# Towards Large-Scale In-Context Reinforcement Learning by Meta-Training in Randomized Worlds

*Fan Wang, *Pengtao Shao, Yiming Zhang, Bo Yu, Shaoshan Liu,
Ning Ding, Yang Cao, Yu Kang, Haifeng Wang

中国科学技术大学 University of Science and Technology of China

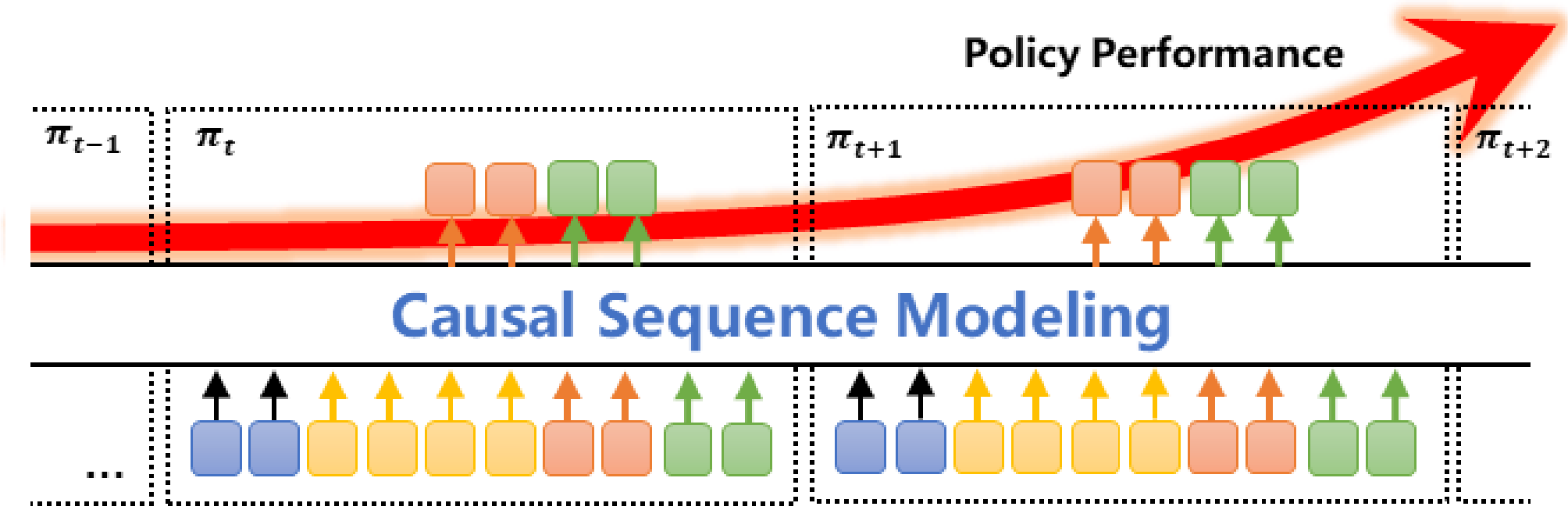
AIRS SHENZHEN INSTITUTE of ARTIFICIAL INTELLIGENCE AND ROBOTICS for SOCIETY 深圳市人工智能与机器人研究院

Baidu 百度

*Corresponding to: fanwang.px@gmail.com, shaoshanliu@cuhk.edu.cn*

# ICRL : In-Context Reinforcement Learning

- Self-directed exploration and exploitation
- Learning from **external feedback** instead of internal reasoning (e.g CoT)
- ICRL is a key to **gradient-free, experience-driven learning** for decision making



# Scaling ICRL is challenging

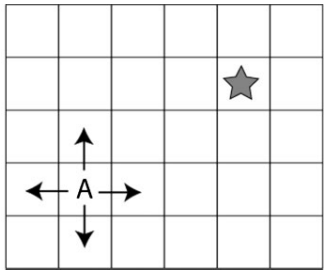
- The lack of large-scale, low-structural-bias collections of decision tasks.
- **Training / Incentivizing ICRL is hard**
  - (1) Self supervision or distillation (Laskin et al., 2022) is ineffective
  - (2) RL2 (Duan et al., 2016) is effective but more expensive

# Contributions

- Scalable task set **AnyMDP**: high quality & diversity
- Incentivizing ICRL with **efficient training frameworks**
- **OmniRL** - a general in-context learner for discrete decision making: scaling up ICRL to 500K tasks and 500K sequence length:

# AnyMDP : high-quality and high-diversity MDP tasks

Bandits

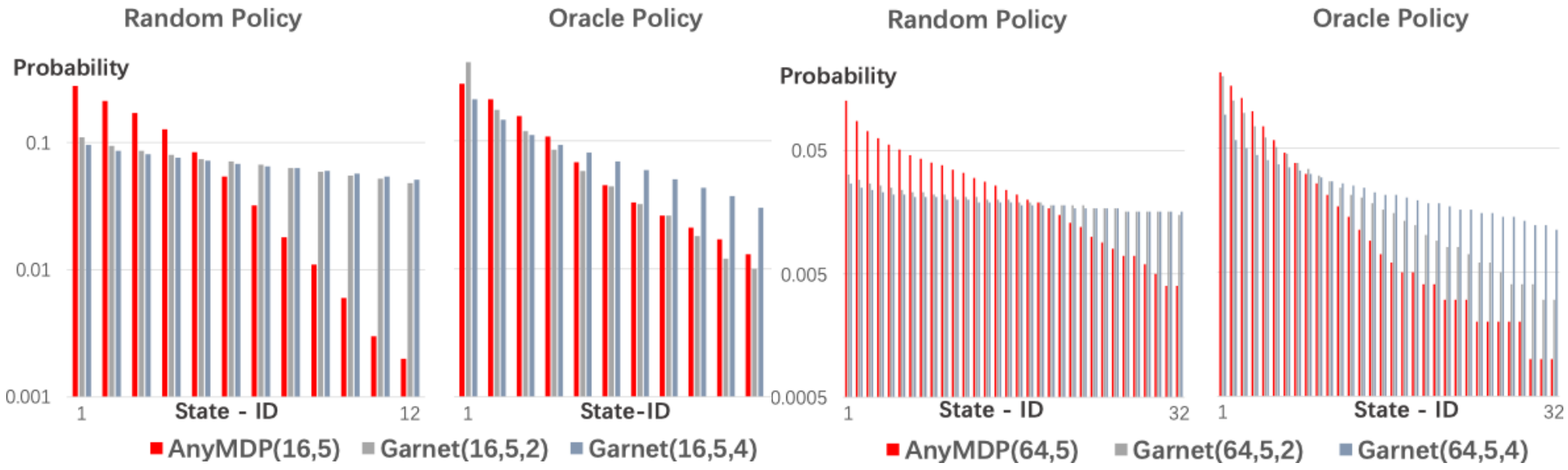



Grid Worlds

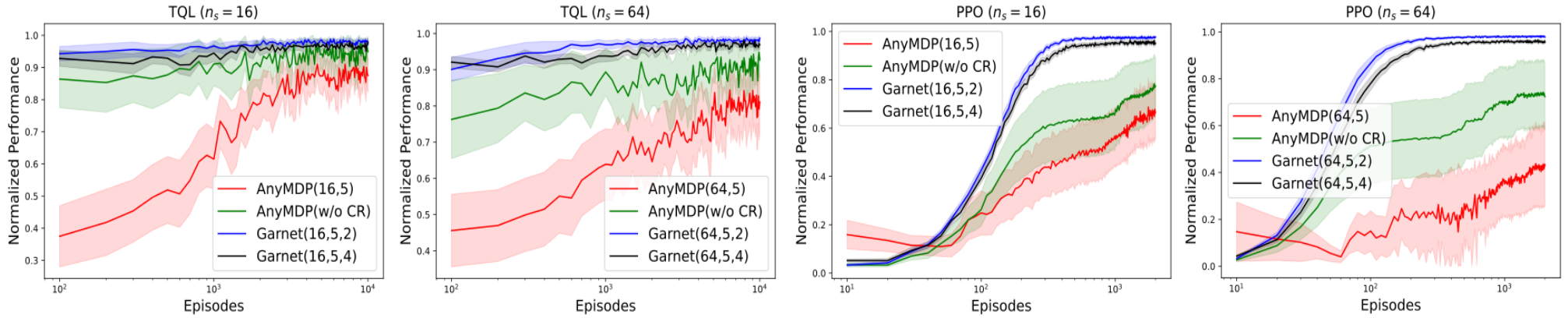
Disturbed Games
(Cobbe et al., 2019)

- Manually designed tasks: high structural bias and inductive bias
- Randomly sample transition & rewards *(Bhatnagar et al 2009):* end up in trivial and low-difficulty tasks mostly
- AnyMDP: lower structural bias, higher quality achieved through the constraint of:
  - (1) **banded transition matrices**
  - (2) **composite reward (CR)** sampling

# Comparison of AnyMDP & Others Tasks



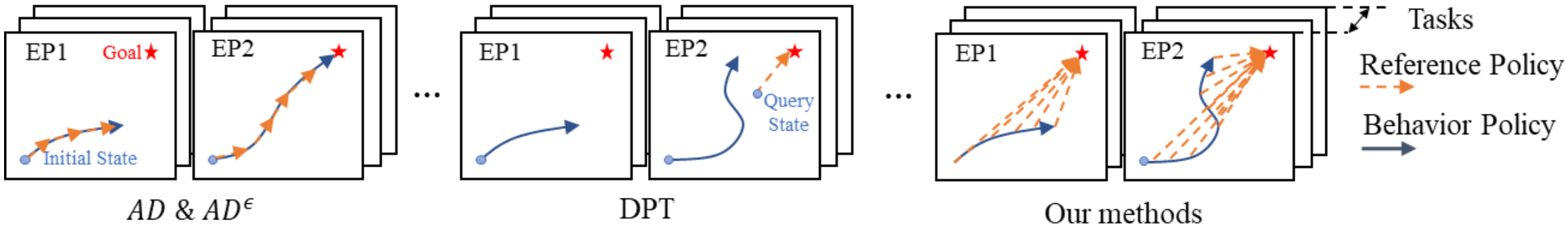

**AnyMDP follows the rule: high-valued states is exponentially less probable to reach by chance**




**AnyMDP tasks raises more challenge to reinforcement learners such as Q-Learning and PPO**

# Efficient Meta-Training of ICRL

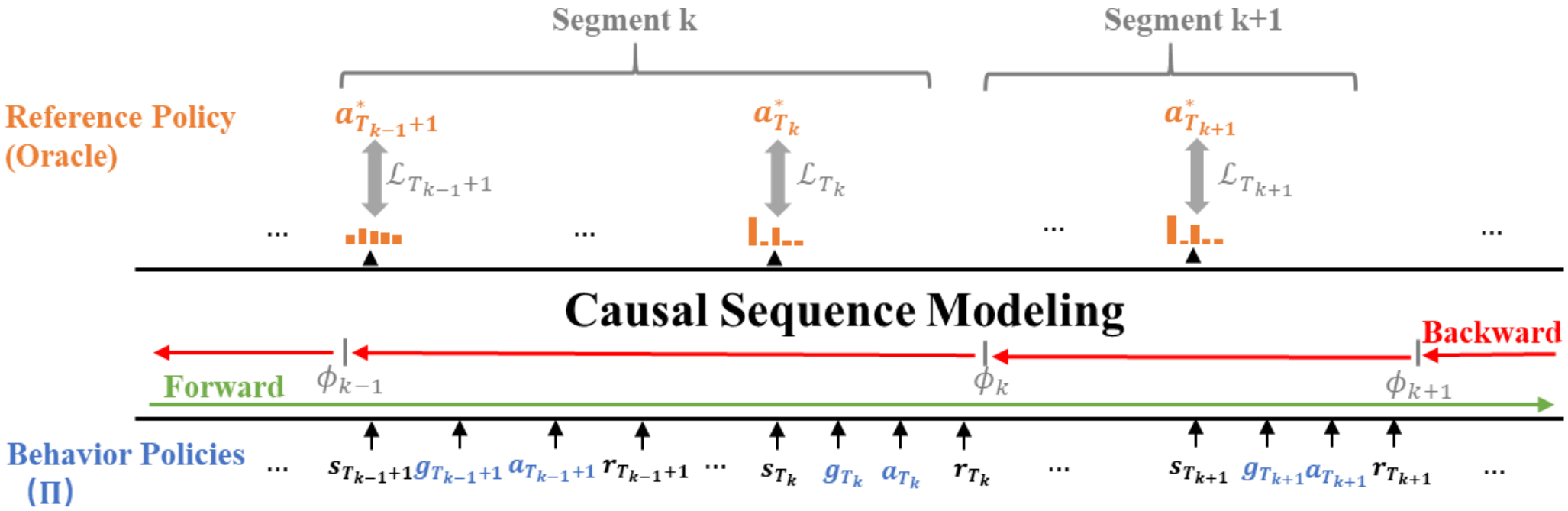
- **Decoupled Policy Distillation (DPD)**: Efficient step-by-step policy distillation via decoupling behavior policy and reference policy
  - Inspired by DPT (*Lee, et al. 2024*) and DAgger (*Ross et al., 2011*)
  - Allowing chunk-wise training just like pretraining
  - Reducing the problem of distribution shift by introducing noisy behavior policy




**DPD explained: step-by-step trajectory generation with noisy behavior, label with oracle reference**

# Efficient Meta-Training of ICRL

- **Prior information integration**: Enable ICL with both posterior (reward) and prior information, thus enable self-directed versatile **In-Context-{RL, Offline RL, Imitation}**
- **Chunkwise Training & Linear Attention**: Overcomes long-context computational limits, enabling training on sequences of arbitrarily length




**Context arrangement, model structure, and training target functions**

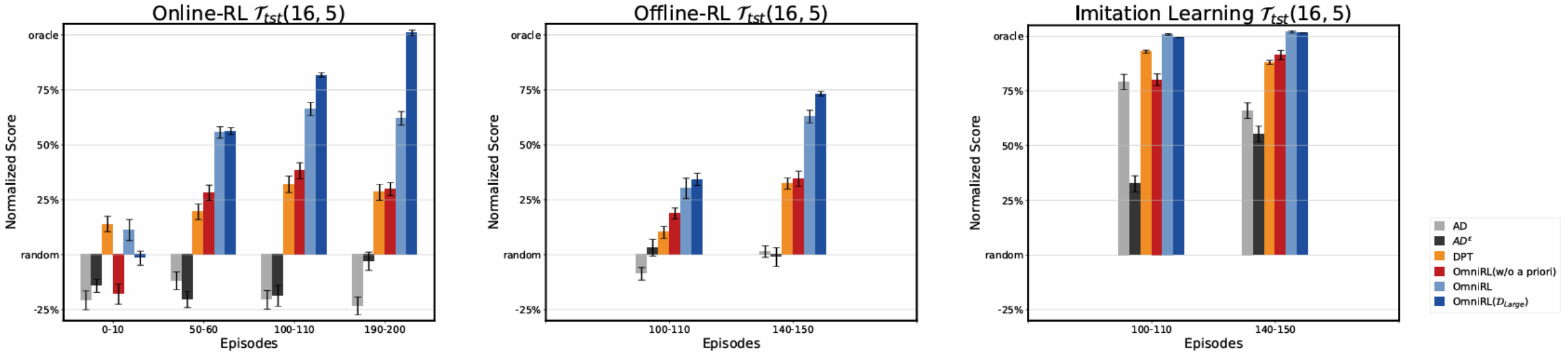
# OmniRL: Towards General In-Context Learner of Decision Making

- OmniRL generalizes to **unseen & unconsidered tasks** in meta-training
- OmniRL generalizes to **multi-agent cooperation** without multi-agent training
- OmniRL is **two orders of magnitude more sample efficient** than canonical RL

| **Environments** | **Performances / AVG. *Steps* cost / AVG. *Episodes* cost** | | | |
|---|---|---|---|---|
| | **TQL-UCB** | **PPO** | **OmniRL** (AnyMDP) | **OmniRL** (GarnetMDP) |
| $\mathcal{T}_{tst}(1,5)$ (Bandits) | $92.1\%/100/100$ | $95.6\%/1.2K/1.2K$ | $82.5\%/103/103$ | $46.6\%/33/33$ |
| $\mathcal{T}_{tst}(16,5)$ | $92.0\%/297K/4.7K$ | $90.6\%/476K/9.7K$ | $95.3\%/2.0K/29$ | $47.8\%/1.6K/24$ |
| $\mathcal{T}_{tst}(32,5)$ | $84.7\%/616K/5.6K$ | $72.2\%/618K/9.7K$ | $90.3\%/6.5K/47$ | $42.0\%/5.0K/44$ |
| $\mathcal{T}_{tst}(64,5)$ | $83.7\%/1.1M/5.1K$ | $58.3\%/1.1M/9.4K$ | $91.3\%/7.7K/25$ | $47.1\%/6.6K/24$ |
| $\mathcal{T}_{tst}(128,5)$ | $73.2\%/1.8M/6.9K$ | $49.0\%/1.3M/8.6K$ | $80.2\%/36.3K/100$ | $32.3\%/9.0K/31$ |
| Garnet$(16,5,2)$ | $98.8\%/241K/2.1K$ | $97.1\%/57K/0.5K$ | $85.9\%/8.2K/71$ | $99.0\%/10.8K/95$ |
| Garnet$(64,5,2)$ | $98.7\%/614K/1.7K$ | $98.1\%/96K/0.26K$ | $80.4\%/8.0K/19$ | $87.3\%/7.4K/23$ |
| CliffWalking | $100\%/3.1K/35$ | $95.9\%/99.3K/2.7K$ | $100\%/3.0K/65$ | $63\%/29K/300$ |
| FrozenLake (non-slippery) | $95.3\%/23.6K/3.7K$ | $96.8\%/18.2K/2.1K$ | $99.8\%/0.3K/35$ | $75.1\%/4.0K/250$ |
| FrozenLake (slippery) | $96\%/208K/10.0K$ | $95.6\%/73.6K/4.7K$ | $79.5\%/7.7K/245$ | $31.3\%/11.8K/800$ |
| Discrete-Pendulum (g=1) | $94.9\%/22K/110$ | $99.3\%/198K/990$ | $90.5\%/8K/40$ | $0.0\%/-/-$ |
| Discrete-Pendulum (g=5) | $99.7\%/426K/2.13K$ | $99.8\%/132K/660$ | $91.8\%/34K/170$ | $0.0\%/-/-$ |
| Discrete-Pendulum (g=9.8) | $90.2\%/2.0M/10.0K$ | $98.3\%/186K/930$ | $73.4\%/33K/165$ | $0.0\%/-/-$ |
| Switch2 (Multi-Agent) | $98\%/3.8K/110$ | – | $80.4\%/2.8K/100$ | – |
| Darkroom (6x6) | $98.1\%/6.2K/481$ | $97.6\%/10.6K/560$ | $95.2\%/845/40$ | $90.5\%/21.3K/440$ |
| Darkroom (8x8) | $96.8\%/24.5K/2.0K$ | $96.7\%/15.9K/930$ | $93.8\%/1.5K/40$ | $88.9\%/30.4K/480$ |
| Darkroom (10x10) | $89\%/31.1K/1.7K$ | $92.3\%/15.7K/570$ | $91.7\%/2.8K/100$ | $75.6\%/20.8K/280$ |

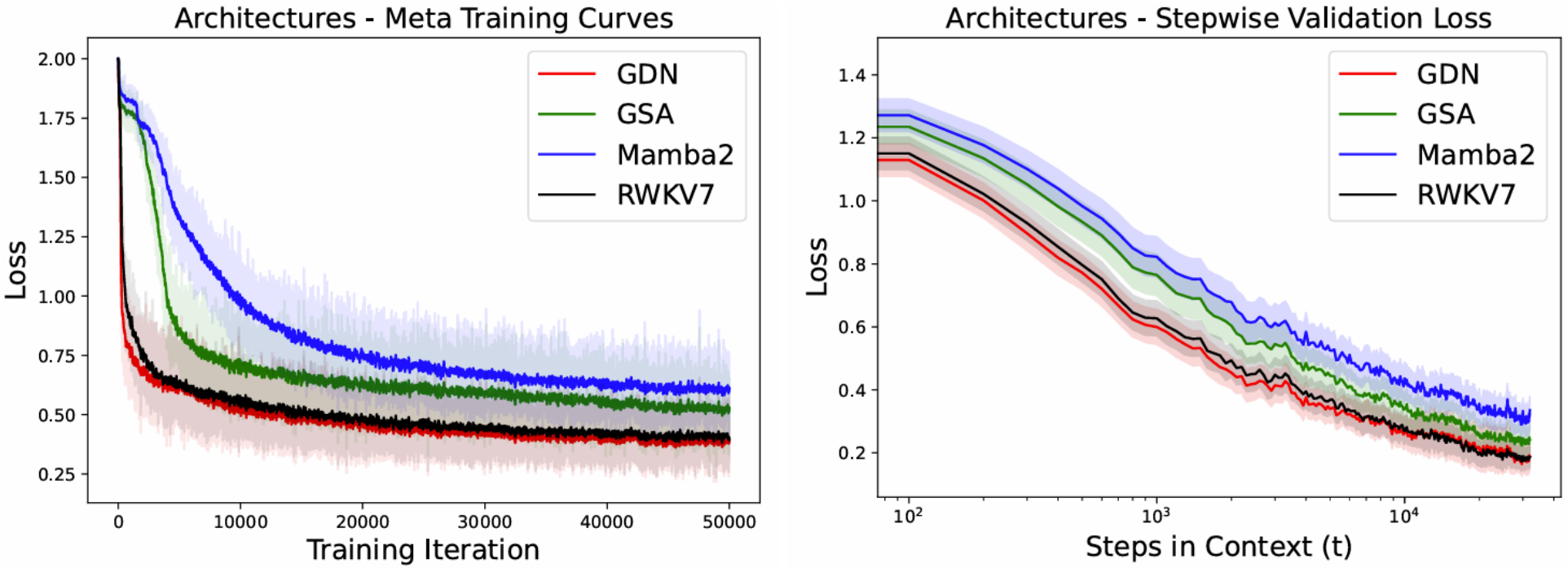
# OmniRL: Towards General In-Context Learner of Decision Making

- OmniRL performs **In-Context-{RL, Offline RL, Imitation} better** than previous ICRL



# Benchmarking Long-Context Modeling

- AnyMDP is an effective benchmark for evaluating **long-context modeling**



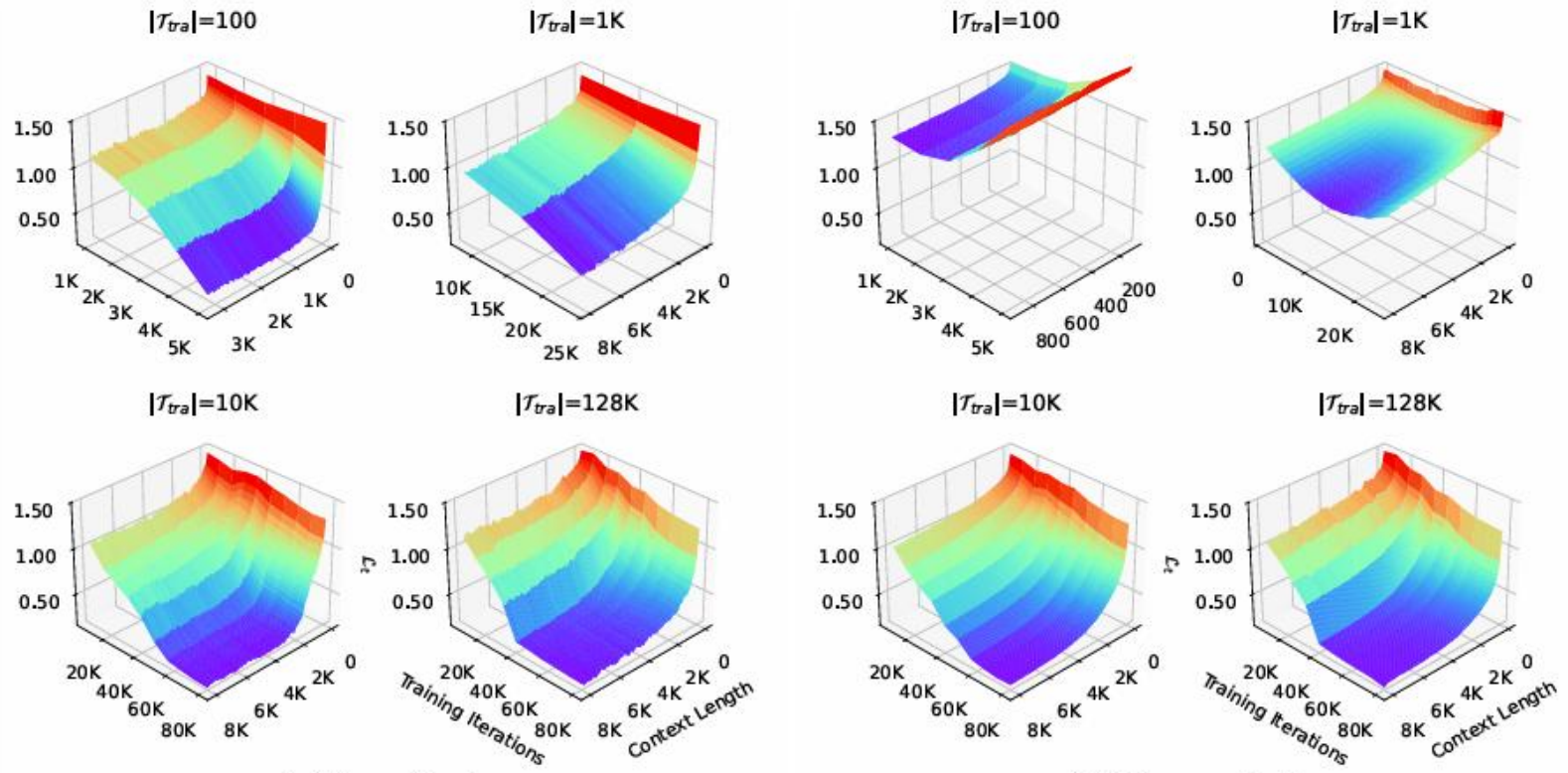

**Comparison of the performance of linear-attention models in AnyMDP dataset**

# Investigating Task Scaling in ICRL

- **Larger task scale** switches ICRL from **task identification** to **task learning** mode
- Longer context is a cost to larger generalization scope

## Ensights & Takeaways

- **Task diversity** (at least 10K) and **Long-context modeling** is essential for ICRL
- ICRL generalization requires longer adaptation periods, prioritizing **asymptotic performance** over few-shot performance in evaluation.




(a) Seen Tasks (b) Unseen Tasks

| Validation set | Metrics | $\lvert\mathcal{T}_{tra}\rvert$ | | | |
|---|---|---|---|---|---|
| | | 100 | $1K$ | $10K$ | $128K$ |
| **Seen tasks** | $\max(d_t)$ | $> 81.0\%$ | $> 65.4\%$ | $\geq 86.6\%$ | $\geq 84.5\%$ |
| | $\min(t)$ s.t. $d_t \geq 80\%$ | 0.88K | - | 2.4K | 5.2K |
| **Uneen tasks** | $\max(d_t)$ | 17.9% | 38.0% | 84.4% | $\geq 84.8\%$ |
| | $\min(t)$ s.t. $d_t \geq 80\%$ | - | - | 3.9K | 5.1K |
| | $\max(d_t)$ achieved at iteration | $2K$ | $14K$ | $72K$ | $\geq 80K$ |

# Thanks for watching

AnyMDP

OmniRL