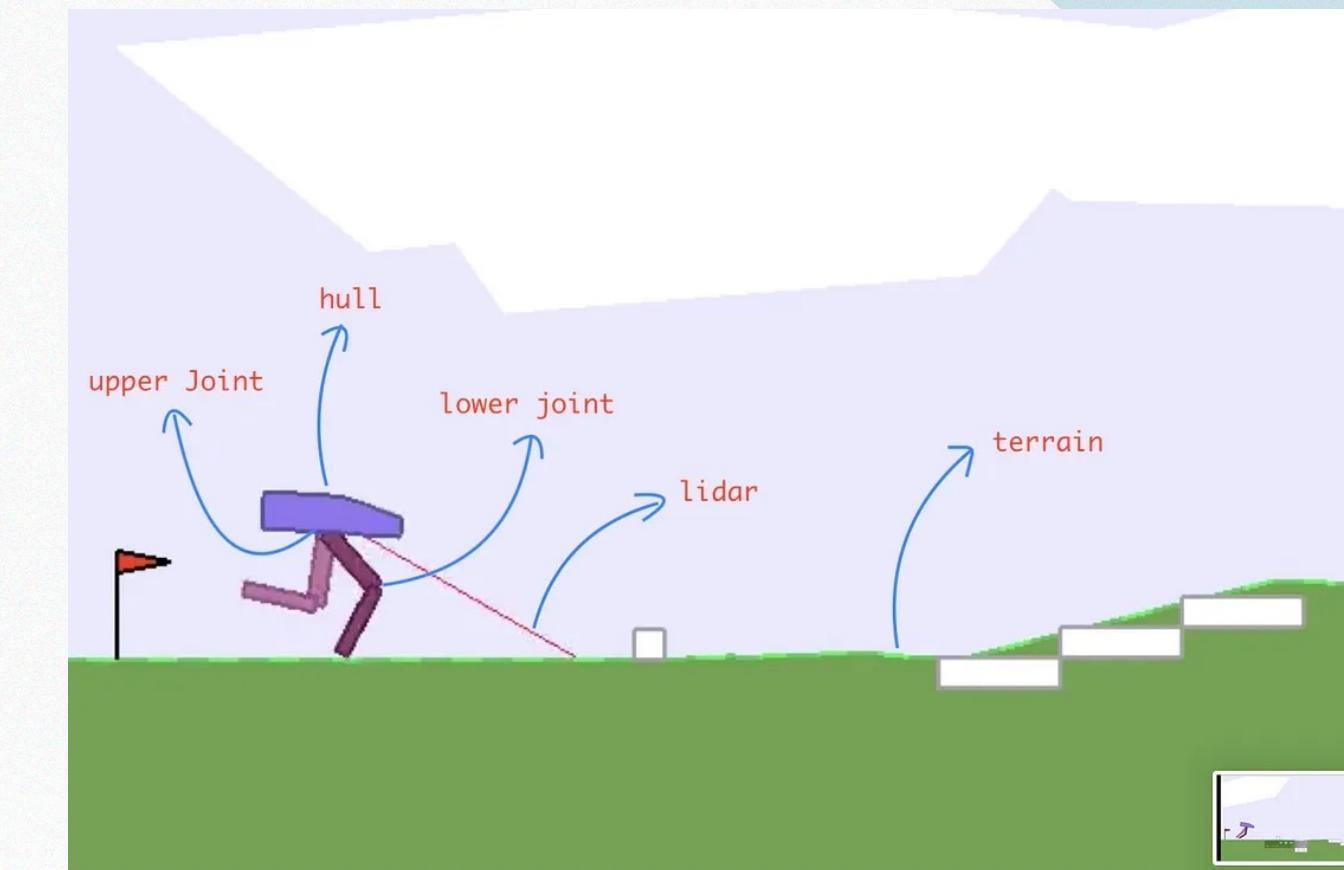
- **POMDP = MDP + 观察模型**
- 观察模型： $O(o|s)$  或  $O(o|s,a)$
- **决策行为只能根据观察信息  $O$  作出，因为无法得到准确的状态  $S$**

# Data POMDP的解法 (数据篇)

叠帧：补足运动信息

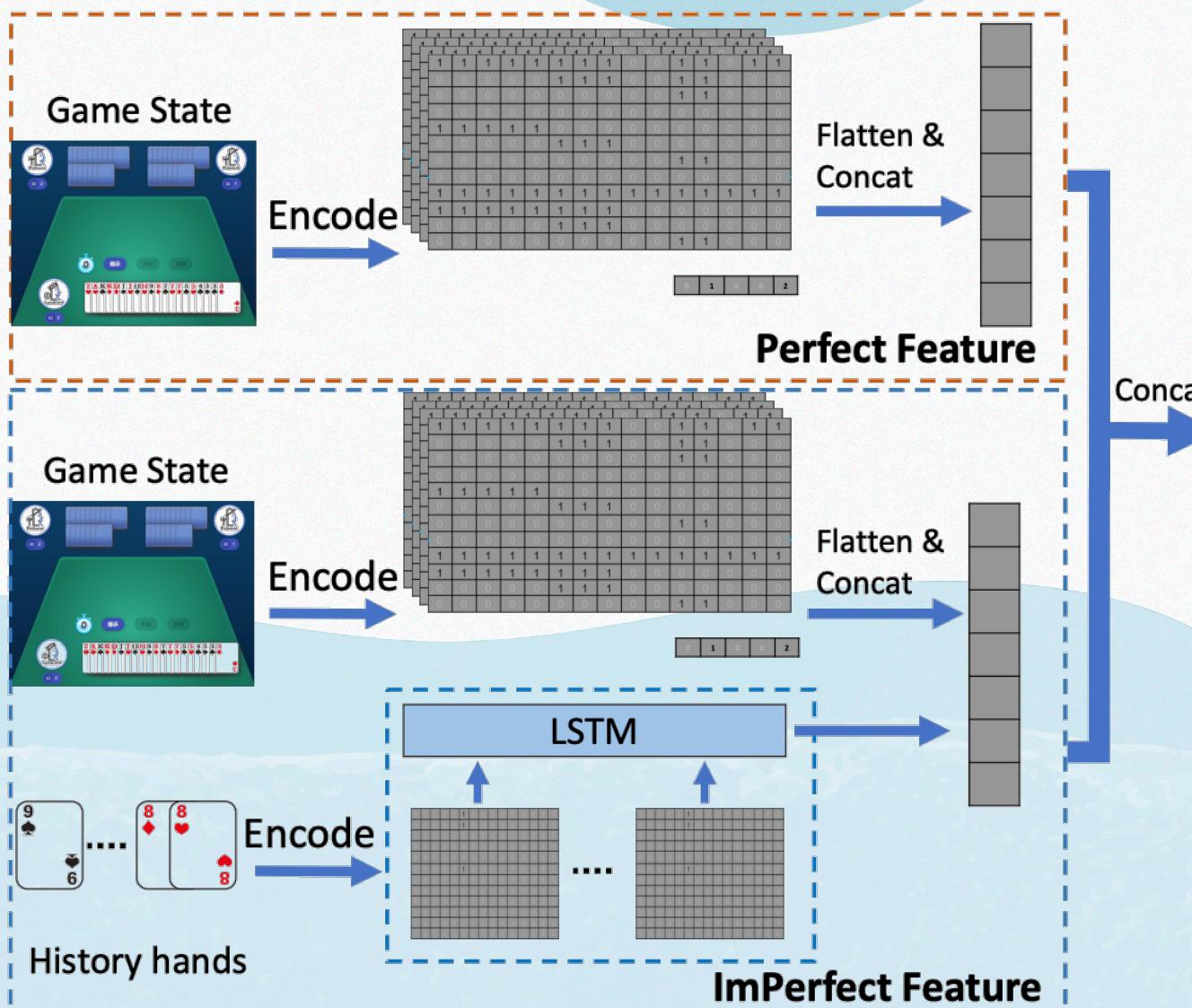


添加参数：补足物理特质信息（力学系数）

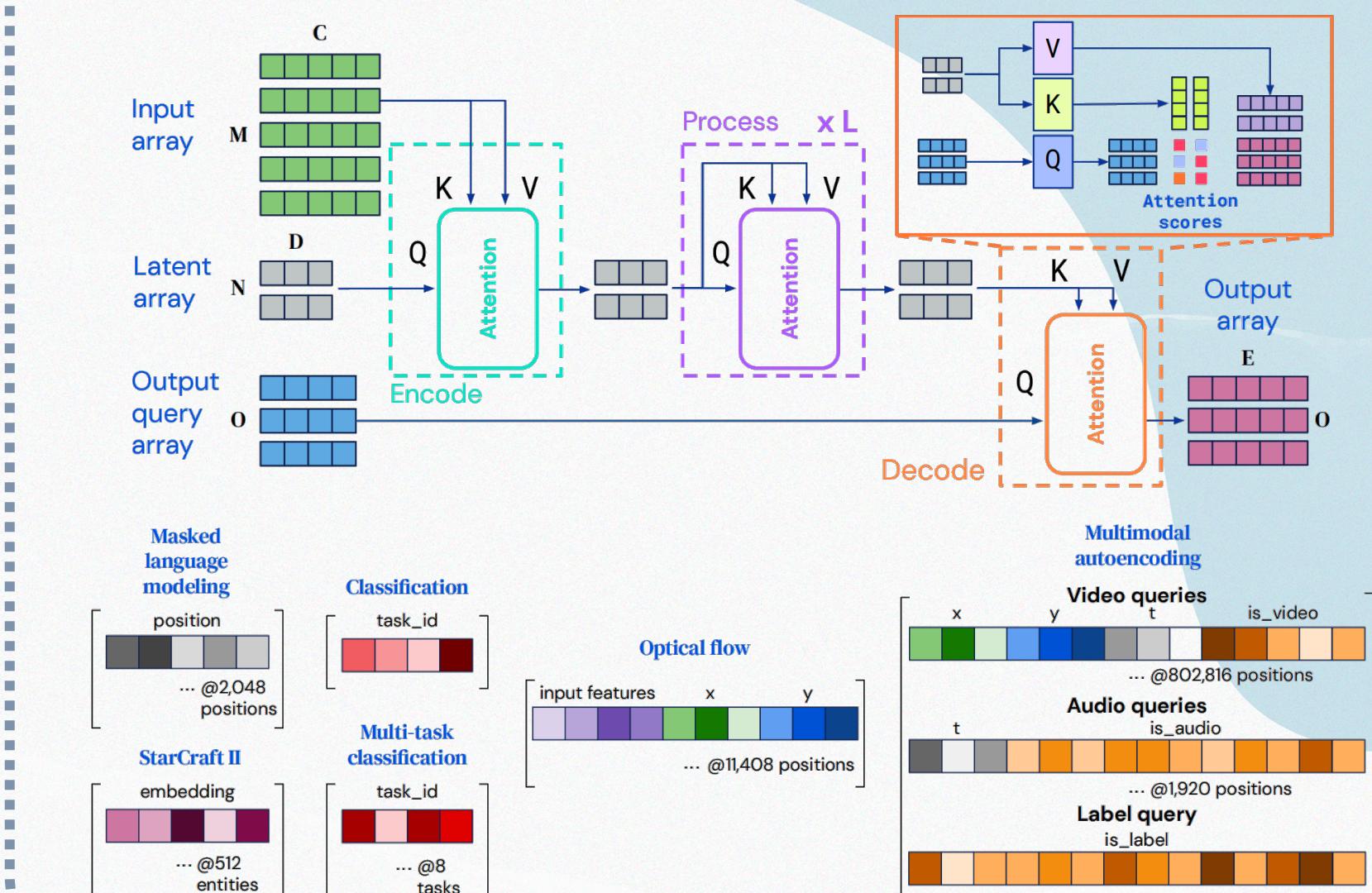


# NetwPOMDP的解法（网络篇）

完美的价值网络：补足博弈信息

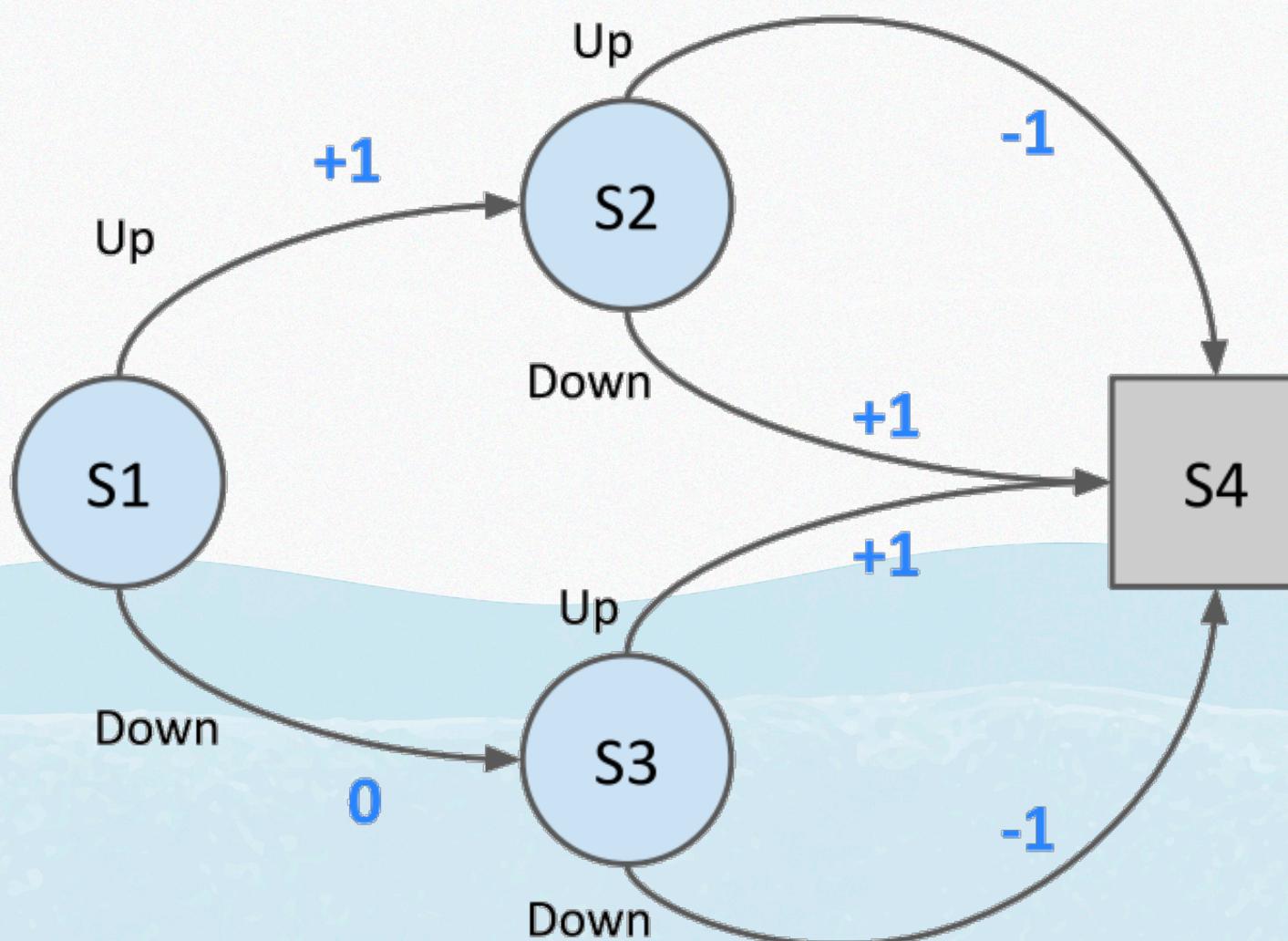


更强大的神经网络：减小预测误差

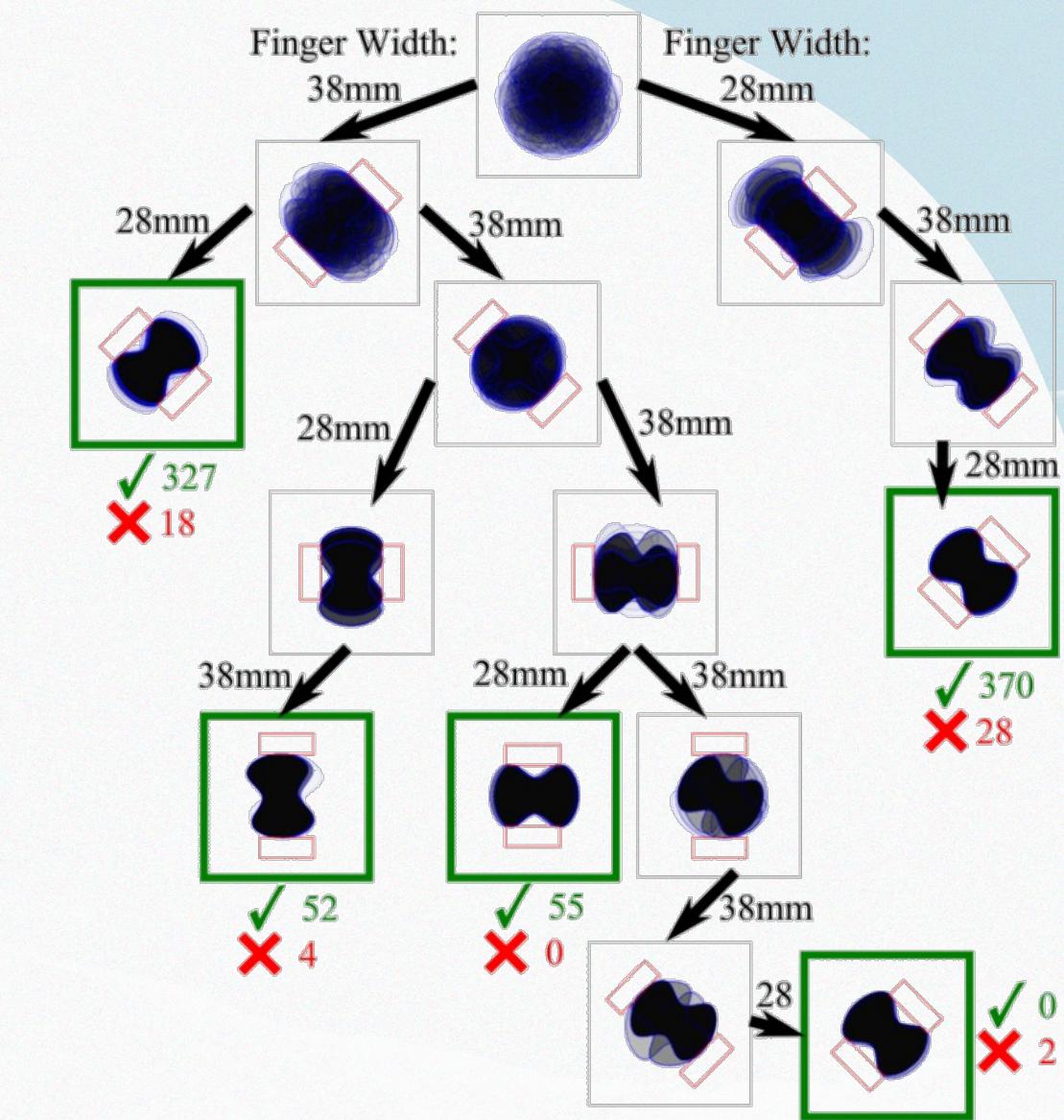


# Optin POMDP 的解法 (优化篇)

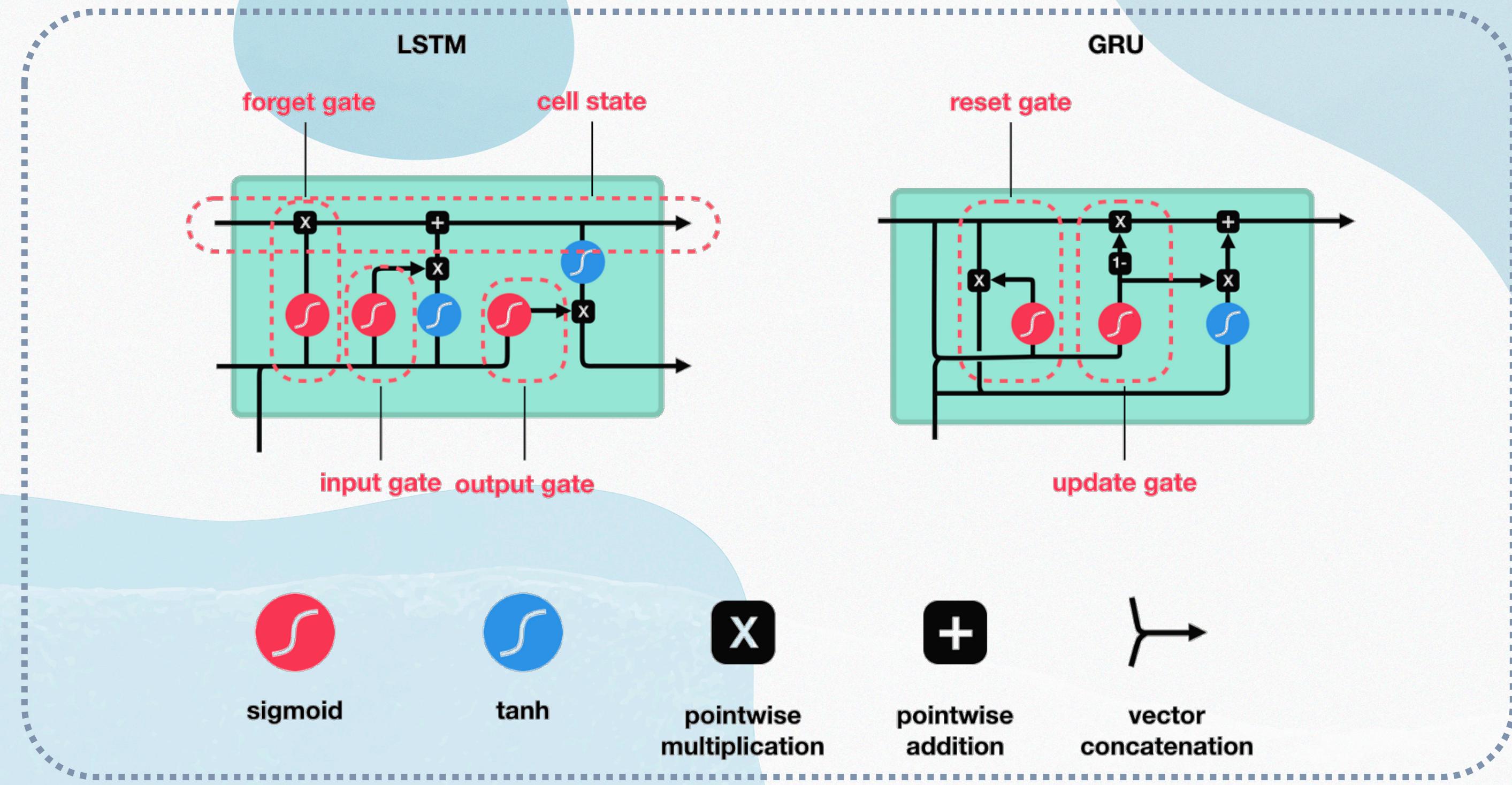
多用 N-step



善用 MCTS

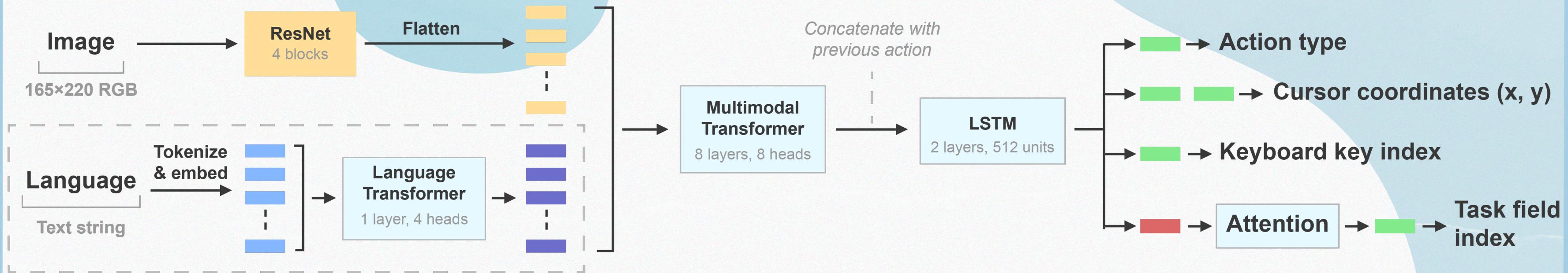


# 有记忆的 PPO：LSTM 篇



# 理论：PPO + LSTM

## ● 架构

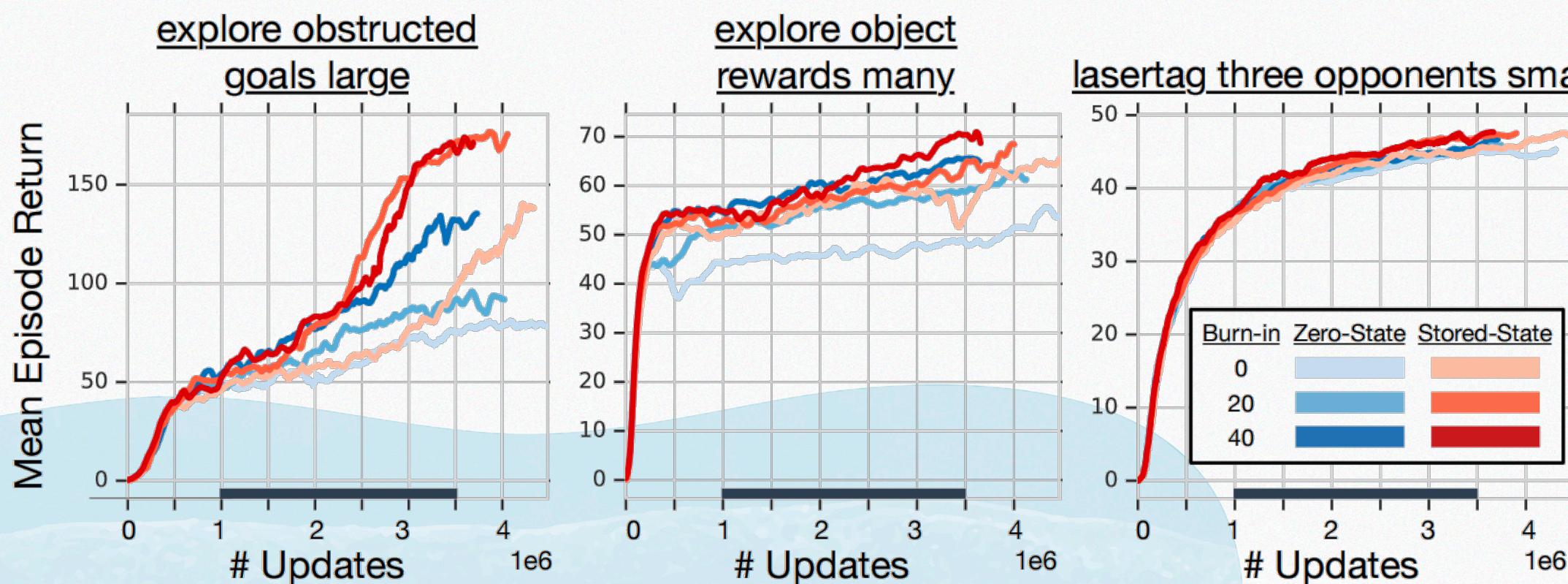


**编码器 ( encoder ) + 时序模块 ( LSTM ) + 预测器 ( head )**

- 编码器和预测器将观察（输入）和动作（输出）空间的复杂性内聚起来，  
LSTM 只在特征向量空间进行时序信息的记忆和建模

# 理论：PPO + LSTM

# 技巧



## 隐状态的传递 + Burn In 机制

- 数据收集时需额外存储隐状态
- 训练开始加载隐状态并 Burn In
- Burn In 即取序列最开头的一部分数据，用训练端最新的网络沿着时间步向前展开，产生更加新的隐状态，但不参与梯度计算

# 理论：PPO + LSTM

# 技巧

反应型环境 (answer quantitative question)

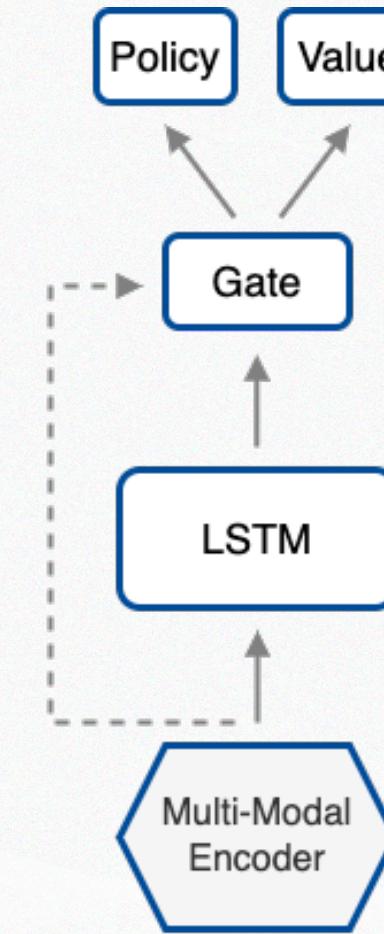


记忆型环境 (obstructed goals large)



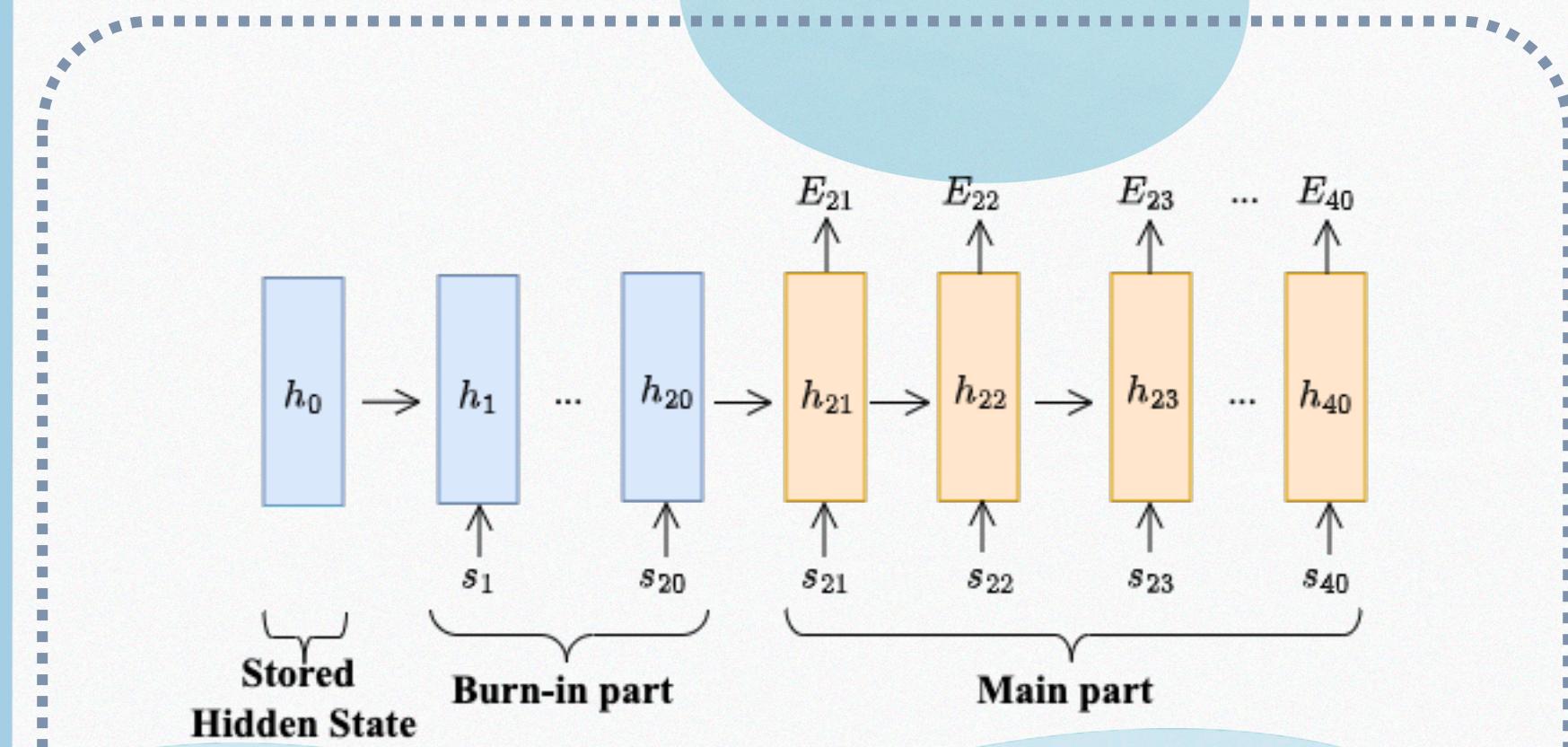
Residual Link :

兼顾反应型与记忆型环境



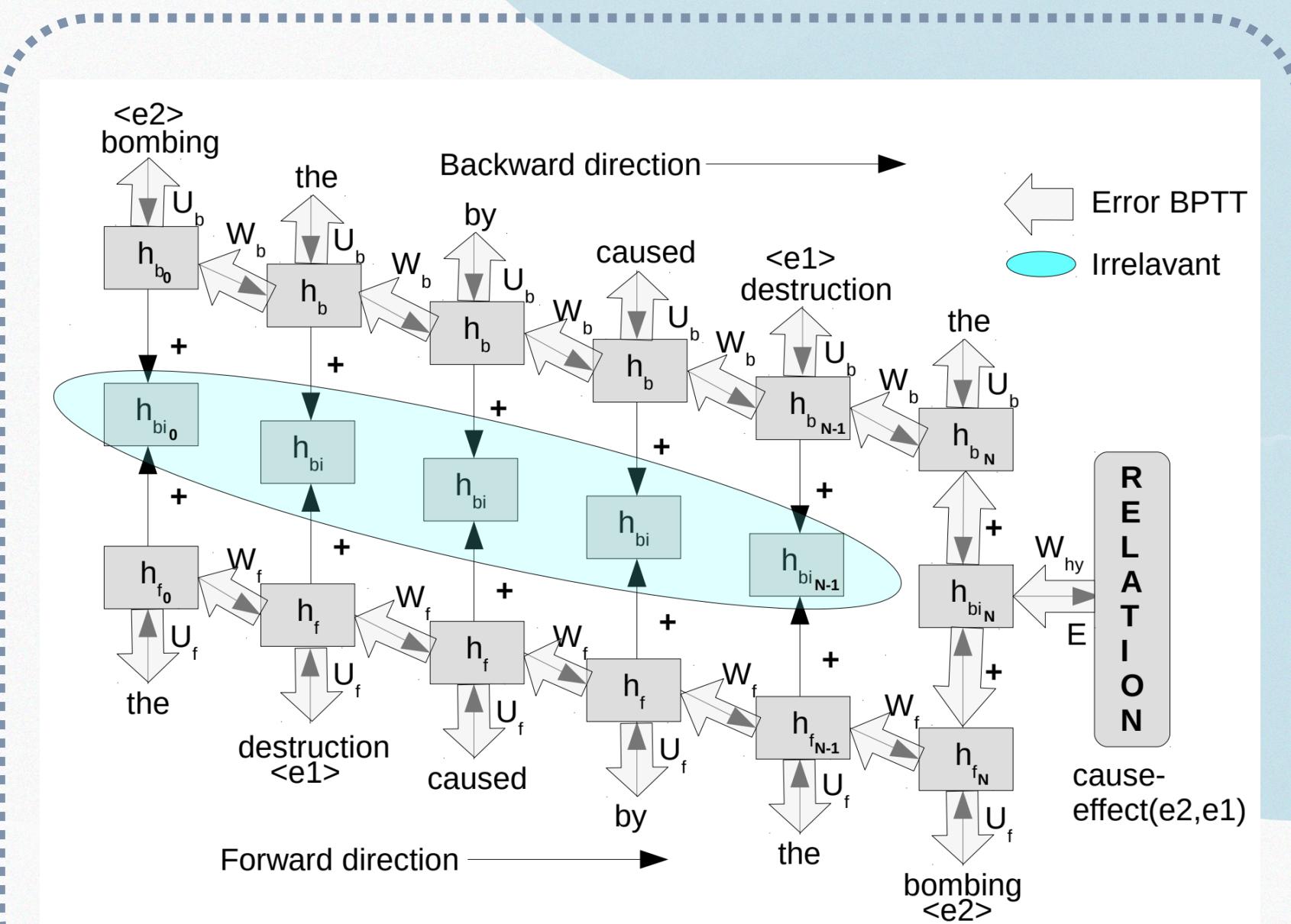
# 理论：PPO + LSTM

● 优化



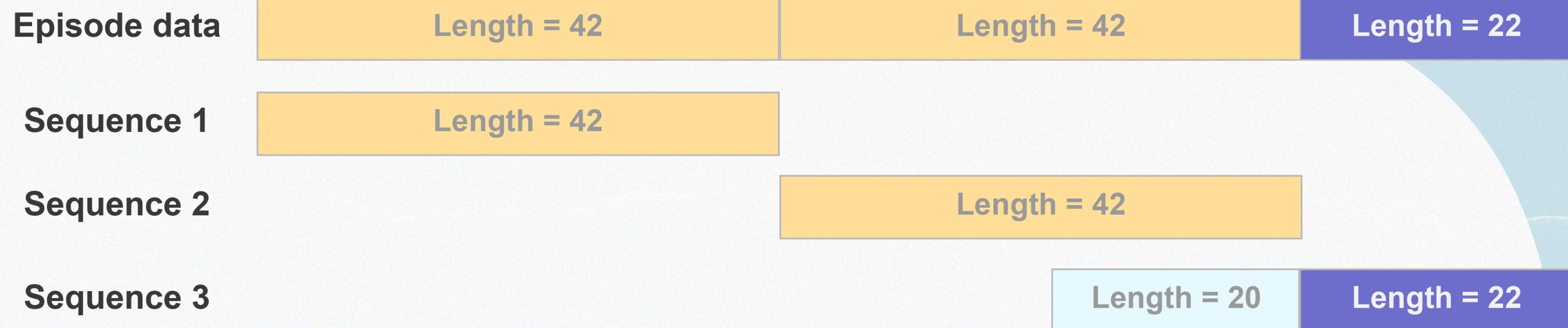
- 预加载和 Burn In 的隐状态都不参与梯度计算
- 使用隐状态和观测信息编码获得当前帧的

Embedding 和传递给下一帧的隐状态

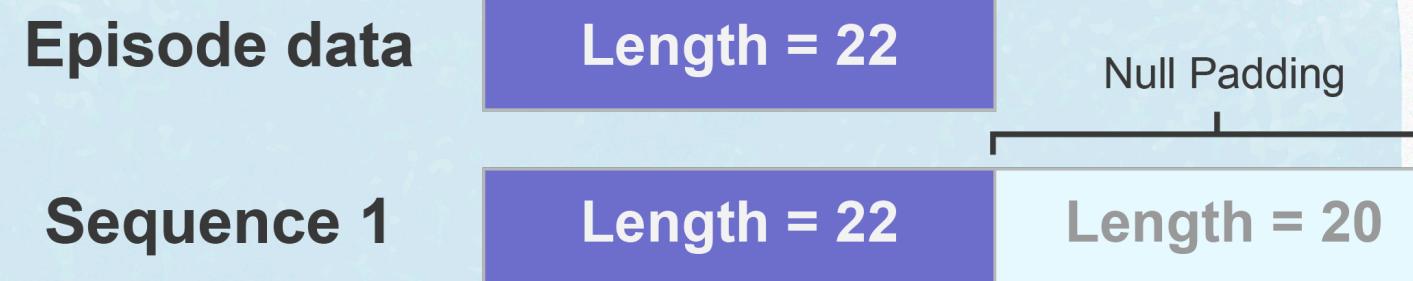


# 代码：如何准备统一的轨迹数据

**Case 1: Episode length > Sequence length**



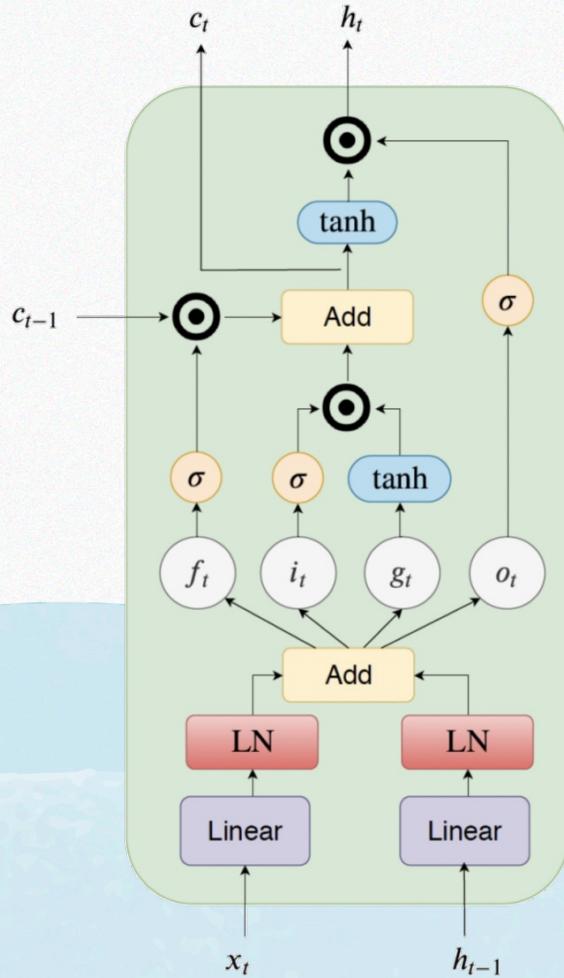
**Case 2: Episode length ≤ Sequence length**



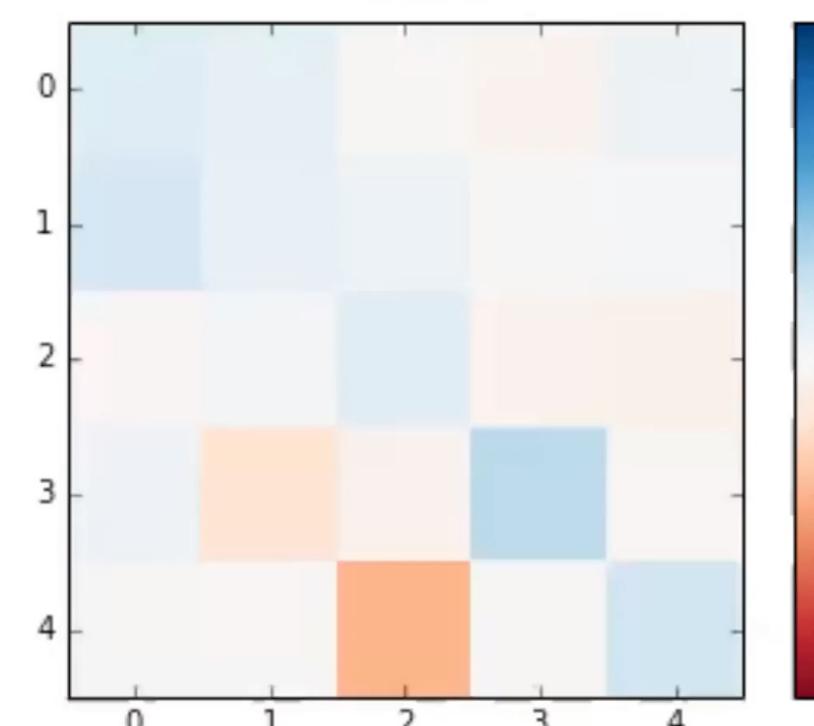
Note: For null padding,  
Observation = zeros\_like (obs), Reward = 0  
Done = True, Hidden State = None

# 代码：RL中应用LSTM的三重境界

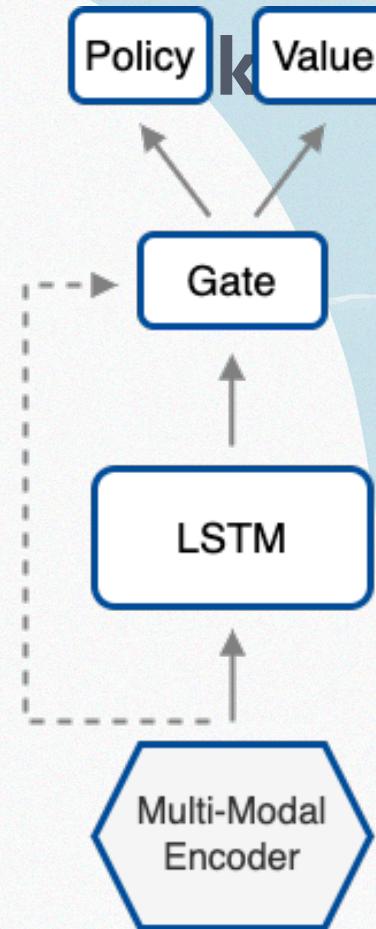
LSTM + LayerNorm



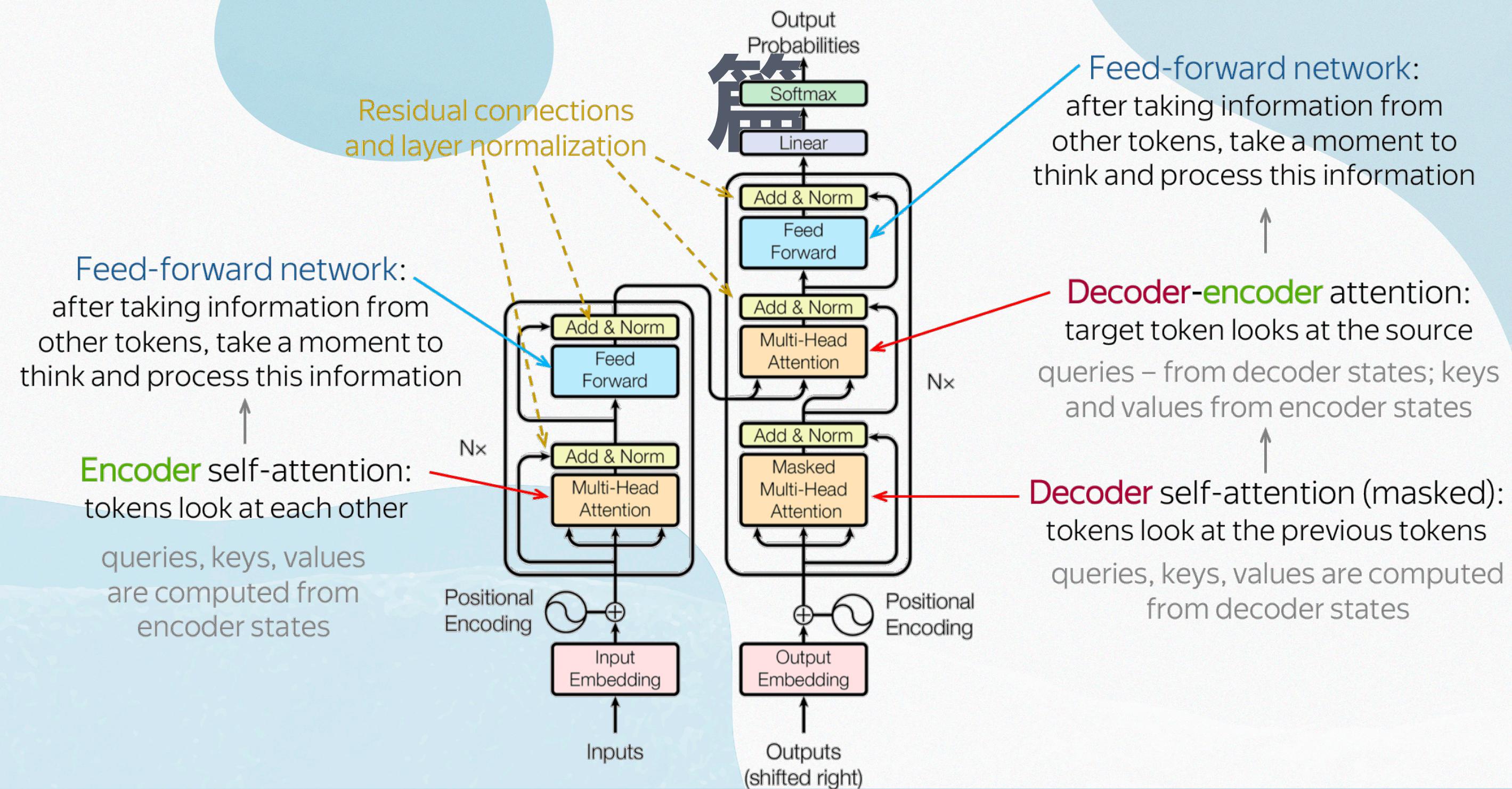
LSTM +



LSTM + Residual



# 有记忆的 PPO : Transformer



# 理论：Transformer + RL 的劣势

## Transformer 的训练

需要一系列复杂的技巧来优化其性能

- 复杂的学习率调整策略（如余弦衰减）
- 复杂的梯度裁剪技巧
- dropout 的各种变体
- 特定的权重初始化方案
- 特殊的 normalization 网络层和使用位置
- 位置编码的各种技巧

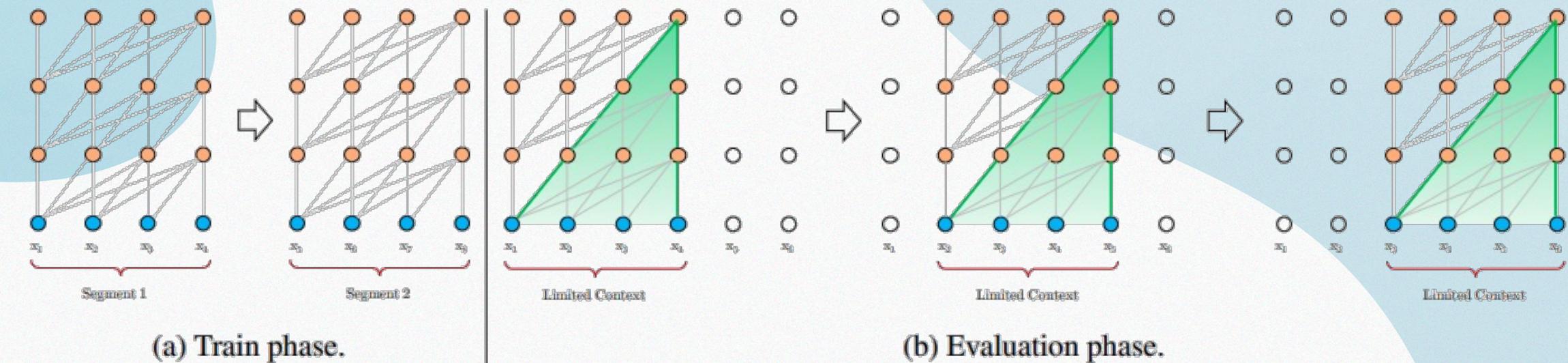
## 强化学习应用 Transformer

还具有监督学习所没有的一些挑战，其中包括

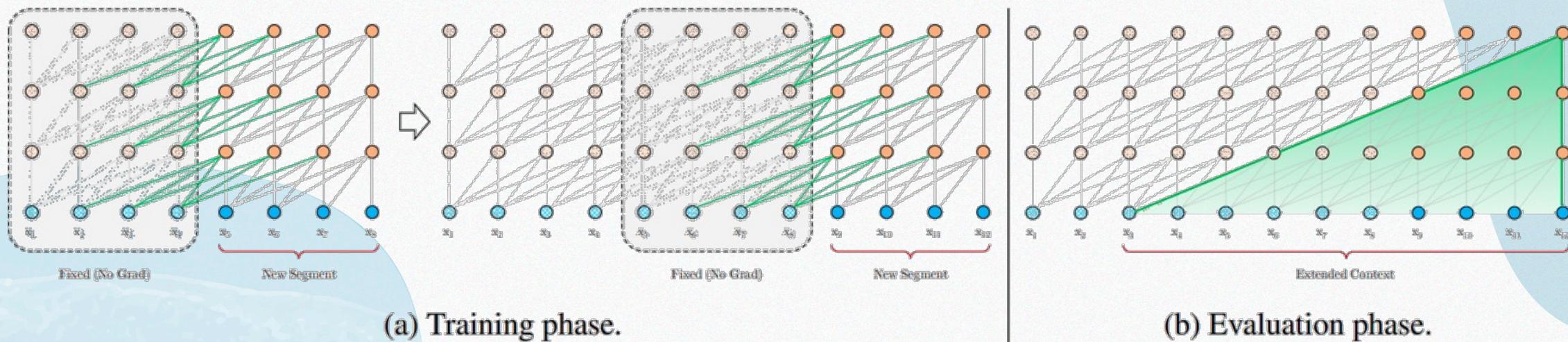
- 长期的时间依赖性
- 环境奖励的稀疏性
- 平衡强化学习的探索和利用

# 理论：Transformer + RL 的优势

标准 Transformer

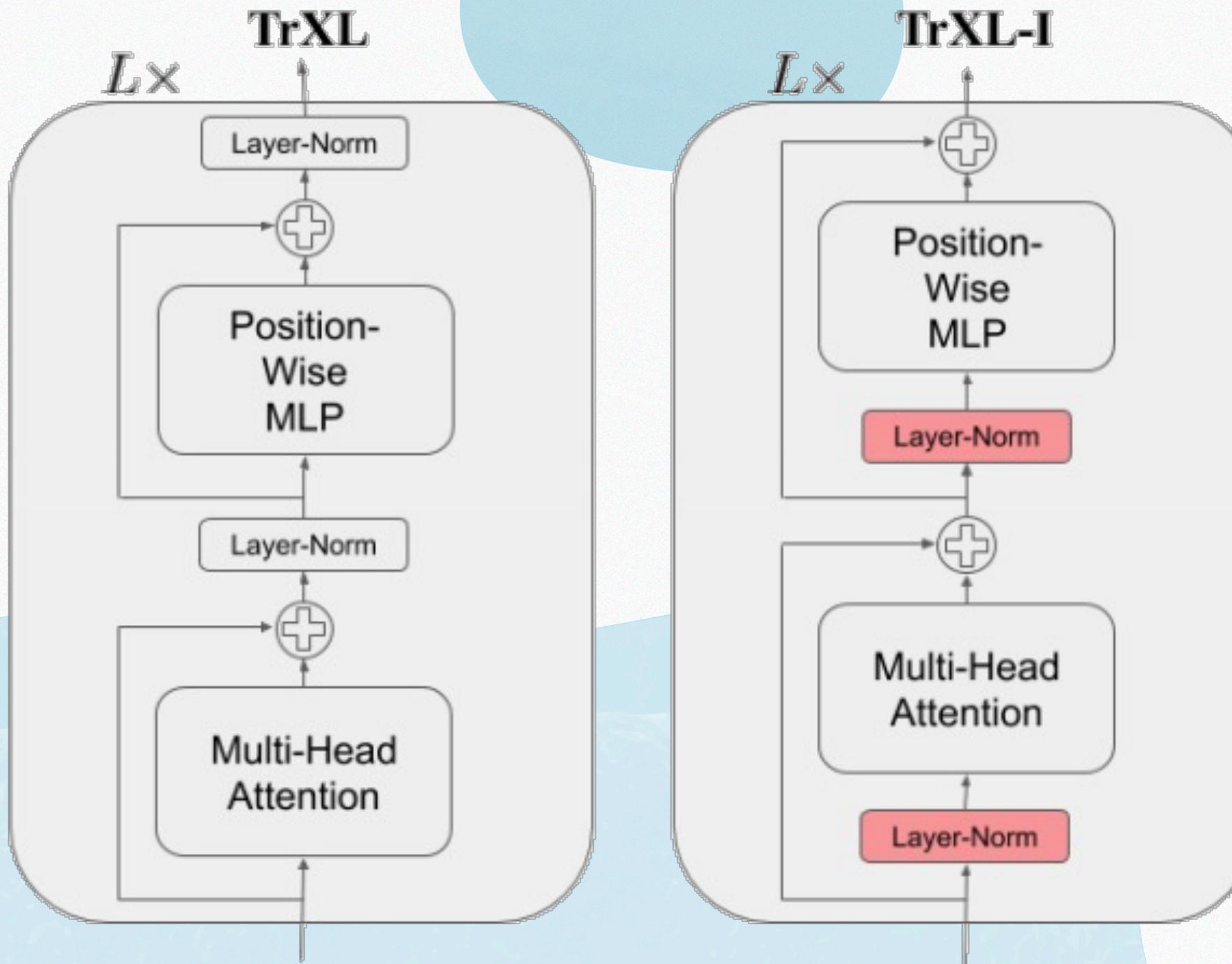


Transformer-XL



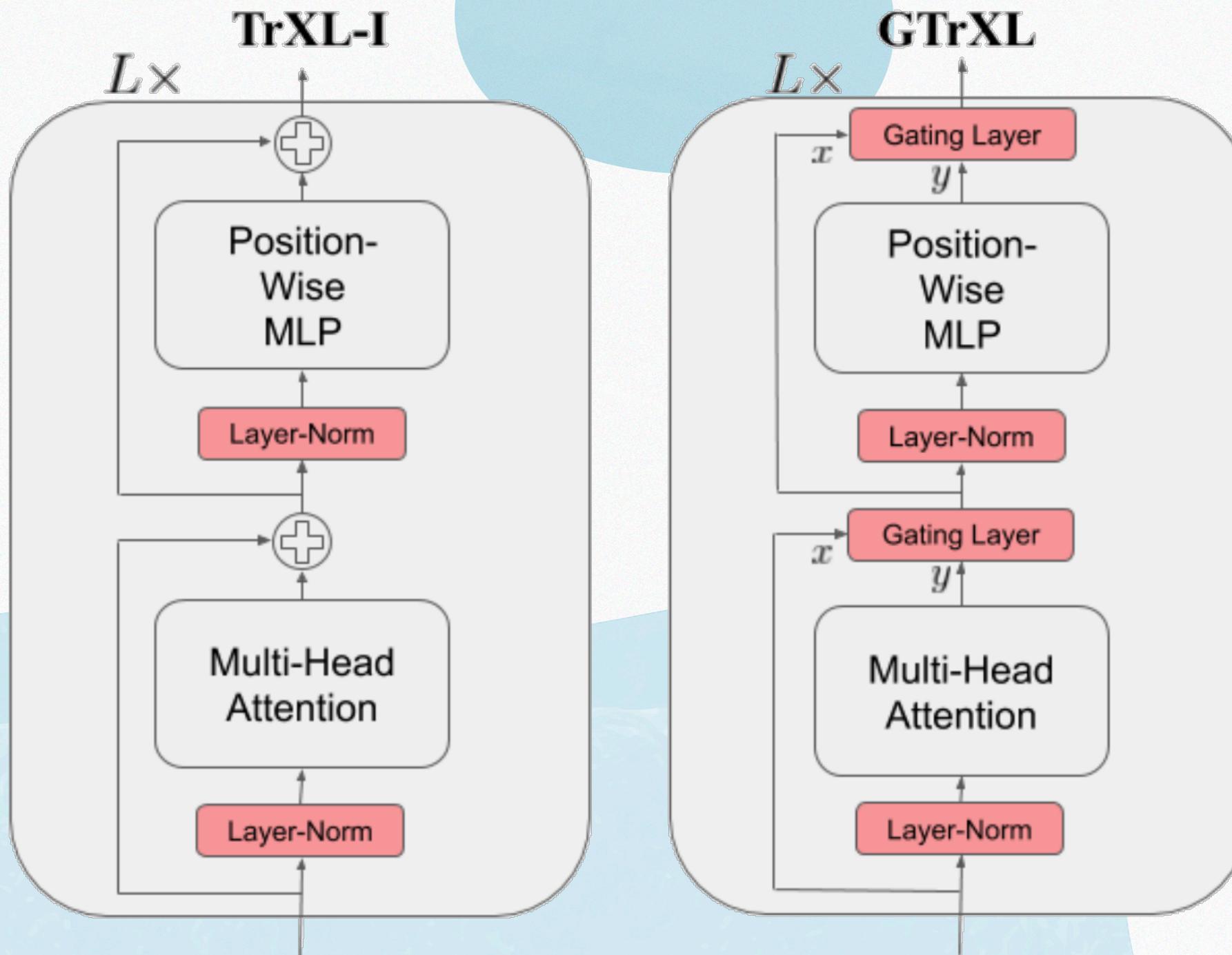
- Transformer 相比 LSTM 可以处理更长期的时序依赖，这对一些长期记忆型决策尤为重要
- Transformer 相比 LSTM 更适应 RL 这种在线学习，高方差的优化过程

# 理论：PPO + GTrXL + Layer-Norm LN



- 思路：让网络初始时更接近马尔可夫假设，即仅用当前状态进行决策，忽略其他时间步的信息
- 原理：让智能体先学习一个符合马尔可夫假设的、较为简单的策略，再学习更复杂的能处理时序信息的策略
- 例子：一个人应当先学会如何走路，再记住走过的路径，更好地规划路径

# 理论：PPO + GTrXL



$$r = \sigma(W_r^{(l)}y + U_r^{(l)}x)$$

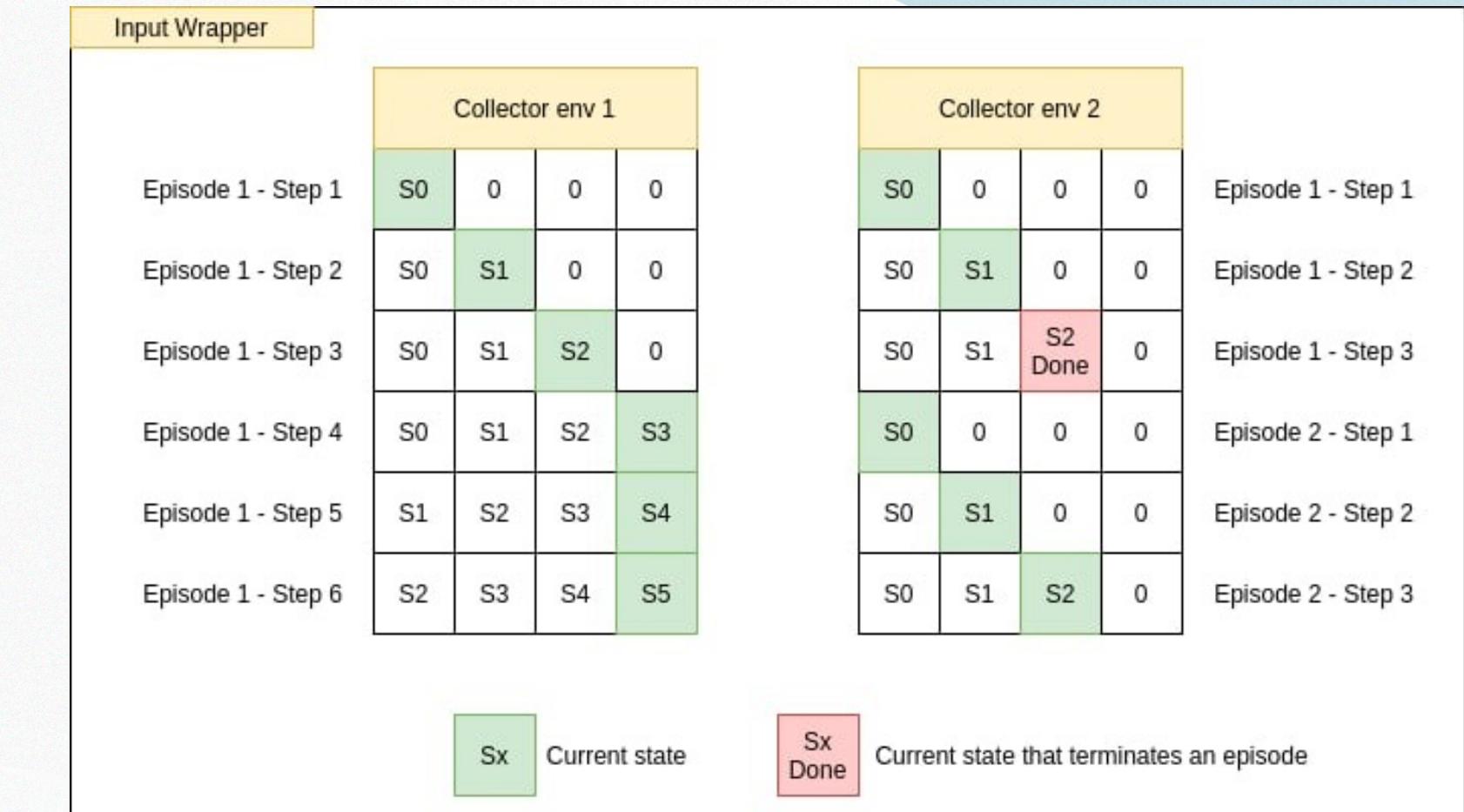
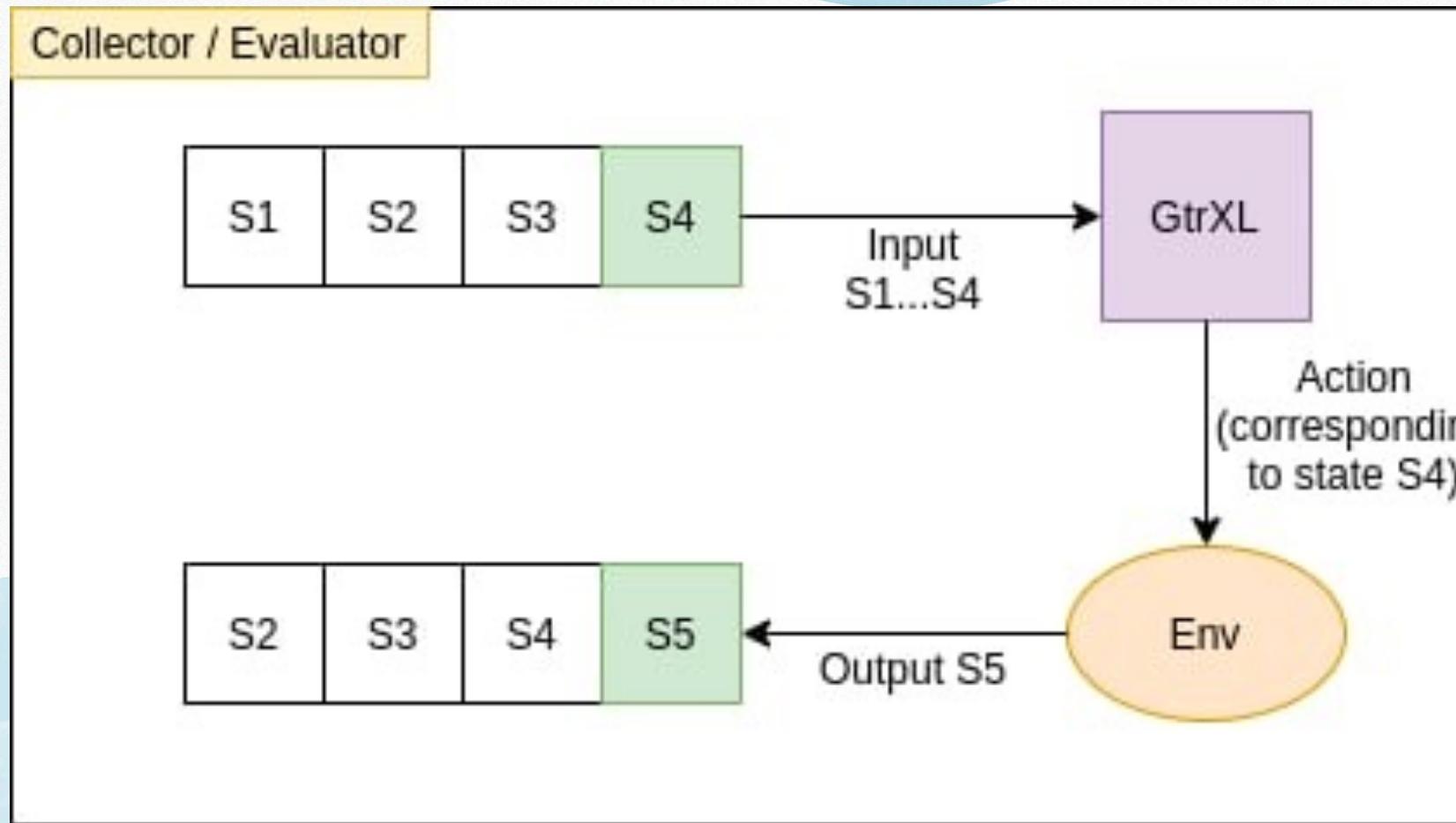
$$z = \sigma(W_z^{(l)}y + U_z^{(l)}x - b_g^{(l)})$$

$$\hat{h} = \tanh(W_g^{(l)}y + U_g^{(l)}(r \odot x))$$

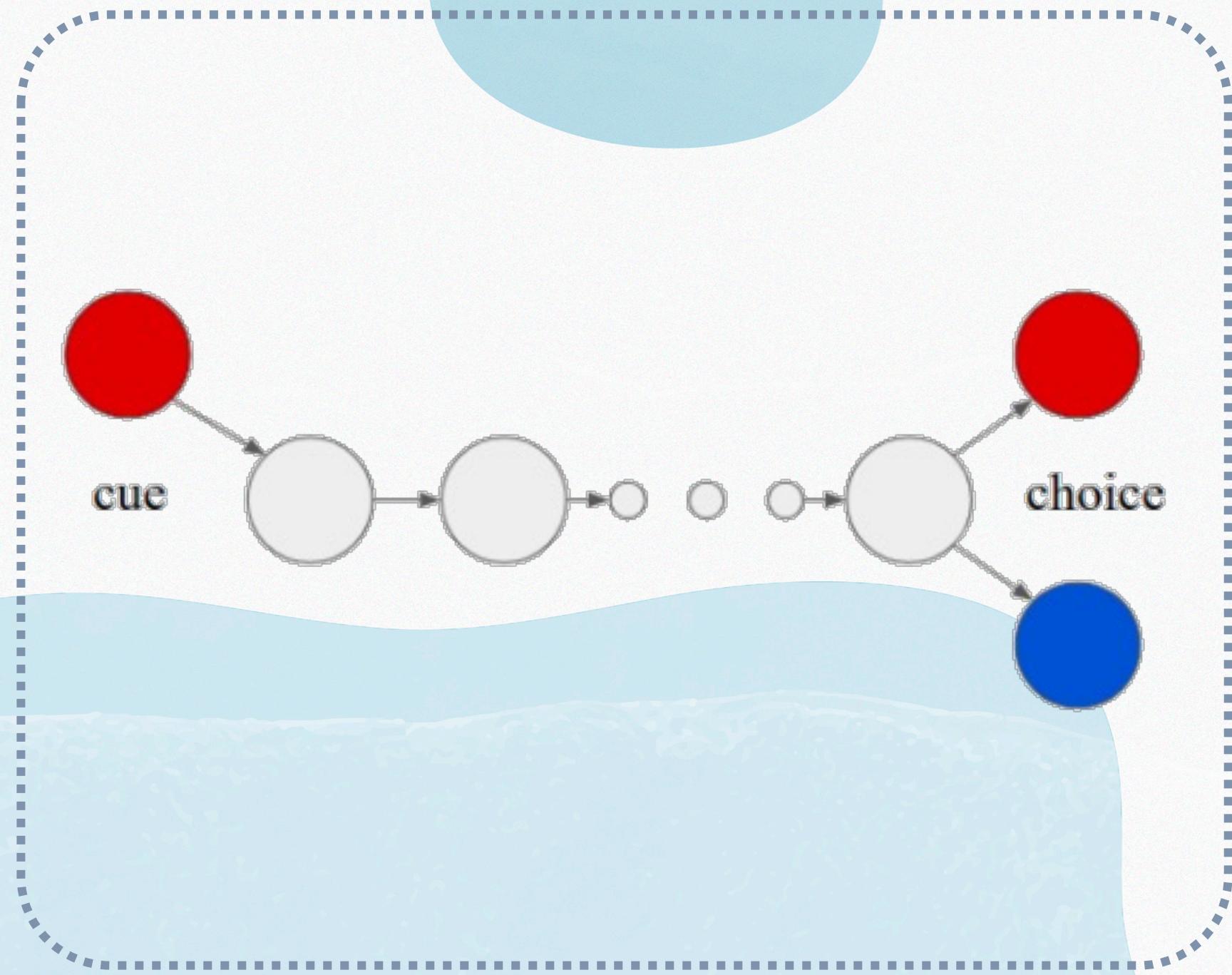
$$g^{(l)}(x, y) = (1 - z) \odot x + z \odot \hat{h}$$

- 思路：通过门控机制更灵活地组合反应型策略和记忆型策略
- 技巧：通过对 bias 的特殊初始化，让 Gate 初始化为恒等映射，即在原版设计的基础上学习

# 代码：GTrXL + PPO 应用小妙招



# 实践：PPO + memory len



## 任务目标

- 测试智能体的长期记忆能力，跨越多步记住一个比特的信息（cue）

## MDP元素

- 观察：时间步和当前信息（随机）
- 动作：2维离散动作，选择 +1 or -1
- 奖励：稀疏奖励，即智能体是否最终选择正确的cue

# 实践：PPO + memory len

Policy	Memory len (N)	收敛性	收敛所用环境交互步数 ( Env Steps )
PPO+LSTM	12	Yes	5.8k
PPO+LSTM	30	Yes	78k
PPO+LSTM	50	Yes	383.5k
PPO+LSTM	100	No	>90M未收敛
PPO+GTrXL	12	Yes	2.1k
PPO+GTrXL	30	Yes	7k
PPO+GTrXL	50	Yes	11.4k
PPO+GTrXL	100	Yes	32.3k

# 总结：PPO + 时序建模

小节	算法要点	代码和实践要点
POMDP 概述	<ul style="list-style-type: none"> <li>• POMDP 的定义和特点</li> <li>• 从数据编码角度应对 ( 叠帧 )</li> <li>• 从网络设计角度应对 ( 完美价值网络 )</li> <li>• 从优化方法角度应对 ( N-step )</li> </ul>	/
PPO + LSTM	<ul style="list-style-type: none"> <li>• 整体网络架构设计</li> <li>• 隐状态维护更新问题</li> <li>• 反应型和记忆型决策</li> <li>• 时序展开如何优化</li> </ul>	准备统一的轨迹数据 RL 中应用 LSTM 的三重境界
PPO + Transformer	<ul style="list-style-type: none"> <li>• Transformer + RL 的优劣势</li> <li>• LayerNorm 的位置之争</li> <li>• 门控 ( Gate ) 调节机制</li> </ul>	Transformer 的记忆模块实现 memory len 环境对比测试

## 下节预告

### (六) 群体之趣：统筹多智能体

---

- Dec-POMDP：多智能体协作的特殊之处
- MAPPO：套用 PPO —— 建立基石
- HAPPO：化用 PPO —— 追寻本源