Goal-Driven Reward by Video Diffusion Models for Reinforcement Learning

Supplementary Material

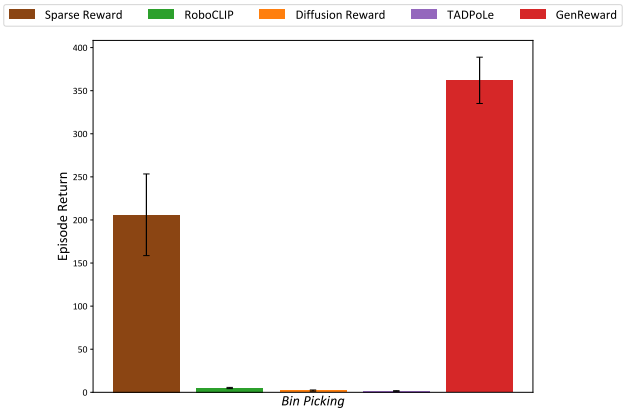

Figure A. Performance on Meta-World *Bin Picking* under sparse reward setting.

Table A. Performance comparison across various environments.

Model	Raw Reward	RoboCLIP	Diffusion Reward	GenReward
Walker Walk	640 ±74	695±94	28±2	**782 ± 110**
Hopper Stand	646 ± 161	589±44	776±58	**821±96**
Adroit Door	60±20	70±20	0±0	**90±10**
MW Reach	25±5	45±5	60±10	**85±5**

A. Additional Results

A.1. Results with Sparse Rewards

Different from dense reward settings, we consider a setup with a sparse reward function for the Meta-World tasks. Under this setting, the agent receives a reward every 64 steps, with the reward set to 0 before that. We present quantitative results of sparse rewards in Figure A. It can be observed that GenReward still achieves consistent improvements compared to other baselines even with sparse rewards, demonstrating its effectiveness.

A.2. Evaluation on Challenging Benchmark

We reimplement GenReward with the model-free *DrQ-v2* [37] backbone on DCS and Adroit [20] (both dense reward), and on Meta-World (limited training steps with *0/1 sparse reward*), reporting *success rates* for Adroit and Meta-World. Notably, we use text-to-video generation *without finetuning* on DCS. As reported in the Table A, GenReward consistently outperforms baselines across all benchmarks and reward settings.

Table B. Performance with hallucination.

Model	Dense Reward	GenReward w/ hallucination	GenReward
Pick Place	454 ±30	574 ± 98	**796 ± 73**

Table C. Impact of individual components.

Model	GenReward w/o CLIP	GenReward w/SigLIP 2	GenReward
Walker Walk	663±16	448±122	**782 ± 110**

Table D. Performance of GenReward variants on DCS *Walker Walk*.

Model	GenReward w/o action	GenReward w/ real	GenReward
Walker Walk	435±249	**784±67**	782 ± 110

A.3. Results on Video Hallucinations

We do not address hallucinations. Instead, we train our method on generated videos exhibiting hallucinations (*i.e.*, object teleportation). Interestingly, as reported in Table B, the agent consistently benefits from the proposed reward mechanism.

A.4. Component Sensitivity

The results of 1) replacing CLIP with random goal image selection, 2) replacing DINOv3 with SigLIP 2 are shown in Table C. Overall, the original design achieves the best performance among these variants

A.5. Comparison of GenReward Variants

Table D shows that our method is competitive with the real-video variant. Moreover, when computing the FB reward, removing the action leads to decreased performance (see Table D).

A.6. VDM Adaption Details.

Finetuning VDM takes 7 days using 16 A100 GPUs. We randomly select 400 samples from finetuning data. The FVD score is 21.6 indicates high fidelity of VDM.

A.7. Computational Complexity

Table E shows that, when training for 1M steps, GenReward trains faster than all baselines, except raw dense reward.

B. Forward-Backward Network Details

The forward-backward objective originates from approximating the successor measure $M^{\pi_z}(s, a, s')$, which de-

Table E. Training time comparison between baselines on Meta-World *Pick Place*.

Model	Dense Reward	RoboCLIP	Diffusion Reward	GenReward
Training time	102 hours	119 hours	109 hours	106 hours

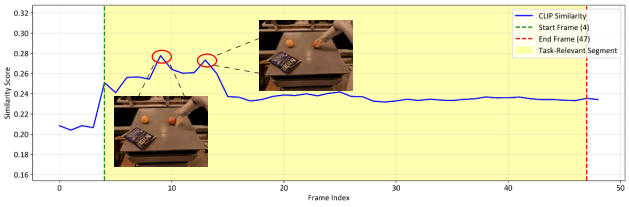

Figure B. Failure Case of CLIP-based frame selection in a generated RT-1 *Pick Apple* video. The most relevant frame does not fully grasp the apple, while the second-most relevant frame actually contains a successful grasp.

scribes the discounted occupancy of future states s' reachable from (s, a) under the policy π_z. In the low-rank factorization (see Eq. (7)), the inner product $F(s, a, z)^\top B(s')$ acts as a learned similarity measure: it is large if s' is likely to be visited from (s, a). Minimizing the Bellman residual on this approximation, as detailed in Eq. (8), therefore encourages future states that are likely to be visited from (s, a) under π_z to receive high similarity scores in the latent space. Additionally, the orthonormalization loss $\mathcal{L}_{\text{norm}}$ regularizes the backward representations to prevent degenerate collapse and ensure feature isotropy. Concretely, $\mathcal{L}_{\text{norm}} = \left\| \mathbb{E}_\rho[B\, B^\top] - I_d \right\|_{\text{F}}^2$. Here $\|\cdot\|_{\text{F}}$ is Frobenius norm. Detailed derivation of FB loss (see Eq. (8)) can be found in [30].

C. Effect of Frame-Level Goal Selection

As shown in Figure B, in some cases, CLIP may select a frame that is not the most relevant as the goal image. Interestingly, GenReward with RT-1 *Pick Apple* video still outperforms DreamerV3 with original reward (see Figure 9). Although not the most relevant, the frame-level goal selected by CLIP can still facilitate fine-grained goal achievement of the agent.

D. Hyperparameters

The final hyperparameters of GenReward are listed in Table F.

Table F. Hyperparameters of GenReward.

Name	**Notation**	**Value**
`Video-Level Reward`		
Reward weight	α	1×10^{-2}
`Forward-Backward Reward`		
Reward weight	β	1×10^{-5}
Train steps	—	1×10^{5}
Observation dimension	—	384
Feature dim	d	512
Hidden dim	—	512
Learning rate	—	1×10^{-4}
Target network soft-update rate	—	0.01
`General`		
Replay capacity	—	1×10^{6}
Batch size	B	16
Batch length	T	64
Train ratio	—	512
Intrinsic reward interval	—	128
`World Model`		
Deterministic latent dimensions	—	512
Stochastic latent dimensions	—	32
Discrete latent classes	—	32
RSSM number of units	—	512
World model learning rate	—	1×10^{-4}
Reconstruction loss scale	β_{pred}	1
Dynamics loss scale	β_{dyn}	0.5
Representation loss scale	β_{rep}	0.1
`Behavior Learning`		
Imagination horizon	H	15
Discount	γ	0.997
λ-target	λ	0.95
Actor learning rate	—	3×10^{-5}
Critic learning rate	—	3×10^{-5}