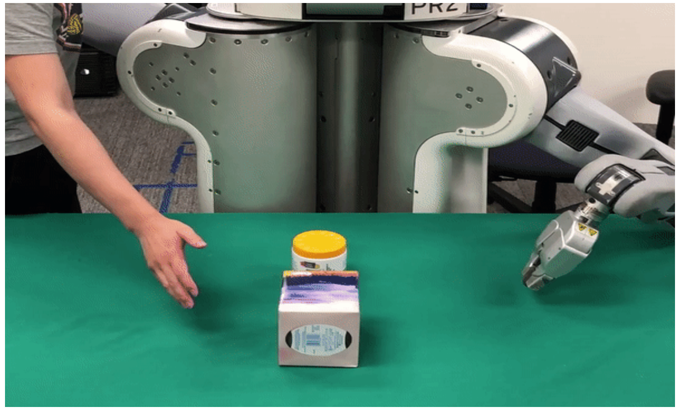
Learning from Visual Observation via Offline Pretrained State-to-Go Transformer

Bohan Zhou, Ke Li, Jiechuan Jiang, Zongqing Lu

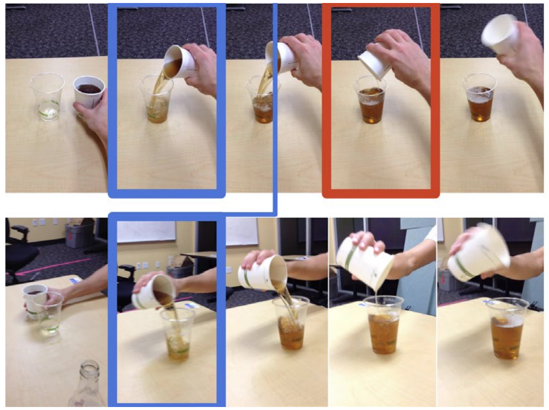
24/10/2023

Motivation

北京大學
PEKING UNIVERSITY



Learning from Demonstrations
(LfD)



Learning from Visual Ovservations
(LfVO)

From LfD to LfVO

- ✓ **Less Supervision**
- ✓ **Enlarging resource**
- ✓ **Biologically reasonable**

LfD:

- \+ Easy to learn
- − Hard & expensive annotations

LfVO:

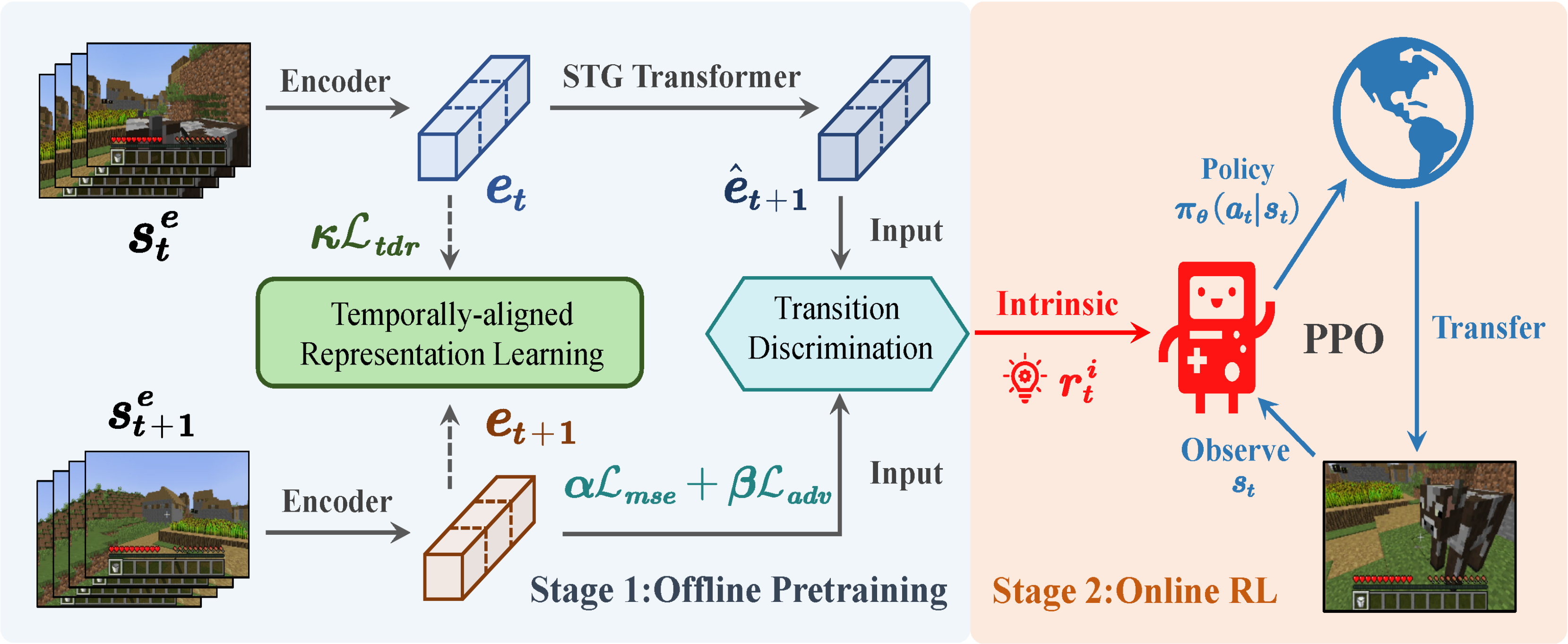
- \+ No actions or rewards
- \+ An ocean of Internet videos
- \+ Explore unknown expert policy
- − Hard to extract useful experience

Previous work

- **IDM-based methods – extra component, compounding error**
- **Adversarial methods – sample-inefficient online learning schemes**
- **Representation-learning-based methods – over-optimistic estimation**
- **Goal-oriented methods – extra task-specific information**

Abundant **video-only** data contain useful behavior patterns. How can we effectively leverage them to tackle downstream **reward-free** visual control tasks?

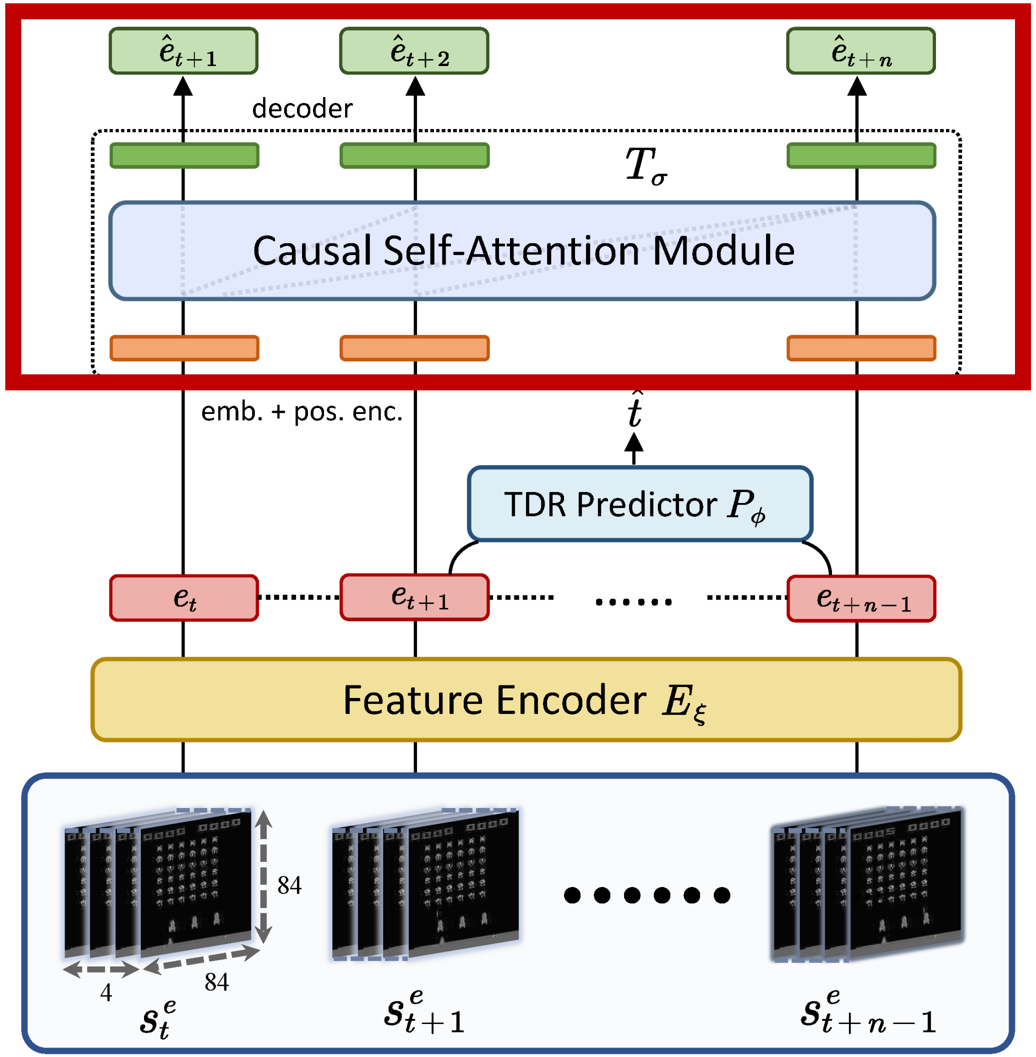
Two-stage framework



- **Pretraining stage**: we simultaneously learn a **GPT** for latent transition prediction, an expert transition **discriminator** for intrinsic rewards and a temporal distance regressor (**TDR**) for temporally-aligned representations.
- **Reinforcecment learning stage**: agents **merely** learn from generated rewards from discriminator without environmental reward signals.

Offline Pretraining

北京大學 PEKING UNIVERSITY



1. Predicting Latent Transition

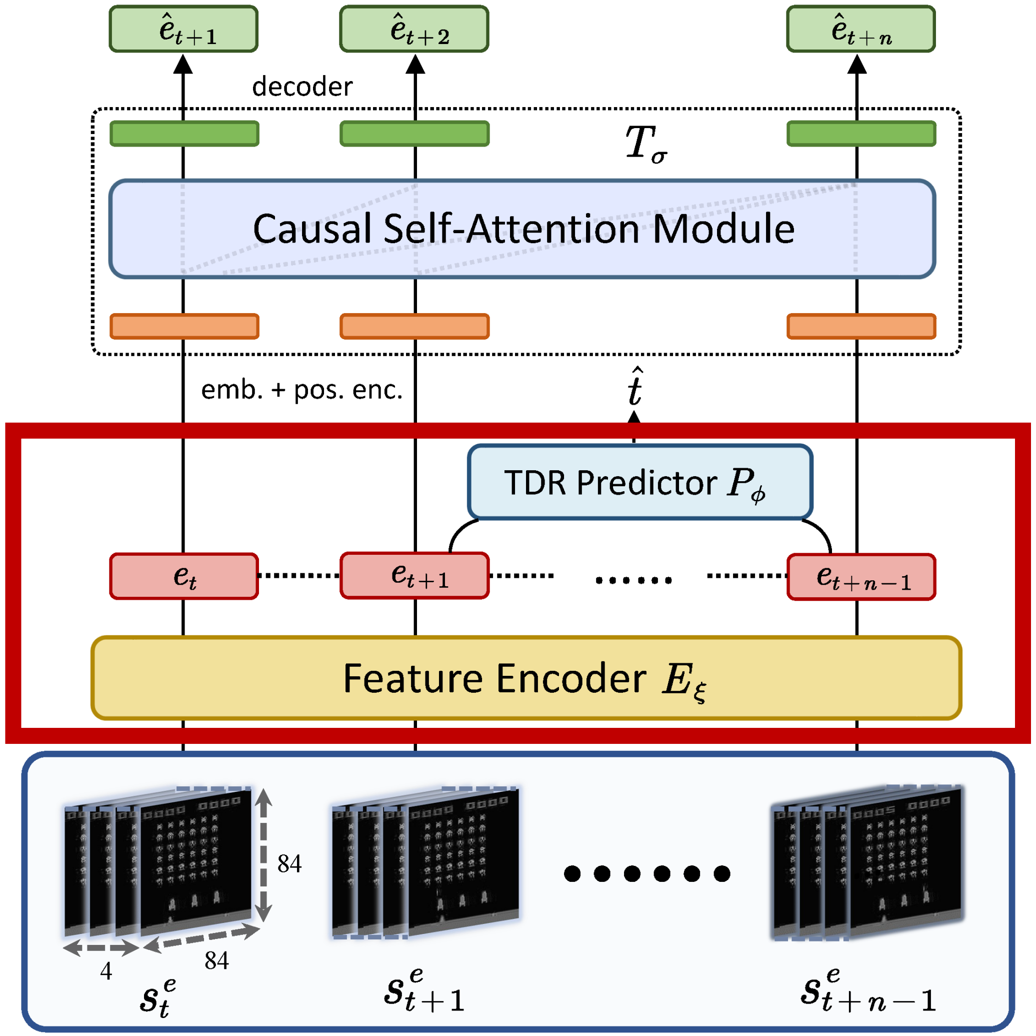
Adversarially learn transition module with L2 regularization as well as a WGAN discriminator

$$\text{for } D_\omega: \min_{w \in \mathcal{W}} \mathbb{E}_{\mathcal{D}^e}[D_\omega(e_t, \hat{e}_{t+1}) - D_\omega(e_t, e_{t+1})]$$

$$\text{for } T_\sigma: \min_{\xi,\sigma} \mathbb{E}_{\mathcal{D}^e}[-D_\omega(e_t, \hat{e}_{t+1}) + \|\hat{e}_{t+1} - e_{t+1}\|_2^2]$$

$$e_t = E_\xi(s_t), \ \hat{e}_{t+1} = T_\sigma(e_t)$$

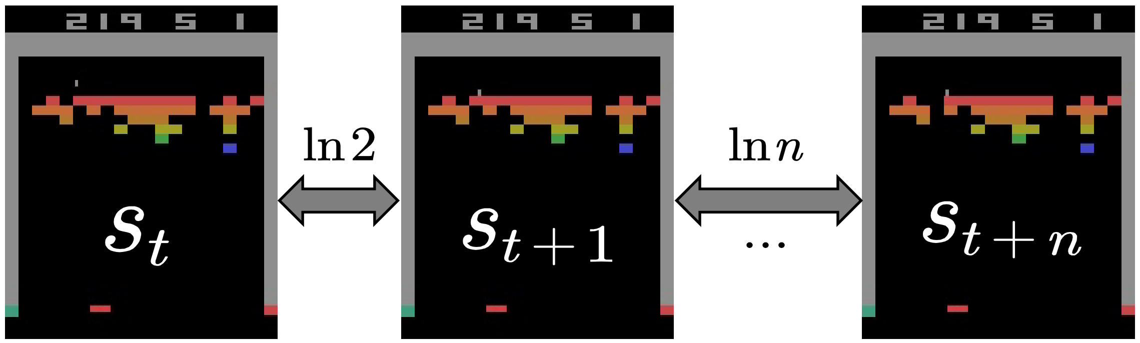
Offline Pretraining



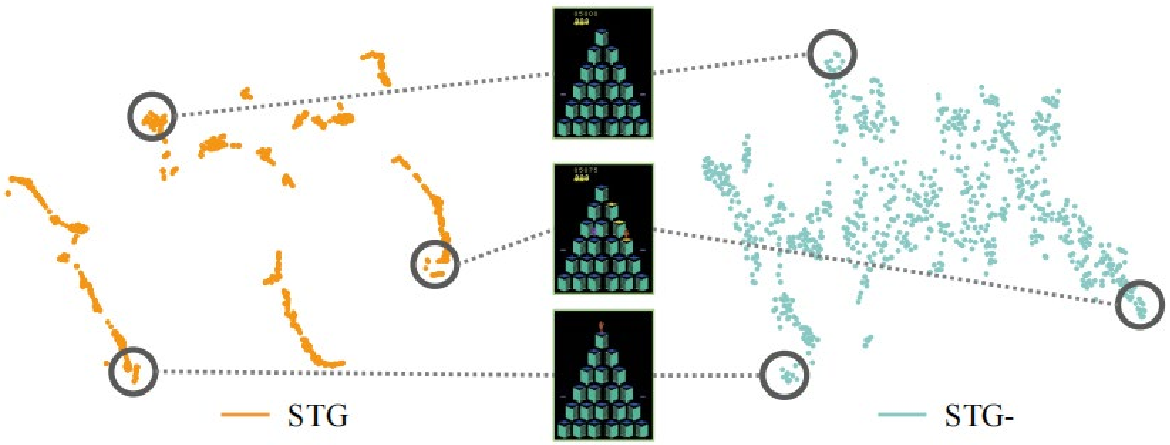
2. Learning Temporally-Aligned Representation

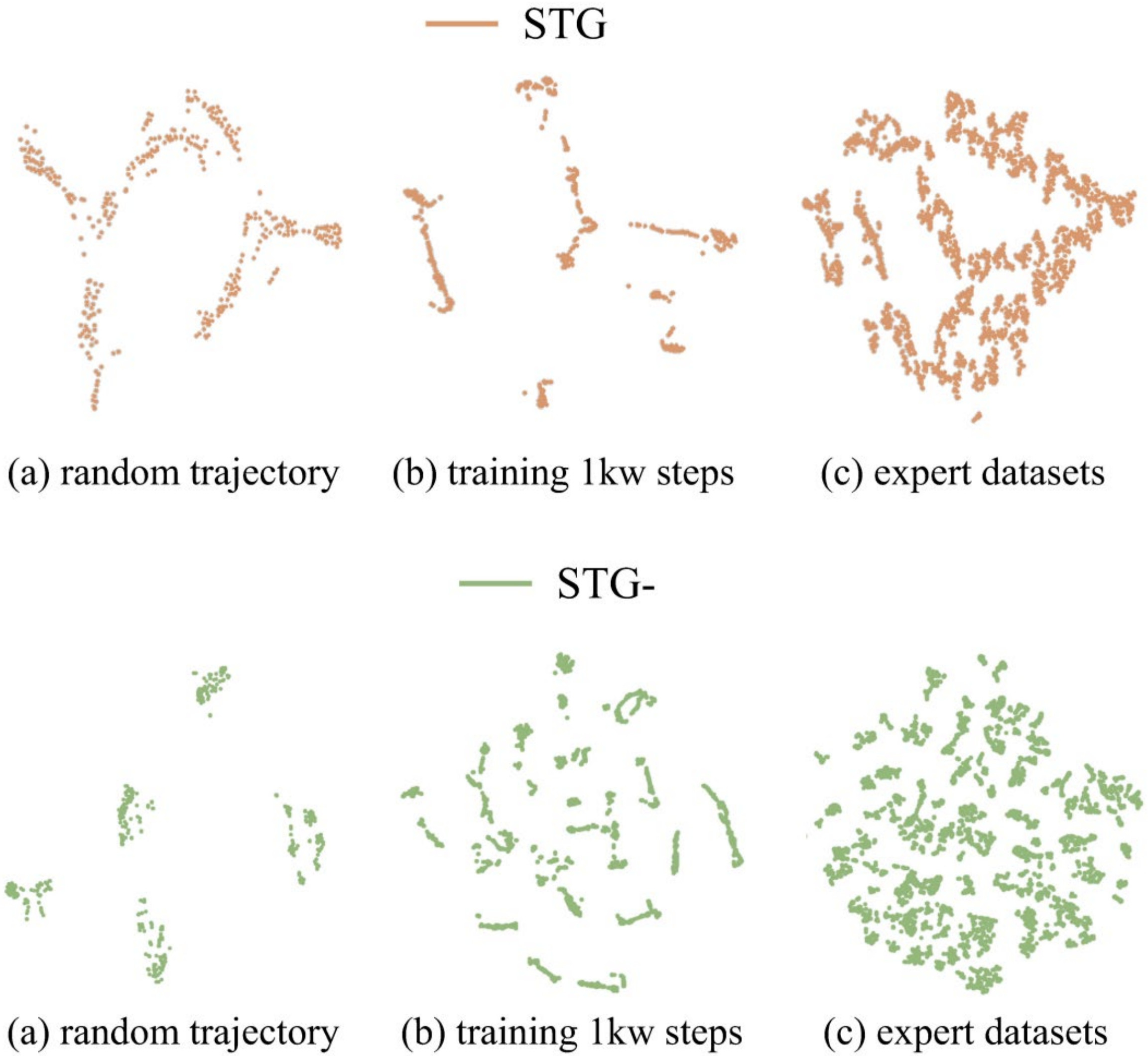
Apply symlog temporal distance prior in low-dimensional representation space

$$\min_{\xi,\phi} \mathbb{E}_{\mathcal{D}^e} \| P_\phi(e_t, e_{t+j}) - \mathrm{sign}(j) \ln(1 + |j|) \|$$



TDR Representation





Algo: STG Pretraining

Algorithm 1 STG Transformer Offline Pretraining

Input: STG Transformer T_σ, feature encoder E_ξ, discriminator D_ω, expert dataset $D^e = \{\tau^1, \tau^2, \dots, \tau^m\}, \tau^i = \{s^i_1, s^i_2, \dots\}$, buffer $\mathcal{B}$, loss weights α, β, κ .

```
Initialize parametric network E_ξ, T_σ, D_ω randomly.
for e ← 0, 1, 2 ... do                                              ▷ epoch
    Empty buffer B.
    for b ← 0, 1, 2 ... |B| do                                      ▷ batchsize
        Stochastically sample state sequence τ^i from D^e.
        Stochastically sample timestep t and n adjacent states {s^i_t, ..., s^i_{t+n-1}} from τ^i.
        Store {s^i_t, ..., s^i_{t+n-1}} in B.
    end for
    Update D_ω: ω ← clip(ω − ε∇_ω L_dis, −0.01, 0.01).
    Update E_ξ and T_σ concurrently by minimizing total loss αL_mse + βL_adv + κL_tdr.
end for
```

Algo: Online RL

Pretrained WGAN discriminator works as reward function:

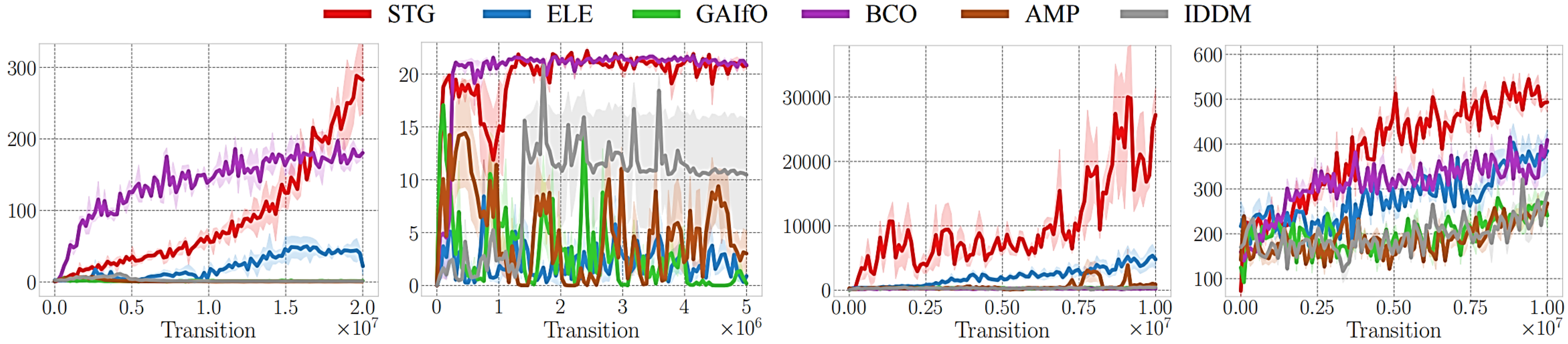
$$r_t^i = -\left[D_\omega\big(E_\xi\left(s_t\right), T_\sigma\left(E_\xi\left(s_t\right)\right)\big) - D_\omega\big(E_\xi\left(s_t\right), E_\xi\left(s_{t+1}\right)\big) \right]$$

Algorithm 2 Online Reinforcement Learning with Intrinsic Rewards

Input: pretrained $E_\xi, T_\sigma, D_\omega$, policy π_θ, MDP $\mathcal{M}$, intrinsic coefficient η.

Initialize parametric policy π_θ with random θ randomly and reset $\mathcal{M}$.
while updating π_θ **do** $\triangleright$ policy improvement
    Execute π_θ and store the resulting n state transitions $\{(s, s')\}_t^{t+n}$.
    Use E_ξ to obtain n real latent transitions $\{(e, e')\}_t^{t+n}$.
    Use T_σ to obtain n predicted latent transitions $\{(e, \hat{e}')\}_t^{t+n}$.
    Use D_ω to calculate intrinsic rewards: $\Delta_t^{t+n} = \{D_\omega(e, \hat{e}')\}_t^{t+n} - \{D_\omega(e, e')\}_t^{t+n}$.
    Perform PPO update to improve π_θ with respect to $r^i = -\eta\Delta$.
end while

Atari Experiments

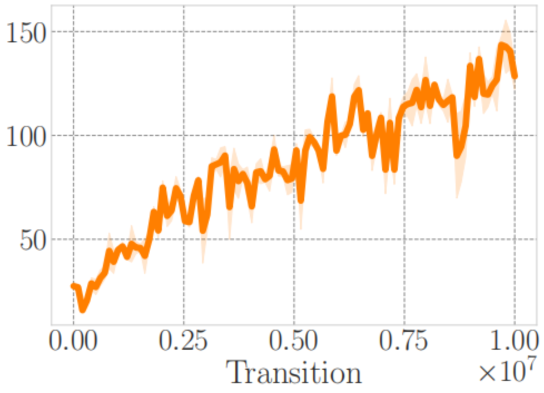


(a) Breakout (b) Freeway (c) Qbert (d) Space Invaders

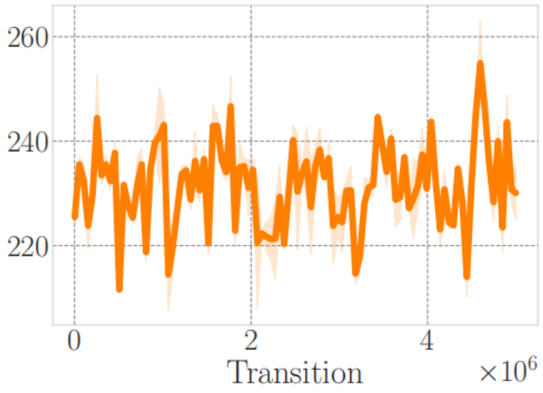
Environment	GAIfO	AMP	IDDM	ELE	BCO	STG	Expert	PPO
Breakout	1.5	0.6	1.2	22.0	180.4	**288.8**	212.5	274.8
Freeway	0.6	3.0	10.5	2.7	21.6	21.8	31.9	32.5
Qbert	394.4	874.9	423.3	4698.6	234.1	**27234.1**	15620.7	14293.3
Space Invaders	260.2	268.1	290.4	384.6	402.2	502.1	1093.9	942.5

Learning from **50** trajectories for each task, STG demonstrates **superiority among baselines** and even **surpass expert level**.

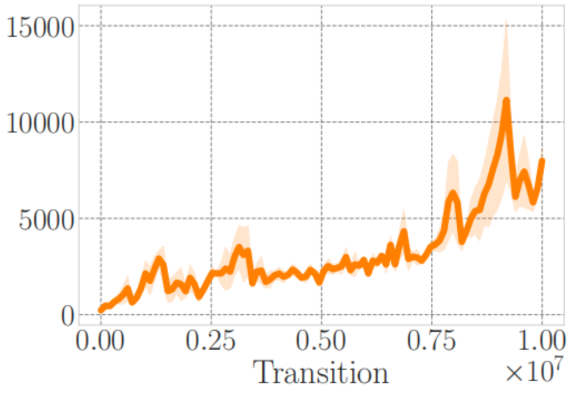
Atari Experiments



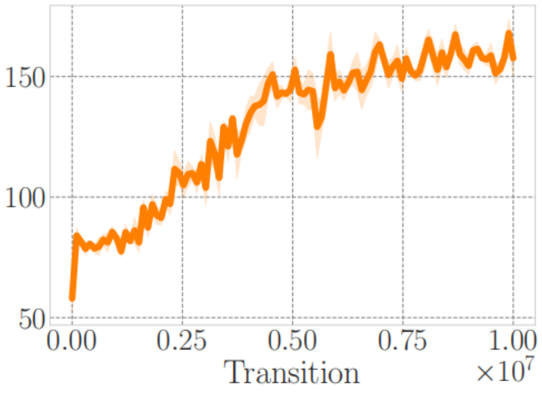
(a) Breakout



(b) Freeway



(c) Qbert

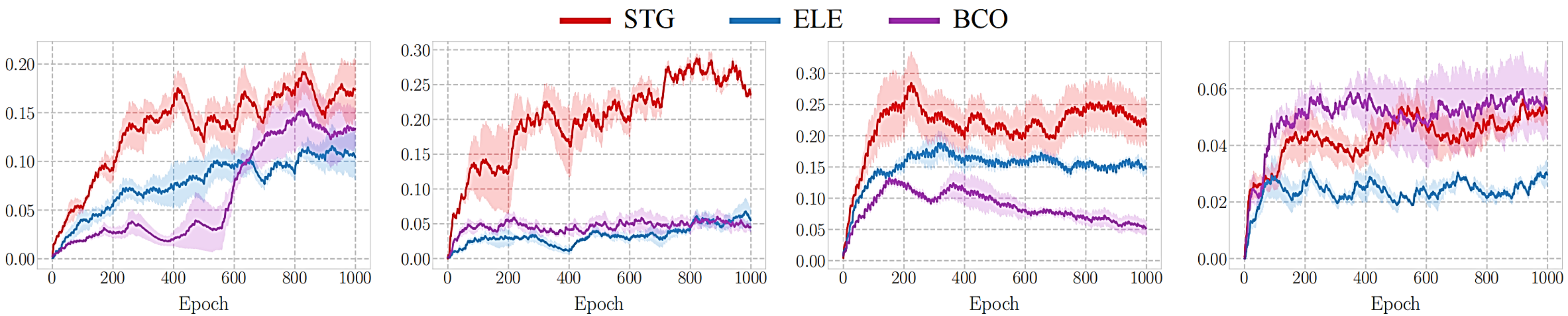


(d) Space Invaders

The rising trend of **intrinsic return** proves that online collected observation distribution is getting **closer** to expert observation distribution during training.

Minecraft Experiments

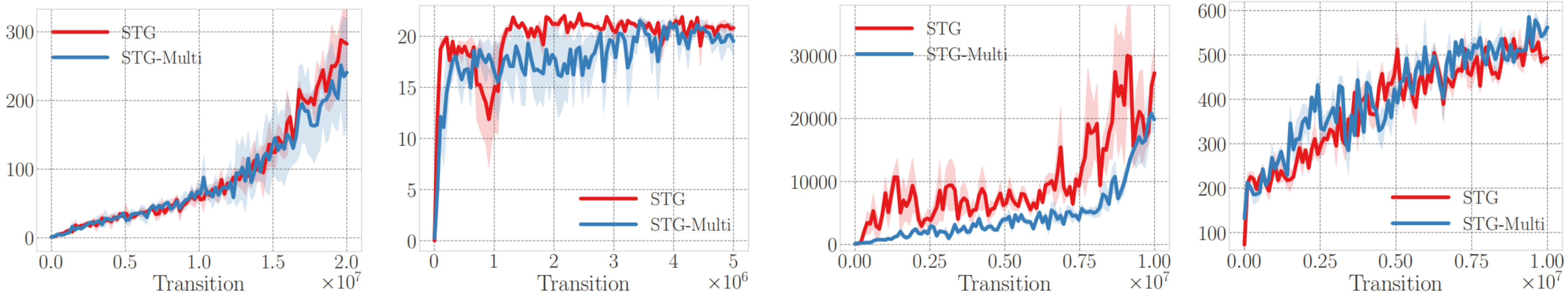
北京大學
PEKING UNIVERSITY



(a) Pick a flower **(b)** Milk a cow **(c)** Harvest tallgrass **(d)** Gather wool

In challenging **open-ended** Minecraft tasks, shows superiority over baselines!

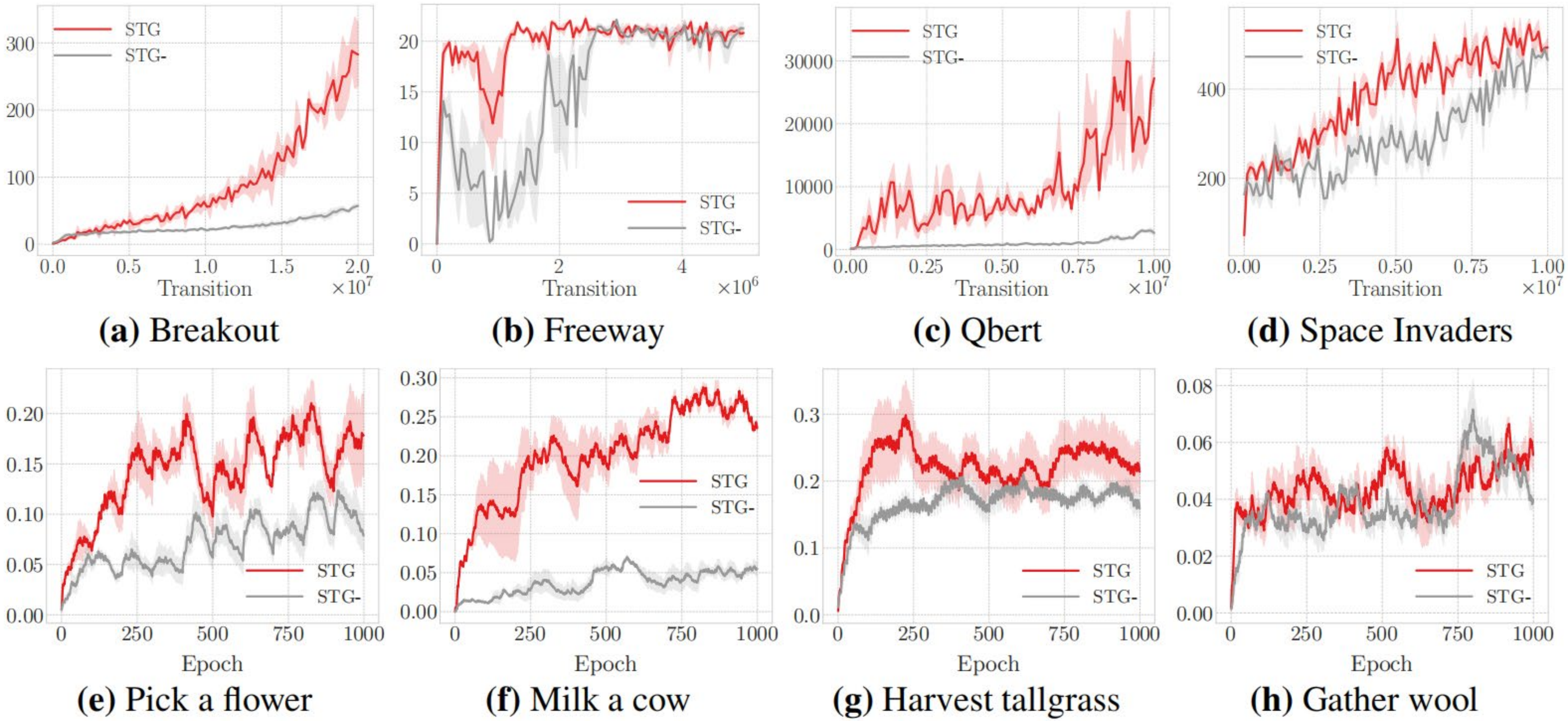
Multi-Task STG



(a) Breakout (b) Freeway (c) Qbert (d) Space Invaders

Pretrained on whole Atari datasets, STG-Multi shows comparable performance.

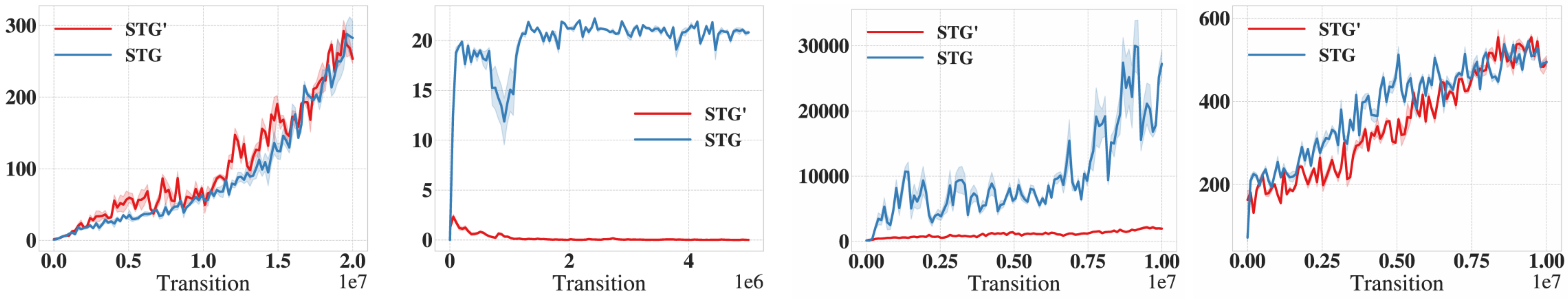
Ablation: TDR removal



Ablation: Reward design

北京大學
PEKING UNIVERSITY

$$r_t^i = D_\omega\left(E_\xi\left(s_t\right), E_\xi\left(s_{t+1}\right)\right) - D_\omega\left(E_\xi\left(s_t\right), T_\sigma\left(E_\xi\left(s_t\right)\right)\right) = r_t^{guide} - r_t^{base}$$



(a) Breakout **(b)** Freeway **(c)** Qbert **(d)** Space Invaders

Figure 4: Atari experiments comparing using r^{guide} (STG') and r^i (STG) as intrinsic reward.

Ablation: Reward design

$$\begin{cases} \text{discrimination reward: } D_\omega(e_t, e_{t+1}) - D_\omega(e_t, \hat{e}_{t+1}) \\ \text{progression reward: } P_\phi(e_t, e_{t+k}),\ k=1 \end{cases}$$

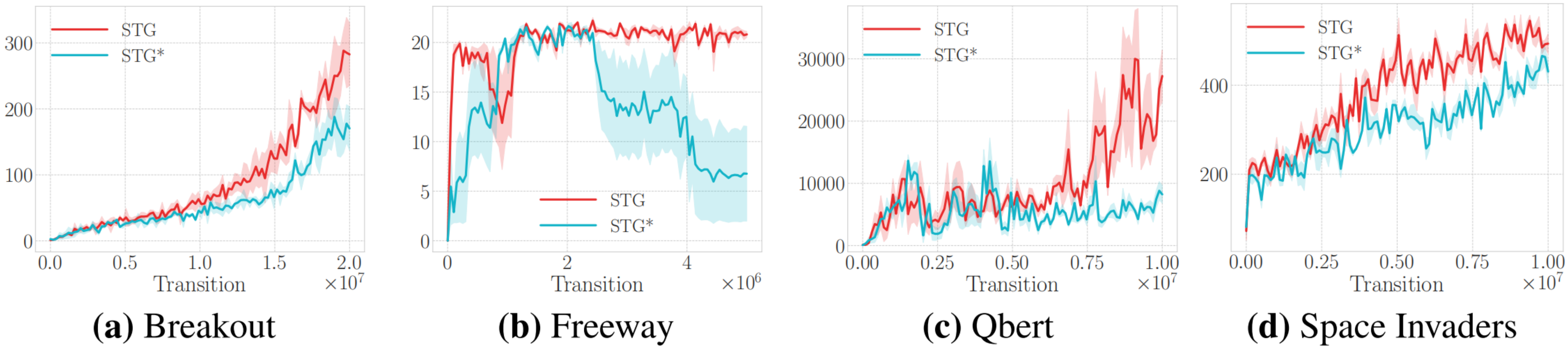
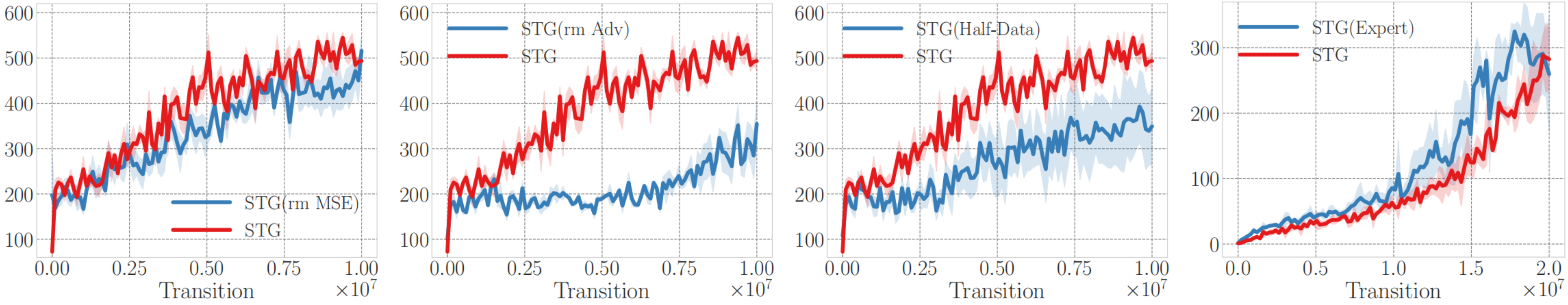


Figure 9: Atari experiments comparing using discriminative rewards (STG) and using both discriminative rewards and progression rewards (STG*).

Ablation : Loss and Dataset

北京大學 PEKING UNIVERSITY



(a) Ablate removing $\mathcal{L}_{mse}$ **(b)** Ablate removing $\mathcal{L}_{adv}$ **(c)** Ablate dataset size **(d)** Ablate dataset quality

Figure 8: Learning curves of four pre-training ablations: (a) removing $\mathcal{L}_{mse}$ in SpaceInvaders; (b) removing $\mathcal{L}_{adv}$ in SpaceInvaders; (c) using half dataset to train STG in SpaceInvaders; (d) using expert dataset to train STG in Breakout.

Extensions

STG offers an effective solution in situations with plentiful video demonstrations, limited environment interactions, and inaccessible labeled action or rewards.

In future work, STG is likely to benefit from:

- more powerful large-scale vision foundation models to facilitate generalization across a broader range of related tasks, domains or embodiments.
- hierarchical framework where one-step predicted rewards can be employed for training low-level policies and multi-step rewards for a high-level policy to tackle long-horizon tasks.

Q&A

北京大學
PEKING UNIVERSITY

Thanks

Bohan Zhou

2023.10.24