Diffusion Reward: Learning Rewards via Conditional Video Diffusion

Tao Huang[1,2,5,*], Guangqi Jiang[1,3,*], Yanjie Ze[1], and Huazhe Xu[4,1,5]

[1] Shanghai Qi Zhi Institute
[2] The Chinese University of Hong Kong
[3] Sichuan University
[4] Tsinghua University, IIIS
[5] Shanghai AI Laboratory
diffusion-reward.github.io
taou.cs13@gmail.com, luccachiang@gmail.com.

Abstract. Learning rewards from expert videos offers an affordable and effective solution to specify the intended behaviors for reinforcement learning (RL) tasks. In this work, we propose Diffusion Reward, a novel framework that learns rewards from expert videos via conditional video diffusion models for solving complex visual RL problems. Our key insight is that lower generative diversity is exhibited when conditioning diffusion on expert trajectories. Diffusion Reward is accordingly formalized by the negative of conditional entropy that encourages productive exploration of expert behaviors. We show the efficacy of our method over robotic manipulation tasks in both simulation platforms and the real world with visual input. Moreover, Diffusion Reward can even solve unseen tasks successfully and effectively, largely surpassing baseline methods. Project page and code: diffusion-reward.github.io.

Keywords: Reward Learning from Videos · Robotic Manipulation · Visual Reinforcement Learning

1 Introduction

Reward specification is a fundamental challenge in reinforcement learning (RL), determining the effectiveness and alignment of an agent's learned behavior with the intended objectives. Manually designing dense reward functions is both labor-intensive and sometimes infeasible, particularly in real-world tasks such as robotic manipulation [40], where obtaining privileged state information is difficult. As an alternative, using sparse rewards is often favorable in these scenarios because of its low requirement for manual effort [29]. Nonetheless, the insufficient supervision provided by sparse rewards hinders the sample efficiency of RL, potentially imposing an unaffordable burden on data collection.

Learning reward functions from expert videos presents a promising solution because of the low effort of video collection and dense task-execution information

---

[*] Equal contribution to this work.
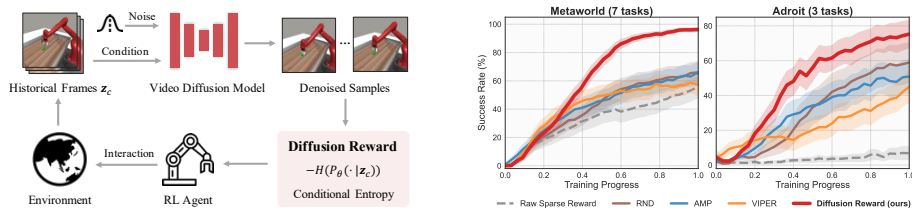
Fig. 1: **Overview of Diffusion Reward.** (*left*) We present a reward learning framework in RL using video diffusion models pretrained with expert videos. We perform diffusion processes conditioned on historical frames to estimate conditional entropy as rewards to encourage RL exploration of expert-like behaviors. (*right*) The mean success rate of 10 visual robotic manipulation tasks demonstrates the effectiveness of our proposed Diffusion Reward over 5 runs. Shaded areas are standard errors.

contained in the videos [6,54]. Among these, generative models have been naturally investigated by researchers to extract informative rewards unsupervisedly for RL training [32,45]. One classical approach builds on generative adversarial learning to learn a discriminative reward that distinguishes between agent and expert behaviors. While straightforward, these methods underutilize the temporal information, whose importance has been shown in solving RL [56], and performance is brittle to the adversarial training. In light of these issues, recent work pre-trains VideoGPT [49] to encode temporal information, and directly use the predicted log-likelihood as rewards [10]. However, they encounter difficulties in modeling complex expert video distributions and accordingly the proposed rewards, especially in tasks with intricate dynamics. As shown in Figure 2, a discernible decline in the learned return is evident for out-of-distribution expert videos, despite their alignment with optimal behaviors in in-distribution videos.

Video diffusion models [17] — which have exhibited remarkable performance in the domain of computer vision — have shown great power in capturing the complex distribution of videos, such as text-to-video generation [11,15] and video editing [5,26]. Recent studies have also investigated their potential in modeling expert videos to serve as generalizable planners in robotic manipulation tasks [9, 19,21]. Despite these advancements, extracting informative rewards from video diffusion models remains an understudied area, while shining great prospects in guiding RL to acquire expertise from videos and generalize to unseen tasks.

In this work, we propose **Diffusion Reward**, a reward learning framework that leverages video diffusion models to capture expert video distribution and pre-train dense reward function for visual RL. Our key insight is that diffusion conditioned on expert-like trajectories exhibits lower generative diversity than that conditioned on expert-unlike ones. This rationale naturally instructs RL exploration on expert-like behaviors by seeking low generative diversity. We accordingly estimate entropy conditioned on historical frames, which formalizes our insight, and augment it with a novelty-seeking reward [4] and spare environment reward to form dense rewards for efficient RL. In addition, to accelerate the reward inference, we perform latent diffusion processes by utilizing vector-quantized codes [13] for compressing high-dimensional observations.
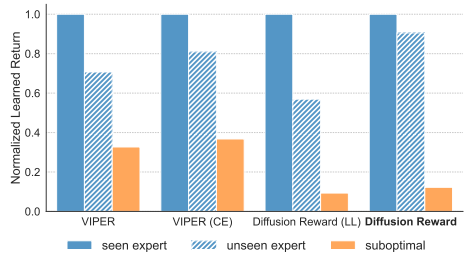
Fig. 2: Analysis of video models and rewards. *VIPER (CE)* and *Diffusion Reward (LL)* replace original rewards with conditional entropy (CE) and log-likelihood (LL), respectively. CE-based rewards assign near-optimal rewards to unseen expert videos. Such a boost is enhanced by the strong modeling ability of diffusion models. Results are averaged over 7 MetaWorld tasks. *Suboptimal* represents videos with 25% randomly-taken actions.
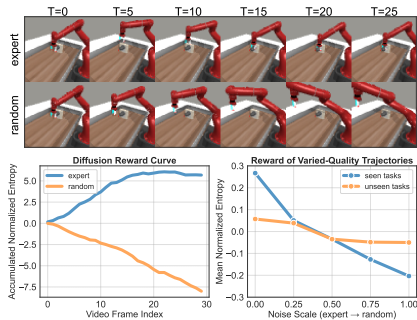
Fig. 3: Reward analysis. (*left*) Our learned rewards are aggregately higher for expert behaviors over 7 tasks from MetaWorld. (*right*) Varied-quality trajectories can be distinguished from seen and unseen tasks. (*top*) Examples of trajectories.

We empirically validate the effectiveness of our framework through experiments on 10 visual robotic manipulation tasks, including 7 gripper manipulation tasks from MetaWorld [51] and 3 dexterous manipulation tasks from Adroit [35], exhibiting 38% and 35% performance improvements over the best-performing baselines, respectively. Additionally, we surprisingly observe favorable zero-shot generalization performance of our learned reward on unseen tasks. Furthermore, reasonable rewards produced on real robot tasks and associated favorable offline RL performance demonstrate the real-world applicability of Diffusion Reward. Figure 1 overviews this work. The primary contributions of this work can be summarized as follows:

- We introduce Diffusion Reward, a novel reward learning framework that leverages the generative modeling capabilities of video diffusion models to provide informative dense reward signals for efficient RL.
- We demonstrate significant outperformance of our framework compared to baselines across robotic manipulation tasks in simulation and the real world.
- We observe that our pretrained reward model can instruct RL on unseen tasks and produce reasonable rewards on real robot tasks.

2   Related Work

Learning rewards from videos. Extracting rewards from expert videos provides an affordable but effective solution for reward specification in RL [6, 37, 38]. Several studies have endeavored to learn dense rewards that indicate task progress by measuring the distance between current observation and goal image in the latent space [3, 25, 54]. While promising, obtaining goal images out of simulation can be challenging, thus limiting their applicability to open-world tasks [24]. In contrast, generative models have been widely explored to extract

rewards without access to future information. One representative approach, belonging to state-only imitation learning [34, 44, 58], builds on generative adversarial learning [32, 45] to distinguish expert-like and expert-unlike behaviors with an online paradigm. This idea is further refined in [10], wherein a reward function is pretrained from videos, and the log-likelihood is predicted as rewards. Nonetheless, it encounters challenges in accurately modeling expert video distributions in complex tasks featuring intricate dynamics, thereby rendering them susceptible to generating unproductive rewards. In contrast, our methods leverage the powerful modeling abilities of diffusion models and estimate the negative of conditional as more discriminative rewards to expedite RL exploration.

Diffusion models for RL. Diffusion models have been widely investigated for RL to, for instance, improve the policy expresiveness [8, 14, 47], and augment experience [7, 52]. Apart from these, some works directly learn the diffusion models from offline data unconditional planners [19], or conditional planners specified by task returns [2]. The idea of conditional diffusion is further investigated in [9, 21] with text as task specification, where video diffusion models serve as planners associated with inverse dynamics. Unlike these methods, we intend to learn informative rewards via video diffusion models conditioned on historical frames to accelerate online RL. Such historical conditioning has been used in [18] to inform trajectory generation, which differs from our focus on reward learning as well. Our work is also close to [30], which extracts rewards from two diffusion models that fit different behaviors in state-based tasks. Differently, we pretrain reward models without action labels and additional task-agnostic data.

Pretraining for RL. Numerous methods have been explored for directly pretraining effective policies in RL from data. These methods include those employing offline RL techniques before online finetuning [22, 23, 27] as well as approaches learning behavioral (skill) priors [33, 36, 39]. However, they require the additional collection of rewards, action labels, or task-agnostic data, which poses certain sacrifices. In contrast, our approach focuses on pretraining rewards solely from action-free expert-only videos, alleviating the need for such additional data sources. Some studies have also explored pretraining visual representations using datasets similar to ours [28, 31, 48, 53]. Nevertheless, these approaches do not extract dense task-execution signals to guide RL.

3   Preliminaries

Problem formulation. We consider an RL agent that interacts with the environment modeled as a finite-horizon Markov Decision Process (MDP), which is defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}$ is the environment transition function, $\mathcal{R}$ is the reward function, and γ is the discount factor. The goal of the RL agent is to learn an optimal policy π that maximizes the expected return $\mathbb{E}[\sum_{k=0}^{K-1} \gamma^k r_k]$ over K timesteps.

In this work, we focus more specifically on high-dimensional state space with a binary sparse (i.e., 0/1 reward as a success indicator of the task). Particularly,

Data	Model	SSIM (↑)	PSNR (↑)	LPIPS (↓)
Real	VideoGPT	0.768 (± 0.141)	23.35 (± 4.13)	0.1017 (± 0.0670)
	VQ-Diffusion	**0.856** (± 0.114)	**27.46** (± 5.24)	**0.0596** (± 0.0450)
Sim.	VideoGPT	0.967 (± 0.021)	32.28 (± 3.15)	0.0067 (± 0.0032)
	VQ-Diffusion	**0.976** (± 0.009)	**33.55** (± 2.31)	**0.0053** (± 0.0029)

Table 1: Quantitative comparison of video models. Results are evaluated on real dexterous robotic manipulation videos and simulation videos from 7 MetaWorld tasks, demonstrating that the video diffusion model generates videos of higher quality than VideoGPT.

ϵ-greedy traj.	Distances as Diversity Metrics		
	SSIM (↑)	PSNR (↑)	LPIPS (↓)
0% (expert)	**0.941** (± 0.054)	**29.91** (± 5.52)	**0.0171** (± 0.0166)
25%	0.901 (± 0.063)	26.01 (± 4.62)	0.0371 (± 0.0304)
50%	0.894 (± 0.065)	25.65 (± 4.29)	0.0421 (± 0.0297)
75%	0.844 (± 0.066)	22.60 (± 2.53)	0.0645 (± 0.0370)
100% (rand.)	0.829 (± 0.730)	22.25 (± 3.43)	0.0597 (± 0.0364)

Table 2: Generative diversities. Expert trajectories over 7 MetaWorld tasks, seen during training, show the lowest generative diversity, which is proportional to distances among generated videos [57].

we consider RGB images $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ as the state observed by the agent. This setting is motivated by the real-world application of RL such as robotics, where vision-based sensory data is more available and specifying sophisticated reward often requires tedious hand-engineering, sometimes even infeasible. However, this poses great difficulty to an RL agent, as a large amount of interaction with environments is often required, known as sample efficiency.

Expert videos. To improve the sample efficiency of RL, we assume a set of unlabeled videos generated by the expert policies are accessible by the agent. Notably, these videos are action-free and gathered from multiple tasks without task identification. We denote it as $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^N\}$, where $\mathcal{D}^i$ is the demonstrated video set of task i. Each video set $\mathcal{D}^i$ consists of multiple expert videos $\tau = \{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_{K-1}\} \in \mathcal{S}^K$. Our goal is to pretrain effective reward functions from expert videos to accelerate the online exploration of RL. We present our reward learning framework in the following section.

4   Method

We introduce Diffusion Reward, a novel framework that learns rewards from expert videos via conditional video diffusion for solving downstream visual RL problems, as illustrated in Figure 1 and Algorithm 1. In Section 4.1, we first model the expert videos with the video diffusion model in the latent space. In Section 4.2, we formalize our key insight of Diffusion Reward by estimating the history-conditioned entropy and using its negative as rewards for RL.

4.1   Expert Video Modeling via Diffusion Model

Diffusion models [41] are probabilistic models that aim to model the data distribution by gradually denoising an initial distribution through a reverse diffusion process [16]. These models showcase their power in capturing highly complex distributions and generating samples that exhibit intricate dynamics, motion, and behaviors in RL literature [2,19]. Unlike prior works that model expert videos as planners, we aim to learn reward functions from the diffusion model trained on expert videos for RL. This motivates our video models to achieve fast inference speed and encode sufficient temporal information.

Algorithm 1 Diffusion Reward for online RL


    // Pretrain reward model from expert videos
1: Collect expert videos $\mathcal{D}$ from K tasks
2: Train diffusion model p_θ on expert videos $\mathcal{D}$
    // Downstream RL with learned rewards
3: **while** not converged **do**
4:     Act $a_k \sim \pi(\cdot|\boldsymbol{x}_k)$
5:     Generate M denoised samples $\tilde{\boldsymbol{z}}_k^{0:T} \sim p_\theta(\boldsymbol{z}_k^{0:T}|\boldsymbol{z}_c)$
6:     Estimate entropy $H(p_\theta(\cdot|\boldsymbol{z}_c))$ with $\tilde{\boldsymbol{z}}_k^{0:T}$ ▷ Eq. (3)
7:     Compute Diffusion Reward $r_k \leftarrow r_k^{\text{diff}}$ ▷ Eq. (5)
8:     Step environment $\boldsymbol{x}_{k+1} \sim \mathcal{T}(\boldsymbol{x}_k, a_k)$
9:     Store transition $(\boldsymbol{x}_k, a_k, r_k, \boldsymbol{x}_{k+1})$
10:    Update policy π and r^{rnd} with RL algorithm
11: **end while**
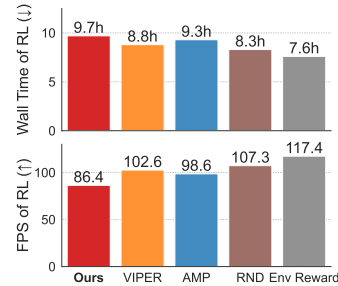
Fig. 4: **Time efficiency** on Door task with a NVIDIA A40 GPU.

Latent diffusion process. Specifically, we first train an encoder unsupervisedly from expert videos to compress the high-dimensional observations. Here we use the VQ-GAN method [12] to represent the image $\boldsymbol{x}$ with a vector-quantized latent code $\boldsymbol{z} = Q(E(\boldsymbol{x}))$, where E is the encoder and Q is the element-wise quantizer. The whole video is then represented by a sequence of latent variables $\tau = \{\boldsymbol{z}_0, \boldsymbol{z}_1, ..., \boldsymbol{z}_{K-1}\}$, where we overwrite the definition of τ without ambiguity. Subsequently, the forward process applies noise ϵ in the latent space at each time step $t \in 0, ..., T$ to the data distribution $\boldsymbol{z}_k$, resulting in a noisy sample $\boldsymbol{z}_k^t$, where $\boldsymbol{z}_k^t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_k^0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, and $\bar{\alpha}$ is the accumulation of the noise schedule over past timesteps. To fit data distribution, we learn a parameterized variant of noise predictor $\epsilon_\theta(\boldsymbol{z}_k^t)$ that aims to predict the noise ϵ during the forward process. Then the parameterized reverse process $p_\theta(\boldsymbol{z}_k^{t-1}|\boldsymbol{z}_k^t)$ can be approximated and performed by iterative denoising an initial distribution.

Historical frames as condition. To utilize the temporal information from expert videos with the power of video diffusion, we further condition the reverse process with historical frames, i.e., $p_\theta(\boldsymbol{z}_k^{t-1}|\boldsymbol{z}_k^t, \boldsymbol{z}_c)$, where $\boldsymbol{z}_c$ is the concatenation of ℓ historical frames $[\boldsymbol{z}_{k-\ell}, ..., \boldsymbol{z}_{k-1}]$ to ensure higher computation efficiency while maintaining enough historical information. This can also be viewed as matching the distribution of expert and agent trajectories [10]. Subsequently, one can perform the reverse process from a randomly sampled noise to generate the latent code of future frames and decode the code for video prediction. In this work, we use VQ-Diffusion [13] as our choice of video diffusion model due to its good performance and compatibility with vector-quantized latent code, but our framework can in principle adopt any off-the-shelf video diffusion models. Specifically, we first tokenize each latent code $\boldsymbol{z}$ indexed by its indices that specify the respective entry in the learned codebook of VQ-GAN, and take as a condition embedding the concatenated tokens. The embedding is then fed to the decoder that contains multiple transformer blocks and softmax layers with cross attention, providing context information for video diffusion.

4.2   Conditional Entropy as Rewards

While previous studies have explored the use of log-likelihood as rewards with video prediction models, exemplified by works such as VIPER [10], this approach encounters two primary challenges. Firstly, it has difficulty in accurately modeling the distribution of complex expert videos featuring intricate dynamics, as shown in Table 1. Secondly, the moderate video modeling ability leads to undesired rewards. This issue is evident in Figure 2, where a significant drop in learned rewards between in-distribution expert videos and out-of-distribution ones, though both sets of videos demonstrate optimal behaviors.

Key insight behind Diffusion Reward. We address these challenges by harnessing the great generative capability of video diffusion models. Our observations indicate increased generation diversity with unseen historical observations (rand.) and reduced diversity with seen ones (expert), as shown in Table 2. This gives rise to the key insight of our proposed Diffusion Reward: diffusion conditioned on expert-like trajectories exhibits lower diversity where the agent ought to be rewarded more, and the opposite holds on unexpert-unlike ones. Such diversity discrimination incentivizes RL agents to chase expert-like behaviors.

Estimation of conditional entropy. To formalize this idea, we seek to estimate the negative conditional entropy given historical frames $\boldsymbol{z}_c$, which in principle captures the conditional generation diversity:

$$-H(p_\theta(\cdot|\boldsymbol{z}_c)) = \mathbb{E}_{p_\theta(\cdot|\boldsymbol{z}_c)}\left[\log p_\theta(\boldsymbol{z}_k|\boldsymbol{z}_c)\right]. \tag{1}$$

One primary challenge is the computation of the entropy in Eq. (1), which is intractable as we have no explicit form of the conditional distribution [43]. Therefore, we instead estimate the variational bound of the entropy. Specifically, we first present the variational bound of conditional log-likelihood as follows:

$$\log p_\theta(\boldsymbol{z}_k^0|\boldsymbol{z}_c) \geq \mathbb{E}_{q(\boldsymbol{z}_k^{0:T}|\boldsymbol{z}_c)}\left[\log \frac{p_\theta(\boldsymbol{z}_k^{0:T})}{q(\boldsymbol{z}_k^{1:T}|\boldsymbol{z}_k^0, \boldsymbol{z}_c)}\right], \tag{2}$$

where $\boldsymbol{z}_k^0$ is the denoised prediction of current observation $\boldsymbol{z}_k$. This bound can be estimated via noise predictor ϵ_θ [16,42], or with the closed-form distribution [20, 41] (e.g., discrete multivariate distribution). We use the latter one as our choice of estimation because of its better compatibility with VQ-Diffusion.

Next, to estimate the whole entropy, we denoise from randomly sampled noise and generate the latent variables $\tilde{\boldsymbol{z}}_k^{0:T} \sim p_\theta(\boldsymbol{z}_k^{0:T}|\boldsymbol{z}_c)$, repeating with M times. Subsequently, we use the generated samples from randomized noise $\tilde{\boldsymbol{z}}_k^T$ (e.g., random tokens) to estimate the whole conditional entropy term as follows:

$$r^{\mathrm{ce}}(\boldsymbol{x}_{k-1}) = \frac{1}{M}\sum_{j=1}^{M}\log \frac{p_\theta(\tilde{\boldsymbol{z}}_k^{0:T})}{q(\tilde{\boldsymbol{z}}_k^{1:T}|\tilde{\boldsymbol{z}}_k^0, \boldsymbol{z}_c)}, \tag{3}$$

We visualize the aggregated entropy-based reward in Figure 3 and present curves of each task in the Appendix. The results show that conditional entropy can successfully capture the varied generative diversity on different videos, echoed with the aforementioned insight of Diffusion Reward. Notably, here we use the standardized entropy reward $\bar{r}^{\mathrm{ce}}$ to mitigate the burden of hyperparameter tuning, as we observe that the scale of conditional entropy varies significantly across different tasks and domains, partially attributed to the varied objects and environment dynamics. Concretely, the conditional entropy is standardized by the empirical mean and standard deviation of the expert videos:

$$\bar{r}^{\mathrm{ce}} = (r^{\mathrm{ce}} - \mathrm{mean}(\mathcal{D}, r^{\mathrm{ce}})) \,/\, \mathrm{std}(\mathcal{D}, r^{\mathrm{ce}}). \tag{4}$$

Exploration reward. As the reward $\bar{r}^{\mathrm{ce}}$ incentivizes the agent to mimic the behavioral patterns of the expert, the exploration may still be prohibitively challenging in complex tasks with high-dimensional input [1, 10]. To alleviate this issue, we incorporate RND [4] as the exploration reward, termed as r^{rnd}.

Diffusion Reward. Lastly, we combine our proposed diffusion-based entropy reward with the exploration reward and the raw sparse task reward r^{spar} from the environment, yielding the following composite reward function:

$$r^{\mathrm{diff}} = (1 - \alpha) \cdot \bar{r}^{\mathrm{ce}} + \alpha \cdot r^{\mathrm{rnd}} + \beta \cdot r^{\mathrm{spar}}, \tag{5}$$

where α and β are the reward coefficients that steer the effect of each reward. The inclusion of r^{spar} is crucial, as the complete absence of environmental supervision can impede progress in tackling complex tasks. Furthermore, we opt for a single coefficient α to balance $\bar{r}^{\mathrm{ce}}$ and r^{rnd}, thereby alleviating the burden of hyperparameter tuning. Meanwhile, the coefficient of task reward β is set as 1 for all tasks. We present negligible empirical distinctions between this formulation and the one featuring an additional coefficient, and effect of β in the Appendix.

4.3   Training Details

We first provide the expert videos from various tasks as the whole dataset for pertaining reward models, which can be generated by scripted policies or other means. Then, we first use VQ-GAN [12] to train the encoder, associated with 8×8 size of codebook across all domains, with an additional perceptual loss [55] calculated by a discriminator to increase perceptual quality. We subsequently use VQ-Diffusion [13] for training the conditional video diffusion model, where the number of historical frames ℓ is set as 2 for all tasks. For reward inference, the reward coefficient is set as 0.95 for all tasks except 0 for the Pen task. We decrease α for the Pen task because a high exploration reward can misguide the robot hand to drop the pen, while a lower one encourages in-hand manipulation. Moreover, during the diffusion process, we also use a DDIM-like sampling strategy [42] to accelerate the diffusion process with denoise steps set as 10 and repetition of diffusion M as 1 for all tasks, which shows fair performance and retains high inference speed in Figure 4. More details on hyperparameters, network architectures, and downstream RL can be found in the Appendix.

(a) MetaWorld (7 tasks)   (b) Adroit (3 tasks)
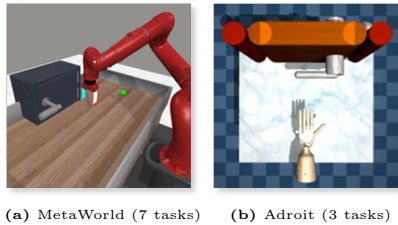
Fig. 5: Task visualization. We evaluate methods on 10 challenging visual RL tasks from MetaWorld and Adroit with visual input and sparse rewards. Tasks are chosen to cover a wide range of manipulation skills.
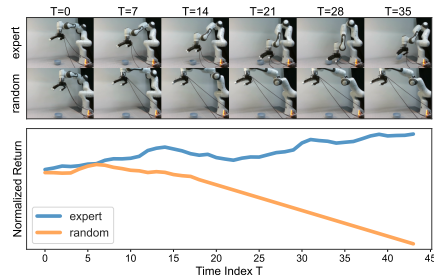
Fig. 6: Reward curve of real robot videos. Our method assigns higher rewards to expert videos than to random ones.

5 Experiments

5.1 Experimental Setup

Simulation environments. We intend to demonstrate the effectiveness of Diffusion Reward on 10 complex visual robotic manipulation tasks, including 7 gripper manipulation tasks from MetaWorld [51] and 3 dexterous hand manipulation tasks from Adroit [35], as visualized in Figure 5. We choose these two simulation environments because of their task diversity and complexity. Each task is associated with 64×64-dimensional RGB images, ± 4 pixel shift augmentation [50], and $0/1$ sparse task reward. For pertaining reward models, we collect 20 expert videos for each Metaword task (5039 frames in total) via the scripted policy provided by the official repository, and 50 for Adroit (3346 frames in total) via the policies trained with performant RL method [46], while our method has emerging low demand for the data scale (see the Appendix). For downstream RL training, the interaction budget of the Adroit task is set as 3 million to ensure convergence of using our rewards, and is respectively set for the MetaWorld task due to the noticeably varied task complexity.

Baselines. We compare our method against the followings:

- **Raw Sparse Reward** that uses the sparse task reward. This comparison tests the benefit of adding learned rewards from videos.
- **Random Network Distillation** (RND, [4]) that encourages exploration with a novelty-seeking reward only. This comparison tests the benefit of rewarding the agent with our pretrained reward.
- **Adversarial Motion Priors** (AMP, [32]) that online learns a discriminator to discern agent behaviors and expert behaviors based on current observations. This comparison tests the benefit of encoding temporal information in learned reward and utilization of novelty-seeking reward.
- **Video Prediction Rewards** (VIPER, [10]) and its standardized variant (VIPER-std) that use VideoGPT [49] as video prediction model and predicted log-likelihood of agent observation as reward. This comparison tests
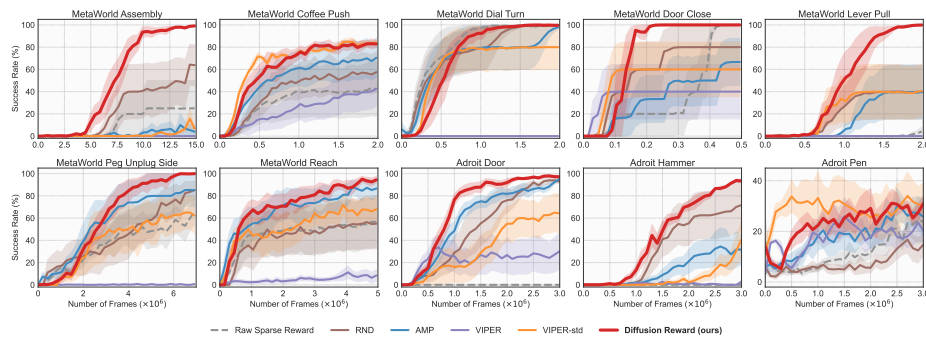
Fig. 7: Main results. Success rate for our method and baselines on 7 gripper manipulation tasks from MetaWorld and 3 dexterous manipulation tasks from Adroit with image observations. Our method achieves prominent performance on all tasks, and significantly outperforms baselines on complex door and hammer tasks. Results are means of 5 runs with the standard error (shaded area).

> the benefit of utilizing the generative capability of video diffusion models and the conditional entropy as a more explorative reward.

For a fair comparison, all methods are augmented with task rewards, which generally improve their performance. Besides, we use DrQv2 [50] as the RL backbone and maintain all settings except reward-pretraining (if exist) identical.

5.2   Main Results

Comparison with non-pretraining methods. We present the learning curves of success rates for each method over two simulation domains in Figure 1 and 7. The results show that solely using sparse task rewards enables progress in relatively straightforward tasks such as Reach and Dial Turn. However, it encounters significant challenges in more complex ones, exemplified by the Door and Hammer within the dexterous hand manipulation domain. The incorporation of pure novelty-seeking rewards (RND) unsurprisingly enhances the RL agent's exploration, particularly evident in addressing moderately complex tasks like Coffee Push and Assembly. Nevertheless, the performance on dexterous hand manipulation tasks is still unsatisfactory due to the lack of expert-instructed exploration at prohibitively large configuration space. Conversely, AMP explicitly employs expert-guided rewards to incentivize the exploration of expert-like behaviors. While it generally outperforms RND in simpler tasks, its efficacy diminishes in more complex ones such as Lever Pull and Hammer.

Comparison with reward pretraining method. The above observations suggest that the combination of expert-instructed rewards and novelty-seeking rewards is likely to perform favorably. Despite the incorporation of such a combination in the VIPER reward, its empirical performance unexpectedly falls short of both RND and AMP. We posit this is attributed to the significantly varied scales of expert-instructed rewards (i.e., log-likelihood), which may eliminate the
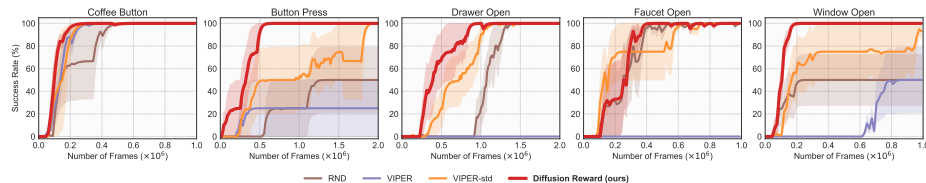
Fig. 8: **Success rate curves on 5 unseen MetaWorld tasks.** Diffusion Reward could generalize to unseen tasks directly and produce reasonable rewards, largely surpassing other baselines. Results are means of 4 runs with the standard errors (shaded area).

effect of novelty-seeking rewards. VIPER-std alleviates this issue and shows better performance than the non-standardized version, especially on Coffee Push and Pen, while still underperforming our method. These results are in line with two limitations outlined in Section 4.2 and further verified in Figure 2, indicating VIPER's struggle in capturing complex video distributions within intricate tasks. In sharp contrast to VIPER, our proposed method not only utilizes the modeling capabilities of diffusion models, but also uses conditional entropy as a reward function to accelerate exploration. The results showcase remarkable performance improvements of **38%** and **35%** over the best-performing baselines with the same training steps across MetaWorld and Adroit, respectively.

5.3 Zero-shot Reward Generalization

Reward analysis. The advances of video diffusion models in generating samples beyond their training data, as exemplified in text-to-image video generation [11, 15], motivates our exploration into the potential of Diffusion Reward to generalize to previously unseen tasks. To investigate this, we first visuale the learned returns of trajectories with varying qualities derived from 15 diverse unseen tasks from MetaWorld (see Figure 3). Expectedly, the distinctions between trajectories of varied qualities are less noticeable than those observed in tasks seen during pretraining. Nonetheless, our pretrained reward model still exhibits a consistent trend where expert-like behaviors receive relatively high learned returns, owing to the generalization prowess of the video diffusion model.

RL performance. Subsequently, we directly apply our pretrained rewards to 5 tasks (more in the Appendix) involving diverse objects without additional tuning. The outcomes, as shown in Figure 8, affirm that our reward effectively guides RL exploration and largely outperforms other baseline methods across all tasks. Notably, our approach proves helpful in constraining the exploration space of RND due to its retained ability to discriminate between expert-like and expert-unlike behaviors. Meanwhile, VIPER generalizes worse than ours on most tasks, especially on complex ones like opening objects, partially attributed to the combined limitations of the adopted video models and log-likelihood rewards. Here we exclude comparison with AMP because it learns reward online, which requires unavailable expert data on the target task. These results not only verify the efficacy of our method, but also point towards the potential of employing
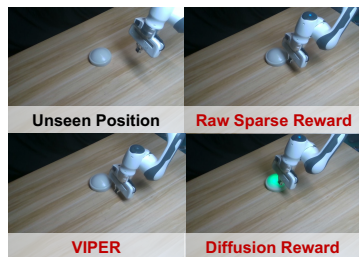
Fig. 9: Generalization in real-world task. Policy trained with our method can successfully press the button at an unseen position.
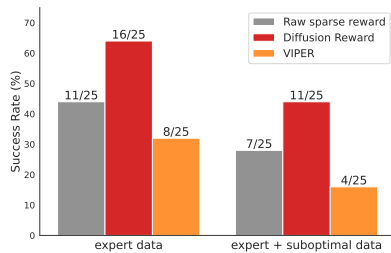
Fig. 10: Success rate of real-world offline RL. Our method significantly outperforms baselines on the Button Press task given different offline data.

larger diffusion models and integrating other modalities (e.g., text-based task specification) to enhance the generalization capabilities of our approach further.

5.4   Real Robot Evaluation

Reward analysis. We consider the real robot task that aims to pick up a bowl on the table. To train and test our reward model, we collect 20 real robot videos (10 expert and 10 random) with an Allegro hand, a Franka arm, and a RealSense, operated by humans. Visualization results in Figure 6 show that Diffusion Reward can appropriately assign expert videos relatively higher rewards and, in contrast, random videos lower rewards, indicating the potential of our method for real-world robot manipulation tasks. We present more curves of multiple trials and reward analysis related to baselines in the Appendix.

Offline RL performance. We conduct offline RL evaluations in the Button Press task using a parallel gripper as the end-effector. Expert videos are used to train each reward model for subsequent offline RL training (see the Appendix for more details). Our results show that Diffusion Reward produces a more generalizable policy for pressing the button at an unseen position, as illustrated in Figure 9. Moreover, Figure 10 quantitatively demonstrates that Diffusion Reward significantly outperforms VIPER in terms of real-world applicability.

5.5   Ablation Studies

As shown in Figure 11, we ablate key design choices in our proposed framework in the previous experiments, aiming to reveal more insights into the quantitative performance of our method. We present more detailed analyses below.

Conditional entropy with diffusion model. As our method has demonstrated better performance than VIPER, we take a further step to see the joint effect of the reward types and the video prediction models. Specifically, we systematically evaluate all possible combinations of conditional entropy and log-likelihood as reward signals, each paired with either diffusion-based or transformer-based video prediction models. Note that the same vector-quantized encoder is
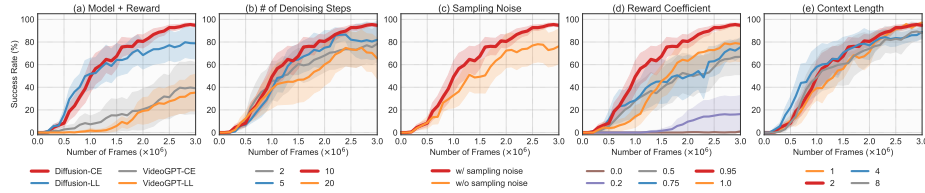
Fig. 11: Ablations. Success rate curves for ablated versions of Diffusion Reward, aggregated over Door and Hammer from Adroit. (a) We test the combinations of generative models and rewards to show the benefit of estimating conditional entropy with the diffusion model. (b) We ablate the choice of the number of denoising steps. (c) We demonstrate that the inherent randomness of Diffusion Reward from the reverse process helps RL exploration. (d) We ablate the choice of reward coefficient. (e) We test the effect of the number of conditional frames. Results are means of 3 seeds with the standard error (shaded area). **Red** is our default.

used for all models, ensuring the variations are solely attributable to the chosen video prediction model and reward. The outcomes consistently align with the observations delineated in Figure 2, indicating two-fold conclusions: (1) video diffusion models are more adept at capturing the complex distribution of expert videos in complex tasks, thus resulting in more informative rewards; (2) employing conditional entropy as rewards prove more productive in RL exploration than using log-likelihood, partially owing to the more generalizable reward inference on trajectories unseen in reward pretraining stage.

Denoising steps. The number of timesteps involved in the reverse process governs the quality and diversity of generated frames [42]. This study seeks to investigate its impact on the derived reward and subsequent RL performance by gradually increasing the number of denoising steps from 2 to 20. The findings indicate that an intermediate choice, approximately around 10 steps, achieves the best performance. This suggests that an intermediate choice balances generative quality and diversity well, thereby producing effective rewards for RL exploration. Furthermore, we observe that the speed of reward inference declines with an increase in denoising steps, suggesting the benefit of adopting advanced techniques to expedite the diffusion process if requiring more denoising steps.

Sampling noise in diffusion process. We hypothesize that the randomness in the diffusion process can accelerate RL exploration, akin to the stochastic characteristics of maximum entropy RL [59]. To substantiate this point, we design a variant of Diffusion Reward wherein the sampling noise is deliberately set as 0 during the reverse process, ensuring the reward becomes deterministic given the identical historical observations. The outcomes show a discernible degradation in performance when employing a deterministic reward. Notably, this decline in performance aligns with results observed when combining the diffusion model with log-likelihood rewards, wherein the rewards are also deterministic. Consequently, our findings demonstrate that the inherent randomness of Diffusion Reward from the reverse process indeed contributes to RL exploration.

Reward coefficient α**.** The reward coefficient α determines the relative importance of the conditional entropy reward against novelty-seeking. We investigate the effect of this parameter by gradually decreasing the value of α from 1 to 0. The results show that α around 0.95 achieves the best performance, while too-large ones (akin to RND only) and too-small ones (akin to no RND) exhibit significant performance drops. This suggests the domination of Diffusion Reward may still result in getting stuck to local optima, while our proposed reward effectively helps the RL agent to narrow down the wide exploration space of novelty-seeking rewards.

Context length. The number of historical frames determines the extent of temporal information being encoded during video diffusion, thus influencing the generating process of video diffusion and induced reward inference. To investigate its effect on downstream RL, we test different choices of context length. The results suggest that opting for 1 or 2 historical frames proves sufficient to generate highly effective rewards, owing to the robust generative capabilities inherent in the diffusion model. Interestingly, a marginal decline in performance is observed when the context length is extended to 4 or 8 frames. This phenomenon may be attributed to potential overfitting to expert trajectories, resulting in inferred rewards that exhibit suboptimal generalization to previously unseen trajectories.

6   Conclusion

In this work, we propose Diffusion Reward, a novel framework that extracts dense rewards from a pretrained conditional video diffusion model for RL tasks. We first pretrain a video diffusion model using expert videos and observe that the entropy of the predicted distribution well discriminates the expert-level trajectories and under-expert-level ones. Consequently, we use its standardized entropy, plus the exploration reward and the sparse environmental reward, as an informative reward signal. We evaluate Diffusion Reward across 10 visual robotic manipulation tasks from MetaWorld and Adroit and observe prominent performance improvements over 2 domains. Additionally, we present results of reasonable rewards generated in real robot tasks and associated promising offline RL performance, highlighting the real-world applicability of our method. Furthermore, we demonstrate that our pretrained rewards can instruct RL to solve unseen tasks successfully and effectively, largely surpassing baseline methods. This underscores the potential of large-scale pretrained diffusion models in reward generalization.

Limitation and future work. Future work will leverage larger diffusion models sourced from a wider dataset to solve diverse simulation and real-world tasks. The incorporation of additional modalities, such as language, will be explored to augment the generalization capabilities of Diffusion Reward. In addition, enhancing the diffusion-based reward itself, including strategies to balance entropy reward and exploration reward, the estimations of conditional entropy, and accelerating reward inference speed, holds promise for yielding better outcomes.

Acknowledgements

References

1. Adeniji, A., Xie, A., Sferrazza, C., Seo, Y., James, S., Abbeel, P.: Language reward modulation for pretraining reinforcement learning. arXiv preprint arXiv:2308.12270 (2023)
2. Ajay, A., Du, Y., Gupta, A., Tenenbaum, J., Jaakkola, T., Agrawal, P.: Is conditional generative modeling all you need for decision-making? In: International Conference on Learning Representations (ICLR) (2023)
3. Alakuijala, M., Dulac-Arnold, G., Mairal, J., Ponce, J., Schmid, C.: Learning reward functions for robotic manipulation by observing humans. In: IEEE International Conference on Robotics and Automation (ICRA) (2022)
4. Burda, Y., Edwards, H., Storkey, A.J., Klimov, O.: Exploration by random network distillation. In: International Conference on Learning Representations (ICLR) (2019)
5. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
6. Chen, A.S., Nair, S., Finn, C.: Learning generalizable robotic reward functions from "in-the-wild" human videos. In: Robotics: Science and Systems (RSS) (2021)
7. Chen, Z., Kiami, S., Gupta, A., Kumar, V.: Genaug: Retargeting behaviors to unseen situations via generative augmentation. arXiv preprint arXiv:2302.06671 (2023)
8. Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. In: Robotics: Science and Systems (RSS) (2023)
9. Du, Y., Yang, M., Dai, B., Dai, H., Nachum, O., Tenenbaum, J.B., Schuurmans, D., Abbeel, P.: Learning universal policies via text-guided video generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
10. Escontrela, A., Adeniji, A., Yan, W., Jain, A., Peng, X.B., Goldberg, K., Lee, Y., Hafner, D., Abbeel, P.: Video prediction models as rewards for reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
11. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
12. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
13. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
14. Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J.G., Levine, S.: Idql: Implicit q-learning as an actor-critic method with diffusion policies. arXiv preprint arXiv:2304.10573 (2023)

15. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
17. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
18. Hu, J., Sun, Y., Huang, S., Guo, S., Chen, H., Shen, L., Sun, L., Chang, Y., Tao, D.: Instructed diffuser with temporal condition guidance for offline reinforcement learning. arXiv preprint arXiv:2306.04875 (2023)
19. Janner, M., Du, Y., Tenenbaum, J.B., Levine, S.: Planning with diffusion for flexible behavior synthesis. In: International Conference on Machine Learning (ICML) (2022)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
21. Ko, P.C., Mao, J., Du, Y., Sun, S.H., Tenenbaum, J.B.: Learning to Act from Actionless Video through Dense Correspondences. arXiv preprint arXiv:2310.08576 (2023)
22. Kun, L., He, Z., Lu, C., Hu, K., Gao, Y., Xu, H.: Uni-o4: Unifying online and offline deep reinforcement learning with multi-step on-policy optimization. In: International Conference on Learning Representations (ICLR) (2023)
23. Lee, S., Seo, Y., Lee, K., Abbeel, P., Shin, J.: Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In: Conference on Robot Learning (CoRL) (2022)
24. Lynch, C., Sermanet, P.: Language conditioned imitation learning over unstructured data. In: Robotics: Science and Systems (RSS) (2020)
25. Ma, Y.J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., Zhang, A.: Vip: Towards universal visual reward and representation via value-implicit pre-training. In: International Conference on Learning Representations (ICLR) (2023)
26. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
27. Nair, A., Gupta, A., Dalal, M., Levine, S.: Awac: Accelerating online reinforcement learning with offline datasets. arXiv preprint arXiv:2006.09359 (2020)
28. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. In: Conference on Robot Learning (CoRL) (2023)
29. Nasiriany, S., Pong, V., Lin, S., Levine, S.: Planning with goal-conditioned policies. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
30. Nuti, F., Franzmeyer, T., Henriques, J.F.: Extracting reward functions from diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
31. Parisi, S., Rajeswaran, A., Purushwalkam, S., Gupta, A.: The unsurprising effectiveness of pre-trained vision models for control. In: International Conference on Machine Learning (ICML) (2022)
32. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. ACM Trans. Graph (ToG). (2021)
33. Pertsch, K., Lee, Y., Lim, J.: Accelerating reinforcement learning with learned skill priors. In: Conference on Robot Learning (CoRL) (2021)

34. Radosavovic, I., Wang, X., Pinto, L., Malik, J.: State-only imitation learning for dexterous manipulation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021)
35. Rajeswaran, A., Kumar, V., Gupta, A., Schulman, J., Todorov, E., Levine, S.: Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In: Robotics: Science and Systems (RSS) (2018)
36. Rao, D., Sadeghi, F., Hasenclever, L., Wulfmeier, M., Zambelli, M., Vezzani, G., Tirumala, D., Aytar, Y., Merel, J., Heess, N., et al.: Learning transferable motor skills with hierarchical latent mixture policies. In: International Conference on Learning Representations (ICLR) (2021)
37. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S.: Time-contrastive networks: Self-supervised learning from video. In: IEEE International Conference on Robotics and Automation (ICRA) (2017)
38. Sermanet, P., Xu, K., Levine, S.: Unsupervised perceptual rewards for imitation learning. arXiv preprint arXiv:1612.06699 (2016)
39. Singh, A., Liu, H., Zhou, G., Yu, A., Rhinehart, N., Levine, S.: Parrot: Data-driven behavioral priors for reinforcement learning. In: International Conference on Learning Representations (ICLR) (2020)
40. Singh, A., Yang, L., Hartikainen, K., Finn, C., Levine, S.: End-to-end robotic reinforcement learning without reward engineering. arXiv preprint arXiv:1904.07854 (2019)
41. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML) (2015)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2022)
43. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
44. Torabi, F., Warnell, G., Stone, P.: Behavioral cloning from observation. In: International Joint Conference on Artificial Intelligence (IJCAI) (2018)
45. Torabi, F., Warnell, G., Stone, P.: Generative adversarial imitation from observation. arXiv preprint arXiv:1807.06158 (2018)
46. Wang, C., Luo, X., Ross, K.W., Li, D.: Vrl3: A data-driven framework for visual deep reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
47. Wang, Z., Hunt, J.J., Zhou, M.: Diffusion policies as an expressive policy class for offline reinforcement learning. In: International Conference on Learning Representations (ICLR) (2023)
48. Xiao, T., Radosavovic, I., Darrell, T., Malik, J.: Masked visual pre-training for motor control. arXiv preprint arXiv:2203.06173 (2022)
49. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
50. Yarats, D., Fergus, R., Lazaric, A., Pinto, L.: Mastering visual continuous control: Improved data-augmented reinforcement learning. In: International Conference on Learning Representations (ICLR) (2022)
51. Yu, T., Quillen, D., He, Z., Julian, R.C., Hausman, K., Finn, C., Levine, S.: Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Conference on Robot Learning (CoRL) (2020)

52. Yu, T., Xiao, T., Stone, A., Tompson, J., Brohan, A., Wang, S., Singh, J., Tan, C., Peralta, J., Ichter, B., et al.: Scaling robot learning with semantically imagined experience. arXiv preprint arXiv:2302.11550 (2023)
53. Yuan, Z., Xue, Z., Yuan, B., Wang, X., Wu, Y., Gao, Y., Xu, H.: Pre-trained image encoder for generalizable visual reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
54. Zakka, K., Zeng, A., Florence, P., Tompson, J., Bohg, J., Dwibedi, D.: Xirl: Cross-embodiment inverse reinforcement learning. In: Conference on Robot Learning (CoRL) (2022)
55. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
56. Zheng, R., Wang, X., Sun, Y., Ma, S., Zhao, J., Xu, H., Daum'e, H., Huang, F.: Taco: Temporal latent action-driven contrastive loss for visual reinforcement learning. arXiv preprint arXiv:2306.13229 (2023)
57. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. Advances in Neural Information Processing Systems (NeurIPS) (2017)
58. Zhu, Z., Lin, K., Dai, B., Zhou, J.: Off-policy imitation learning from observations. Advances in Neural Information Processing Systems (NeurIPS) (2020)
59. Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K., et al.: Maximum entropy inverse reinforcement learning. In: Association for the Advancement of Artificial Intelligence (AAAI) (2008)