CVPR JUNE 3-7, 2026 DENVER COLORADO

Goal-Driven Reward by Video Diffusion Models for Reinforcement Learning

Qi Wang* Mian Wu* Yuyang Zhang* Mingqi Yuan Wenyao Zhang Haoxiang You

Yunbo Wang Xin Jin† Xiaokang Yang Wenjun Zeng

*Equal contribution †Corresponding author

Motivation

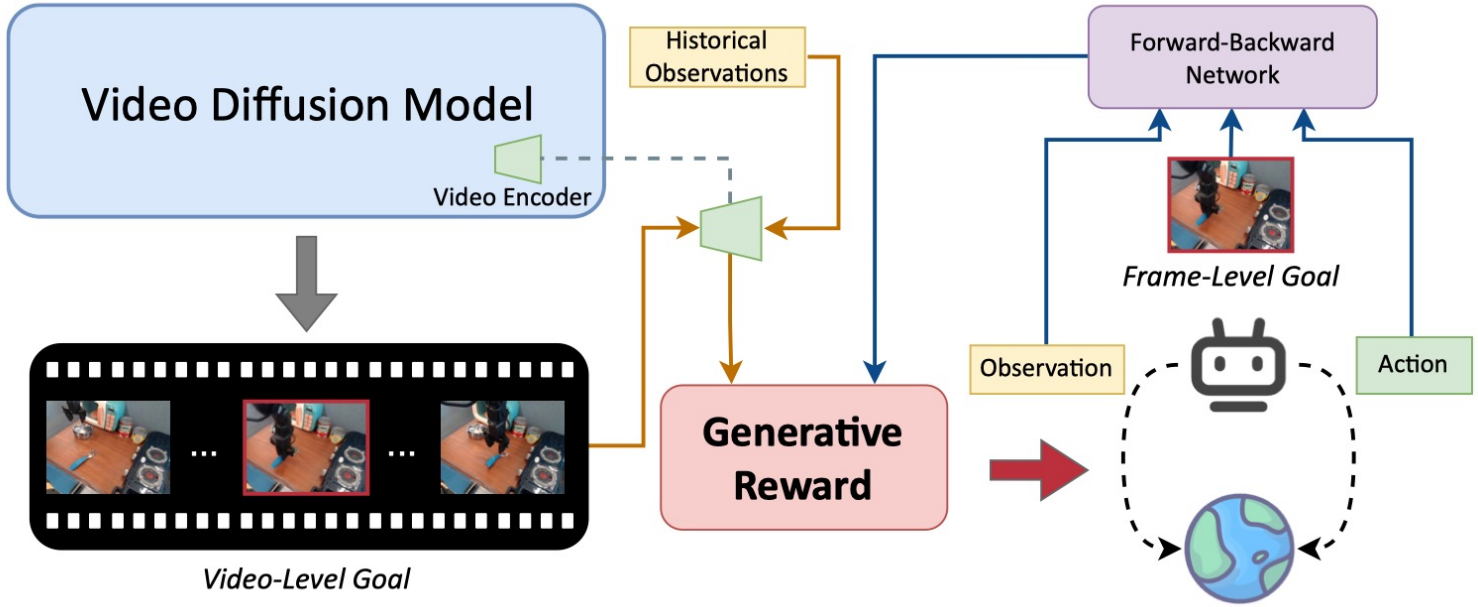
- Traditional reinforcement learning relies heavily on manually designed reward functions, but reward engineering often requires substantial domain expertise and human effort, and it generalizes poorly when tasks change
- Although prior work uses expert videos, vision-language models, or diffusion models to construct rewards, these methods generally do not fully use generated goal videos as goal-driven rewards, and therefore fail to fully exploit the world knowledge embedded in generative models

Compared to other competitive reward models

Model	Demo Free?	Generated?	Action-aware?
RoboCLIP [26]	✗	✗	✗
VLM-RMs [21]	✓	✗	✗
LIV [17]	✓	✗	✗
VIPER [8]	✗	✗	✗
Diffusion Reward [12]	✗	✓	✗
TADPoLe [15]	✓	✓	✗
GenReward (Ours)	✓	✓	✓

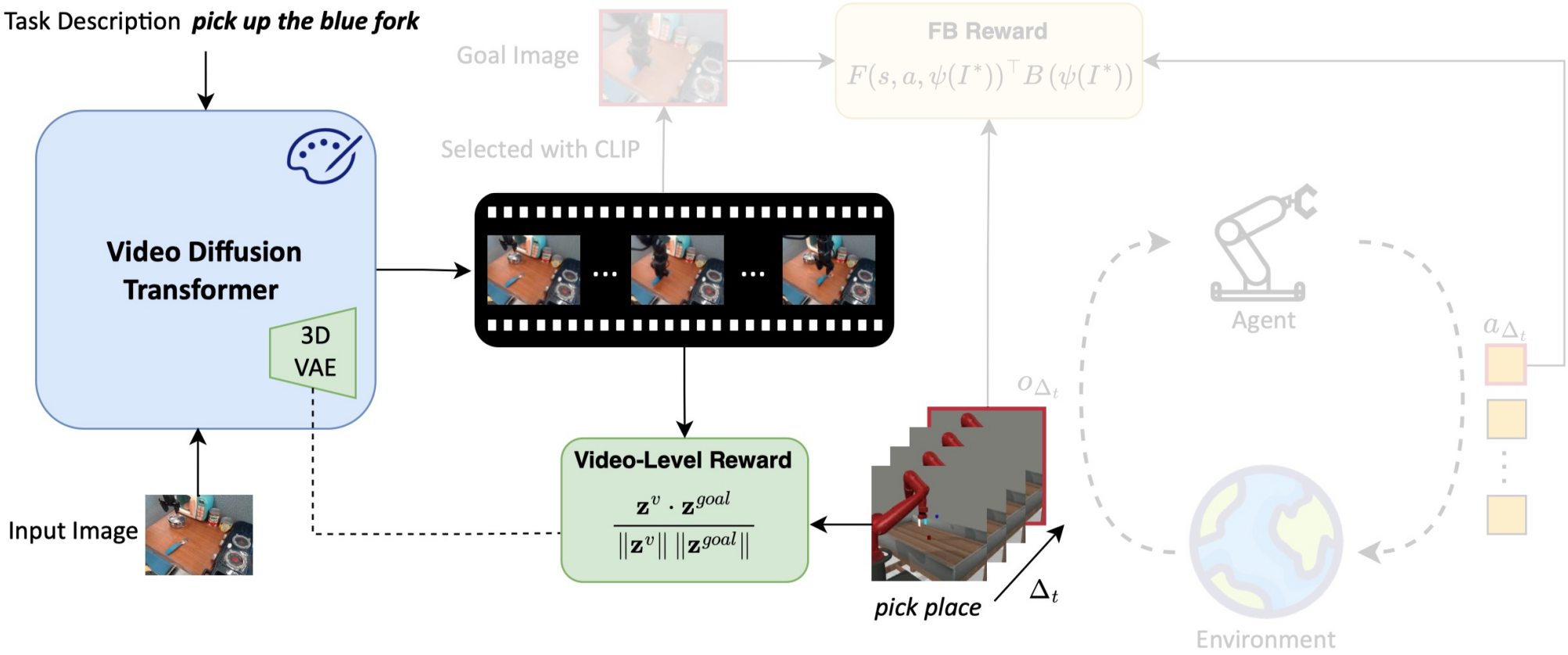
Main idea

- Use a pretrained video diffusion model as a goal-driven reward generator for reinforcement learning
- More specifically, the framework provides reward at two levels: video-level reward and frame-level reward



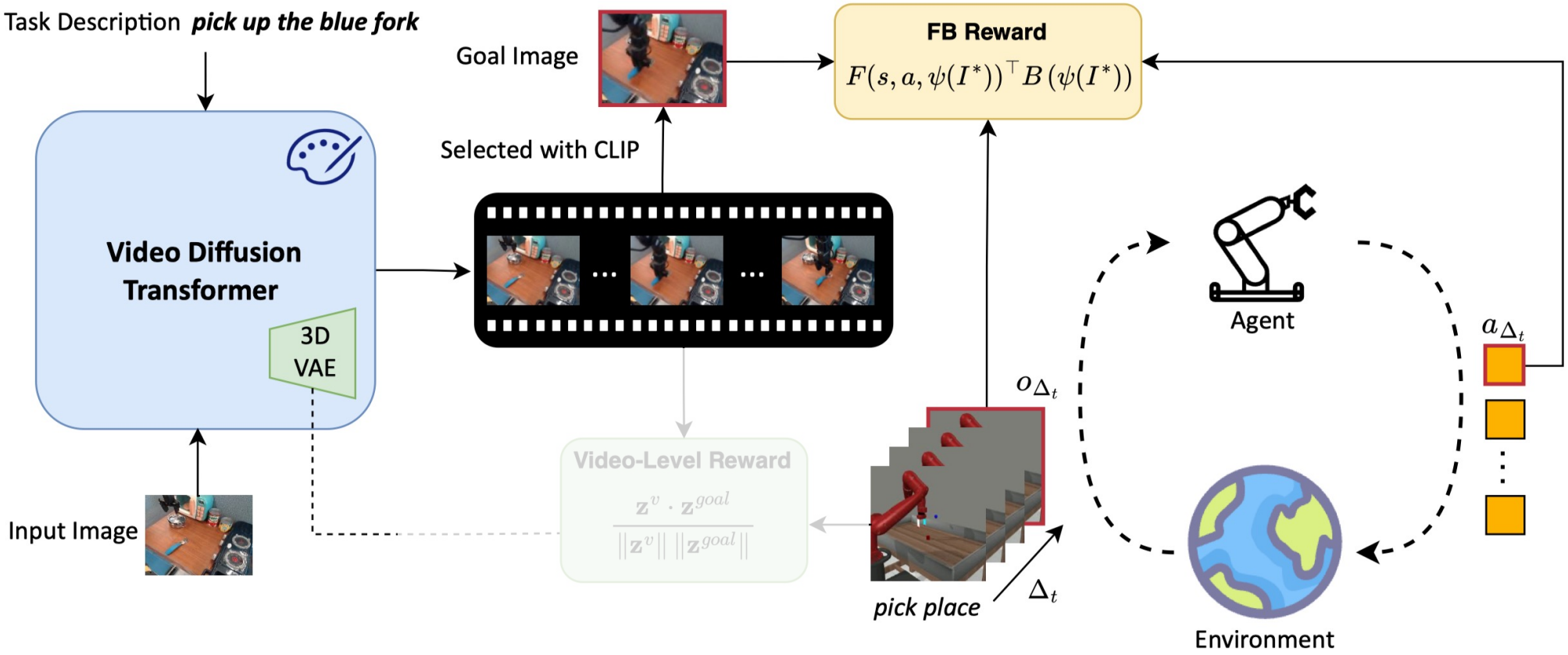
GenReward: Video-Level Reward

- Video-level reward: it compares the latent representation of the agent's trajectory with the latent representation of the generated goal video, so the agent is encouraged to follow the overall desired behavior



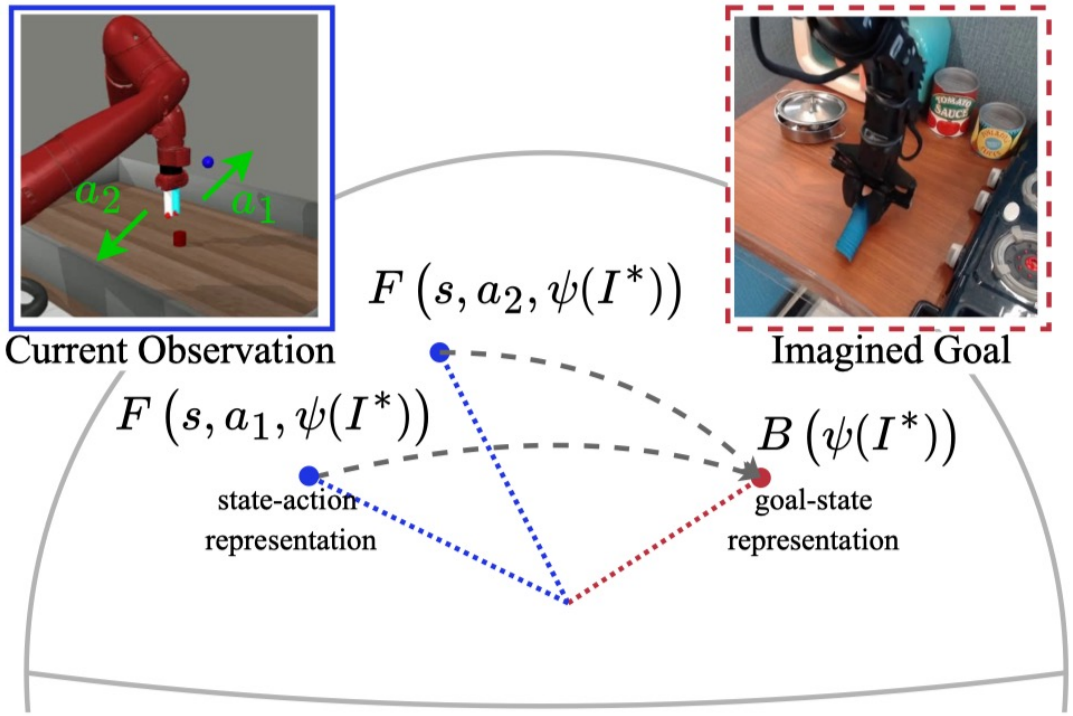
GenReward: Frame-Level Reward

- Frame-level reward: it first selects a key goal frame from the generated video using CLIP, and then uses a forward-backward representation to estimate how likely a state-action pair is to reach that goal state, providing finer-grained and action-aware guidance



GenReward: Frame-Level Reward

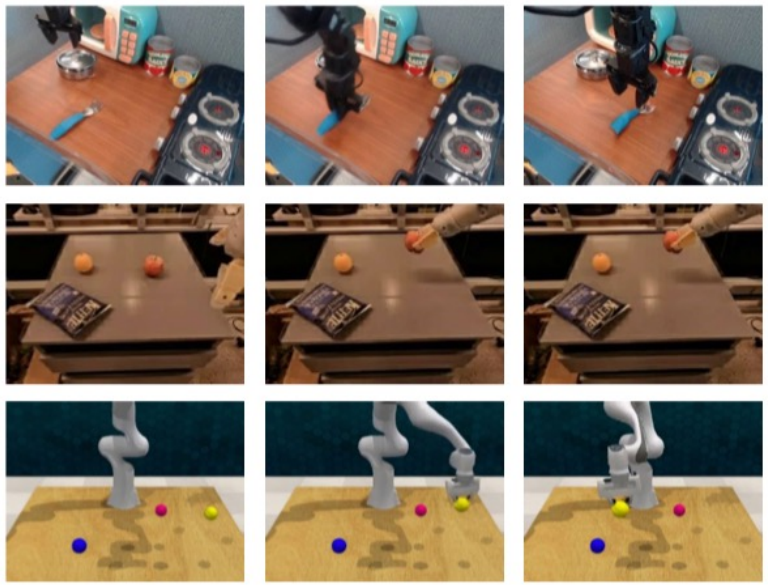
- Frame-level reward: it first selects a key goal frame from the generated video using CLIP, and then uses a forward-backward representation to estimate how likely a state-action pair is to reach that goal state, providing finer-grained and action-aware guidance



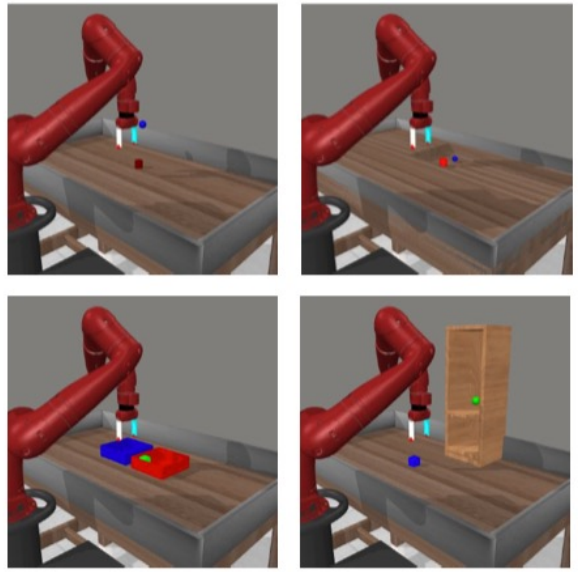
$$M(s, a, s', \pi_z) \approx F^\top(s, a, z)\, B(s')$$

Experimental Setups

- Evaluate the method on Meta-World, a standard benchmark for robotic manipulation
- Consider five medium-to-hard tasks: Pick Place, Pick Out of Hole, Bin Picking, Shelf Place, and Disassemble. To make the setting more challenging, each episode is limited to 256 steps

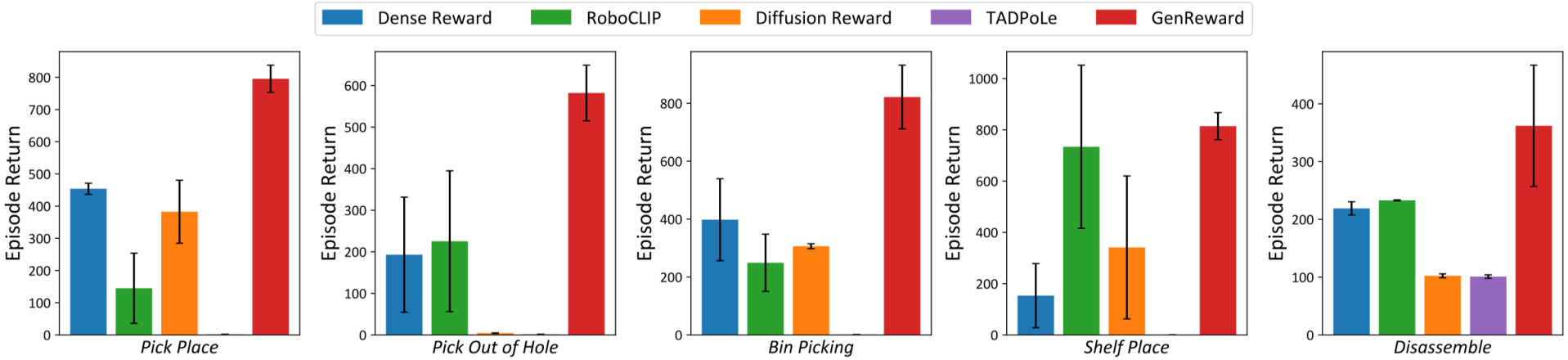


Generated Videos using CogVideoX



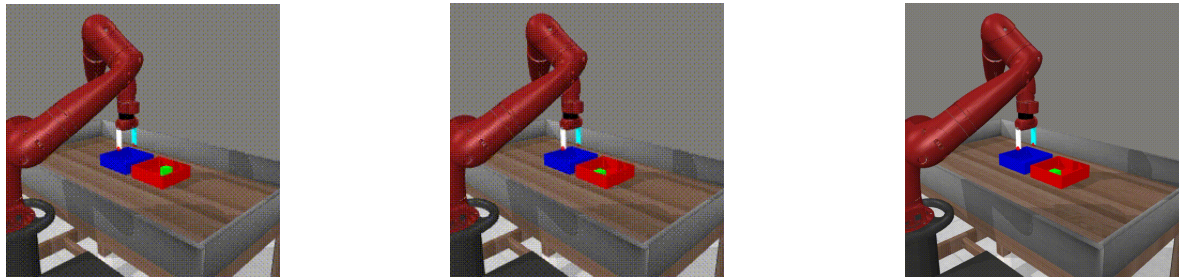
Downstream Manipulation Tasks

Results

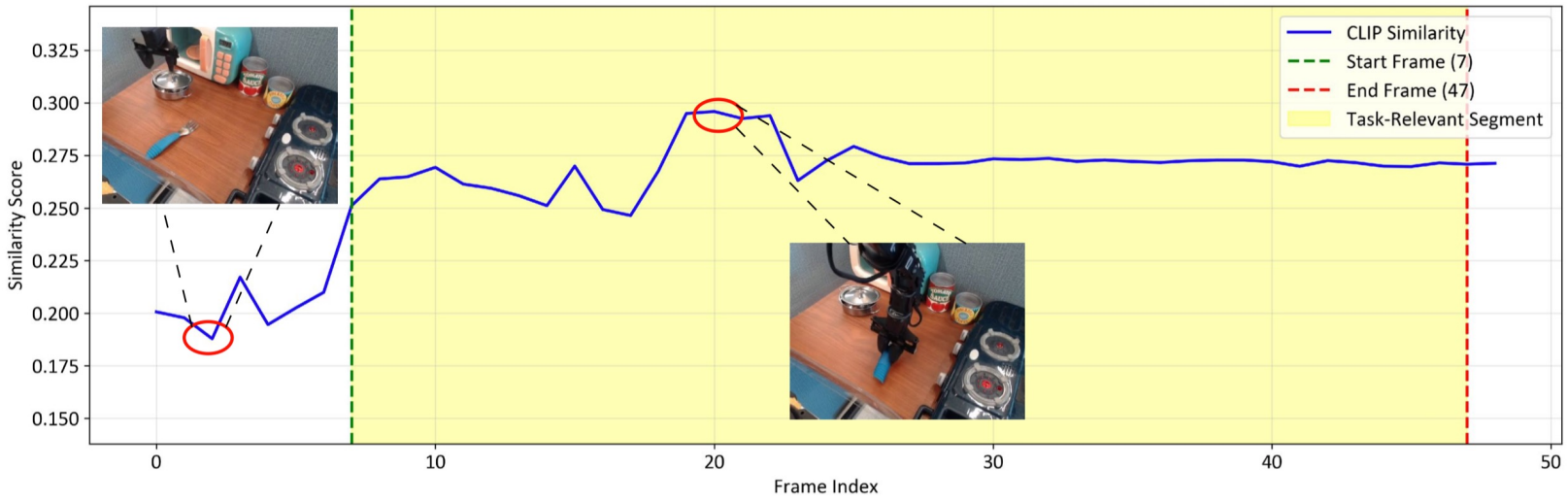


TADPole | Diffusion Reward | **GenReward**

Brown *CUHK* *Ours*



Results

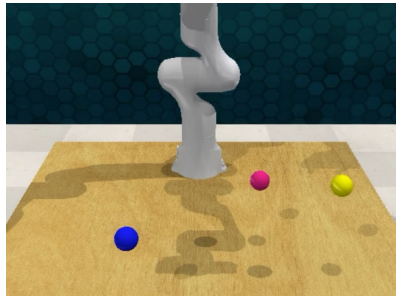


- Select the most task-relevant frame from the generated video on MetaWorld Pick Place based on CLIP similarity
- The selected frame provides a more precise goal image for forward-backward representation learning

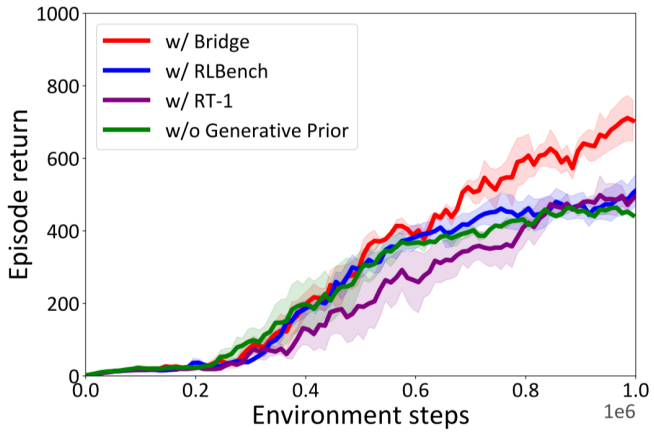
Results

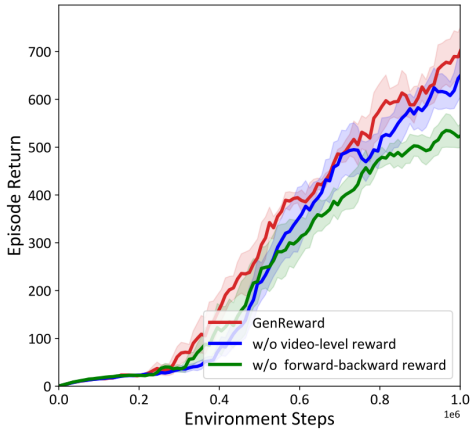
Bridge

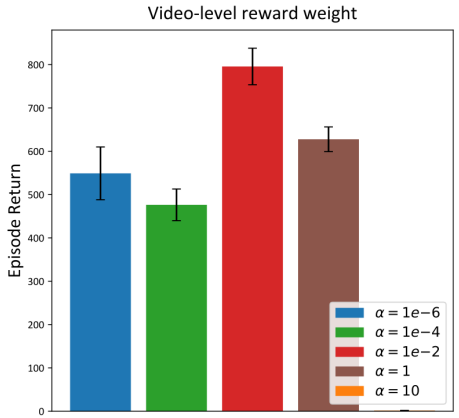
RT-1

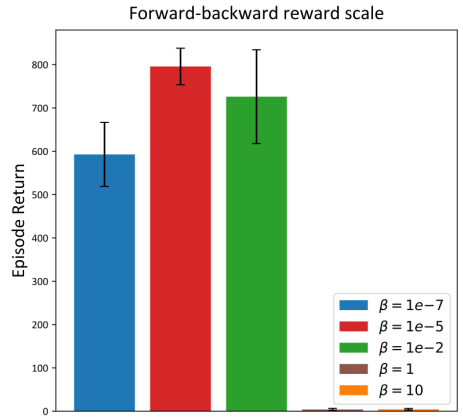


RLBench









- GenReward consistently improves over the baseline agent without generative priors
- It transfers world knowledge from the video diffusion model to guide downstream policy learning