Pre-Trained Video Generative Models as World Simulators

Haoran He[1], **Yang Zhang**[2], **Liang Lin**[3], **Zhongwen Xu**[4], **Ling Pan**[1]*
[1]Hong Kong University of Science and Technology
[2]Tsinghua University
[3]Sun Yat-sen University
[4]Tencent AI Lab
haoran.he@connect.ust.hk

Abstract

Video generative models pre-trained on large-scale internet datasets have achieved remarkable success, excelling at producing realistic synthetic videos. However, they often generate clips based on static prompts (e.g., text or images), limiting their ability to model interactive and dynamic scenarios. In this paper, we propose **D**ynamic **W**orld **S**imulation (DWS), a novel approach to transform pre-trained video generative models into controllable world simulators capable of executing specified action trajectories. To achieve precise alignment between conditioned actions and generated visual changes, we introduce a lightweight, universal action-conditioned module that seamlessly integrates into any existing model. Instead of focusing on complex visual details, we demonstrate that consistent dynamic transition modeling is the key to building powerful world simulators. Building upon this insight, we further introduce a motion-reinforced loss that enhances action controllability by compelling the model to capture dynamic changes more effectively. Experiments demonstrate that DWS can be versatilely applied to both diffusion and autoregressive transformer models, achieving significant improvements in generating action-controllable, dynamically consistent videos across games and robotics domains. Moreover, to facilitate the applications of the learned world simulator in downstream tasks such as model-based reinforcement learning, we propose prioritized imagination to improve sample efficiency, demonstrating competitive performance compared with state-of-the-art methods.

Introduction

The field of video generation has experienced remarkable progress in recent years, with models such as Brooks et al. (2024); Zheng et al. (2024); Wan (2025); Yang et al. (2024c) demonstrating an exceptional ability to generate high-fidelity and temporally consistent videos conditioned on various inputs, most notably text and initial frames. However, these models are limited to support interactive simulation scenarios, as they are trained for one-shot generation with static prompts, lacking frame-level interactivity and frame-to-frame dynamic modeling. To fill this gap, the community is increasingly focusing on building action-conditioned video models (Yang et al. 2023; Bruce, Dennis, and et. al. 2024; Xiang et al. 2024; Wu et al. 2024; Valevski et al. 2024; Decart et al. 2024; Che et al. 2025; Yang et al. 2024a).

These action-conditioned models effectively act as interactive environment simulators ("world models" or "world simulators"), which leverage advanced transformers or diffusion model architectures to predict future visual outcomes based on the agent's actions. Their goal is to encapsulate an understanding of the underlying dynamic transitions and commonsense knowledge about how the world works, enabling action-driven imagination analogous to the human cognition process. These world models open exciting possibilities, particularly in model-based reinforcement learning (MBRL), where agents can learn new skills more efficiently by interacting with world models, avoiding the risks and costs that arise from real-world trials.

In this work, we review recent advances in interactive world simulators, highlighting key challenges that currently limit their broader adoption. (i) These models often require vast computational resources for training from scratch. For example, Genie (Bruce, Dennis, and et. al. 2024) required 125k training steps on 256 TPUv5p cores (roughly equivalent to 226 NVIDIA A100 GPUs) to learn a relatively simple Platformer game simulator. Similarly, GameNGen (Valevski et al. 2024) consumed 700k steps with 128 TPU-v5e cores. (ii) While fine-tuning a pre-trained video generative model offers a more efficient alternative, existing approaches (Rigter et al. 2024; Yu et al. 2025) are inherently architecture-dependent. This poses challenges to adapting these methods across different model architectures to benefit from rapid advances in video generation architectures, as each new model architecture requires substantial engineering efforts for adaptation. (iii) Unlike general video generation tasks, action-conditioned world simulators require precise capture of fine-grained dynamic changes (Zhu et al. 2024; Yang et al. 2023), which requires frame-level action alignment. This requirement is crucial for applications in model-based reinforcement learning, where capturing frame-to-frame dynamic/motion changes takes precedence over modeling static visual elements (e.g., background, object details).

To address the aforementioned challenges, we propose a novel framework, **D**ynamic **W**orld **S**imulation (DWS), which is a unified, architecture-agnostic approach for efficiently converting pre-trained video generative models into world simulators. By leveraging pre-trained priors learned from internet-

*Corresponding author

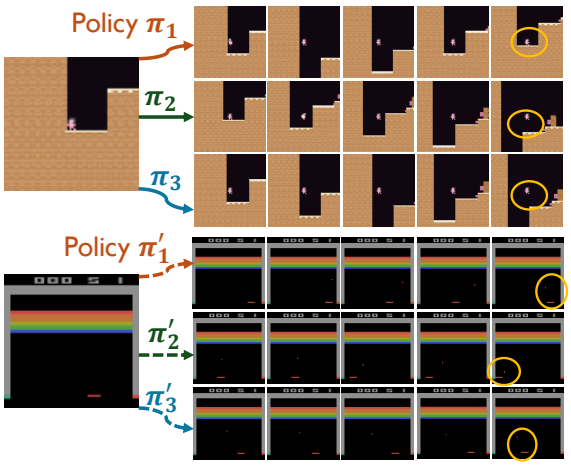


Figure 1: Our fine-tuned video model serves as an effective world simulator for interacting with different policies and generating diverse trajectories. As highlighted with yellow circles, different policies lead to different terminal states.

scale datasets, the fine-tuned world simulators can demonstrate a basic understanding of physical rules and commonsense knowledge. However, they are not inherently equipped for interactive simulation and lack the key mechanisms for precise frame-level action conditioning. DWS introduces a minimalist yet powerful add-on action-conditioned module that improves frame-level action awareness while maintaining architectural flexibility. This module comprises just two linear layers and can be integrated into any network architecture through carefully designed scale and shift operations. It also strengthens the alignment between predicted visual changes and conditioned actions. Furthermore, we observe that traditional supervised learning loss functions lead video models to focus uniformly across all visual elements and complex details, including static backgrounds and irrelevant details, compromising their ability to capture the frame-to-frame dynamic/motion changes crucial for world simulator construction. DWS presents a motion-reinforced loss to address the dynamic modeling challenge, a simple yet effective method to explicitly prioritize the modeling of inter-frame changes during training, resulting in significantly improved temporal consistency and more reliable dynamic predictions. As shown in Fig. 1, the learned world simulators by DWS can interact with diverse policies across different domains while maintaining accurate dynamic prediction and action responsiveness. Finally, to enhance the practical utility of world simulators in model-based reinforcement learning, we introduce prioritized imagination, a novel sampling strategy that focuses on the most informative transitions rather than wasting computational resources on well-understood state transitions, leading to improved sample efficiency during agent-world model interactions.

We summarize our contributions as follows: (i) We introduce DWS, a novel and architecture-agnostic framework that effectively converts pre-trained video generative models to world simulators with low training costs, leveraging the pre-trained prior knowledge for physics grounding. (ii) We introduce two simple yet effective techniques: a lightweight action-conditioned module that enables precise frame-level control and improve action-following ability, and a motion-reinforced training that redirects model attention from static visual details to action-induced dynamic changes for improving temporal, dynamic consistency. (iii) We advance the practical utility of world simulators in model-based reinforcement learning through prioritized imagination, which improves sample efficiency and policy performance when applying the trained world simulators to downstream model-based RL. (iv) Through comprehensive evaluation across challenging game and robotics tasks, we demonstrate that DWS significantly improves the quality and dynamic consistency of generated action-conditioned videos. DWS-trained world simulators with prioritized imagination enable more efficient and effective learning compared to previous SOTA MBRL approaches (Hafner et al. 2023; Alonso et al. 2024).

Related Work

Video World Models. With the development of internet-scale datasets (Bain et al. 2021; Chen et al. 2024) and advanced model architecture (Peebles and Xie 2023; Brooks et al. 2024), significant progress has been made in realistic video generation conditioned on text descriptions and initial frames (Blattmann et al. 2023; Lin et al. 2024; Ma et al. 2024; Yang et al. 2024c; Zheng et al. 2024). Building upon these foundations, current research has increasingly focused on action-controllable video generation, aiming to develop generalist world simulators (Xiang et al. 2024; Yang et al. 2023; Bruce, Dennis, and et. al. 2024; Feng et al. 2024; Valevski et al. 2024; Parker-Holder, Ball, and et. al. 2024; Zhu et al. 2024; Che et al. 2025; Gao et al. 2025) that can effectively model both physical dynamics and action consequences. However, these models typically require training from scratch on large-scale datasets and involve millions (or billions) of parameters, resulting in substantial computational overhead and slow inference speed. In contrast, we propose to adapt publicly available pre-trained video generative models (Zheng et al. 2024; Wu et al. 2024) into action-driven world simulators. Concurrent works (Rigter et al. 2024; Yu et al. 2025) have also explored leveraging pre-trained models for action-conditioned video generation. However, their investigations are limited to diffusion-based models, and neither work validates the effectiveness of their approaches in facilitating downstream tasks such as model-based RL.

Model-Based RL Model-based RL aims to build world models in which the trial-and-error can take place without real cost. With a sufficiently accurate world model, agents can develop imagination abilities, allowing them to simulate interactions and generate synthetic experience data (Ha and Schmidhuber 2018). This simulated data can then be leveraged to learn optimal policies for diverse decision-making tasks, effectively reducing the need for real-world interactions. Built upon Recurrent State Space Models (RSSM) (Hafner et al. 2019), the Dreamer series has demonstrated impressive performance across diverse domains, including Atari games (Machado et al. 2018), DeepMind Control Suite (Tassa et al. 2018), and Minecraft. To address the

limitation of RNNs in expressing complex patterns, recent works have explored leveraging transformer models for enhanced sequence modeling and long-term dependency capture (Micheli, Alonso, and Fleuret 2023a; Robine et al. 2023; Zhang et al. 2023, 2024), and incorporating diffusion models to better represent multi-modal distributions in dynamic learning (Ding et al. 2024; Alonso et al. 2024). However, although these works also employ transformer or diffusion models for world model learning, they predominantly rely on training from scratch and fail to leverage pre-trained knowledge for enhanced dynamics understanding, making them overly task-specific and limiting their ability to generalize across diverse tasks. Furthermore, while existing methods treat all imagined samples with uniform importance during training, our proposed DWS introduces a novel prioritization mechanism that selectively focuses on significant samples, thereby improving sample efficiency.

Preliminaries

Problem Formulation

The conditional video generation framework can be adaptable to instantiate a world simulator (or a world model) (Yang et al. 2023). The world model takes in some action as input and produces the visual consequence of the action as output, which aims to simulate the environment. This environment can be represented as a Partially Observable Markov Decision Process (POMDP), encapsulated within the tuple $(\mathcal{S}, \mathcal{O}, \phi, \mathcal{A}, p, r, \gamma)$. Here, $\mathcal{S}$ is the state space, and $\mathcal{O}$ is the observation space which only provides incomplete information of $\mathcal{S}$. At each timestep t, the agent chooses an action a_t by following a policy $\pi : \mathcal{O} \to \Delta_{\mathcal{A}}$, the environment updates the state following the dynamics, $s_{t+1} \sim p(s_{t+1}|s_t, a + t)$, the next observation $o_{t+1} = \phi(s_{t+1})$ is received and a scalar reward r_t is computed as $R(s_t, a_t, s_{t+1})$. The goal of the agent is to learn a policy $\pi^* = \arg\max_\pi \mathbb{E}_{a_t \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$ by maximizing the γ-discounted cumulative rewards.

A well-trained world model can replace the environment to interact with the agent, and thus benefit downstream policy learning by providing infinite experiences. Concretely, given a history observation o_{T_0}, at each timestep $t = T_0, \cdots, T-1$, the agent takes an action a_t based on its policy and previous imagined observations, and then the world model predicts the transition $p(o_{t+1}, r_{t+1}|o_t, a_t)$ to feedback the agent.

Pre-Trained Video Generative Models

By formulating learning world models for visual control as an interactive video generation problem, we can harness the widely available video data, which embeds broad knowledge that is generalizable across different domains (Yang et al. 2024b). Video data not only contains semantic visual details but also includes motion movements that capture the dynamic rules in the physical world. However, training such video world models on internet-scale video datasets from scratch is expensive and time-consuming. We propose to fine-tune pre-trained advanced video models to enable them to simulate interactions. Specifically, we adopt three different pre-trained video generative models as our base models, which are the diffusion models with different architecture designs, i.e., Open-Sora (Zheng et al. 2024) and Wan2.1 (Wan 2025), and the autoregressive transformer model, i.e., iVideoGPT (Wu et al. 2024). Open-Sora and Wan2.1 are rectified flow-based models that are fully open-sourced and pre-trained on millions of internet videos. We consider using them because they require only a few sampling steps, benefiting from flow-matching training. A recent work named iVideoGPT is an autoregressive transformer built upon LLaMA (Touvron et al. 2023) architecture. It compresses training data from different modalities (e.g., including visual outcomes, actions, and rewards) into a sequence of tokens for interactive video prediction.

Method

In this section, we first introduce our proposed action-conditioned module, and the motion-reinforced loss for enhancing dynamic modeling. After presenting methods for fine-tuning general video generative models into world simulators, we introduce the prioritized imagination technique for improving model-based reinforcement learning performance.

Action-Conditioned Module

Recent video generative models have achieved significant success in generating realistic videos correlated with conditioned text prompts or initial frames. These text-to-video tasks operate with static prompts that globally describe the entire video without specifying what the next frames should be. This design paradigm, while effective for general video generation, presents fundamental challenges when adapting them as world models that aim to simulate action-rich interactions, where the conditions are frame-level, fine-grained action trajectories. To address this requirement and ensure each generated frame matches its corresponding action in the trajectory, we leverage an action-conditioned module that conditions the generation of each frame by its corresponding action individually. Unlike previous text-to-video models that compress the entire action trajectory into a single embedding, our approach, similar to IRASim (Zhu et al. 2024), implements a more granular action encoding mechanism. We introduce a lightweight add-on module, consisting of two linear layers within each transformer block, to encode individual actions separately. This design ensures that each frame's content is directly modulated by its corresponding action, rather than being guided by a global description, and leads to a direct correspondence between actions and generated frames.

Action Representation. A key challenge in adapting video generative models for action-based control lies in the representation of actions. In discretized action spaces, actions are typically represented as integer values, which lack the rich semantic context present in text prompts used in traditional text-to-video models. This semantic gap can limit the model's ability to interpret and respond to different actions effectively. To bridge this gap, we propose to represent the actions using language templates that depict the meanings of the actions. Specifically, given an action trajectory $y = \{a_t, a_{t+1}, \cdots, a_{t+H-1}\}$, where H is the horizon of the trajectory, we develop a mapping function ψ to translate abstract action integers into meaningful languages, i.e.,

$\psi : \mathcal{A} \to L$, where L is the language space. This mapping enables us to leverage the text encoder in pre-trained video generative models to obtain rich feature embeddings $c \in \mathbb{R}^{n_H \times n_d}$, where n_H and n_d represent the horizon and the dimension of each token respectively. For continuous action spaces, following Wu et al. (2024); Zhu et al. (2024), we use a trainable linear action embedder to directly generate feature embeddings c without language translation.

Frame-Level Condition. In the context of video generative models serving as world simulators, precise temporal control is important as each action should directly modulate the visual content of its subsequent frame. To explicitly model and enhance this action-frame correspondence, we incorporate a frame-level action-conditioning module within each transformer block, drawing inspiration from IRASim (Zhu et al. 2024). While IRASim's implementation was limited to specific diffusion models with temporal-spatial transformer architectures, we significantly extend this concept by developing a versatile add-on module that generalizes across different model architectures, including both diffusion-based and transformer-based frameworks. Therefore, our design offers enhanced architectural flexibility and broader applicability. Our minimalist architecture, implemented with just two linear layers, enables lightweight integration and efficient fine-tuning with minimal computational overhead. Specifically, for each video embedding $x \in \mathbb{R}^{T \times C \times H \times W}$, we process them as follows before feeding them into the transformer block:

$$x^i = x^i + \mathrm{FFN}(\mathrm{LayerNorm}(x^i) \times (1 + \alpha^i) + \beta^i), \quad (1)$$

where α^i and β^i denote the scale and shift parameters for the i-th frame. They are regressed from the action embedding c^i. We illustrate our proposed module in Fig. 2, which can be seamlessly integrated into any network block (e.g., attention block).

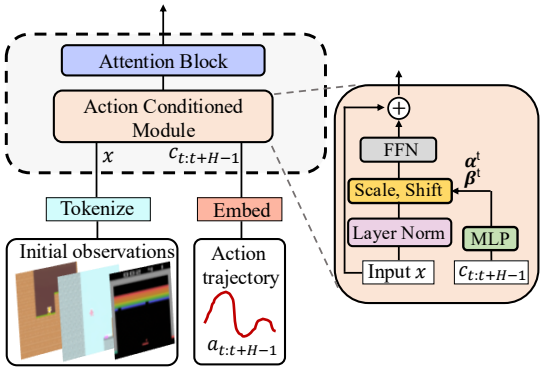


Figure 2: Illustration of the action-conditioned module, which can be incorporated with any type of transformer attention block.

Motion-Reinforced Loss

Traditional video generative models commonly employ the squared l_2 distance (for rectified flow (Liu, Gong, and qiang liu 2023; Zheng et al. 2024) in the continuous space) or cross-entropy loss (for next-token-prediction transformers (Vaswani et al. 2017; Wu et al. 2024) in discrete space) as their training objectives. While these training objectives have demonstrated effectiveness in general video generation tasks (Zheng et al. 2024; Lin et al. 2024; Brooks et al. 2024; Yan et al. 2021; Tian et al. 2024), they typically consider each pixel equally, which may compromise the model's ability to capture action-dependent state changes. Therefore, it is inefficient for them to function effectively as world simulators for RL agents, where accurate modeling of dynamic transitions is more crucial for learning than maintaining high-fidelity background details. This limitation arises because RL agents predominantly learn from action-induced state changes rather than static visual elements.

To tackle this problem and enable more precise dynamic modeling, we introduce a new motion-reinforced loss to improve the action-following ability of video models. At each training step, we sample a random batch of ground-truth video embeddings $x = \{x^0, \cdots, x^k, \cdots, x^j, x^{H-1}\}$, where H denotes the horizon of the video clips. We then compute the differences $\omega = \cup_{i=0}^{H-1} \omega_i$ between consecutive frames, denoted as

$$\omega_{i+1} = c^{\mathrm{Softmax}(|x_{i+1} - x_i|)}, \quad (2)$$

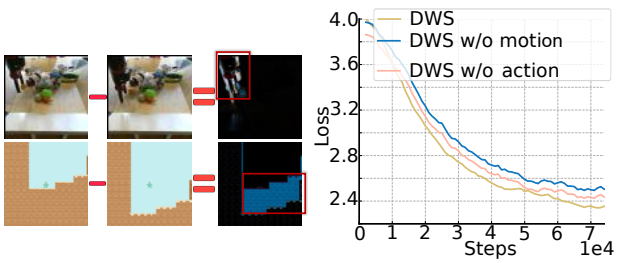
where $\omega_i \to [1, c]$, and we set $\omega_0 = 1$ for the initial frame x^0 since it serves as a conditioned frame. Here, c denotes a hyperparameter that modulates the motion-reinforced strength. After obtaining ω, we integrate it as pixel-wise weights into the supervised training loss. The resulting motion-reinforced loss function can be formulated as:

$$\mathcal{L}_{\mathrm{motion}} = \mathcal{L}_{\mathrm{prev}} * \omega, \quad (3)$$

where $\mathcal{L}_{\mathrm{prev}}$ represents either the original MSE loss used in diffusion models, or the cross-entropy loss function in transformer-based architectures. We include more implementation details in Appendix in He et al. (2025). Through this formulation, pixels that change across frames will have a greater impact on loss backpropagation. These pixels typically correspond to motion-related elements in videos, which undergo continuous changes, while the background, which remains stable, will have less influence during training. This mechanism inherently attenuates the impact of static background elements that contribute minimally to action-conditioned prediction.

By emphasizing dynamic transitions, our approach enhances the world model's capability to capture action-state causal relationships, thereby facilitating more effective policy learning in reinforcement learning contexts.

As illustrated in Figure 3(a), inter-frame differences predominantly correspond to motion-related and dynamic elements in videos, leading ω to assign higher weights to these pixels during the training process. Figure 3(b) presents the SFT loss ($\mathcal{L}_{\mathrm{prev}}$) curves from fine-tuning iVideoGPT (Wu et al. 2024) on the BAIR dataset (Ebert et al. 2017), comparing different variants of our proposed method. The empirical results demonstrate that both the motion-reinforced loss and the action-conditioned module are crucial components, as the absence of either component significantly degrades the model's performance in predicting action-conditioned videos with frame-to-frame dynamics.



(a) frame-to-frame differences (b) SFT loss $\mathcal{L}_{\text{prev}}$ comparison

Figure 3: Motion-reinforced loss enhances the action-controllability of video generative models, helping capture dynamically changing contents.

Therefore, the video generative models fine-tuned by $\mathcal{L}_{\text{motion}}$ will focus more on the dynamic/motion prediction instead of complex visual details that are challenging to learn. Moreover, dynamic/motion consistency is more important than static background details for world simulators, as required to predict action-induced visual changes.

Model-Based Reinforcement Learning

Given the video-based world simulators fine-tuned from pre-trained video generative models, one of the most promising applications is to utilize them as world models for policy learning in model-based reinforcement learning (MBRL). To enable effective policy training in MBRL, the world model should predict both transition dynamics and rewards. We now complete our world model with a reward prediction model. Since estimating the reward is a scalar prediction problem, we introduce a separate model R_ψ consisting of linear layers, self-attention blocks, and cross-attention blocks to estimate the reward given past observations and actions. The RL agent involves an actor-critic network parameterized by a shared CNN backbone that branches into separate policy and value heads. Building upon the MBPO framework (Janner et al. 2019; Wu et al. 2024), we augment the replay buffer with synthetic rollouts to train a standard actor-critic RL algorithm. We adopt PPO (Schulman et al. 2017) as our base algorithm.

Prioritized Imagination. Imagination by a world model needs to start from initial observations, which are sampled from the experiences collected from the environment. These initial states serve as starting points for the world model to generate synthetic trajectories. Previous MBRL methods (Hafner et al. 2020a,b, 2023; Alonso et al. 2024; Micheli, Alonso, and Fleuret 2023b) employ uniform sampling of initial observations for imagination. However, this strategy neglects the varying importance of different states for policy learning, leading to learning inefficiency. We highlight that imagined transitions originating from different initial observations exhibit substantial heterogeneity in their importance and task relevance for MBRL policy optimization. To better unlock the world simulation ability of fine-tuned video generative models, we propose a prioritized imagination method that selectively focuses on more valuable transitions. Our key insight is that initial observations leading to transitions with higher learning potential and learn-ability should be sampled more frequently. We maintain a buffer $\mathcal{B}$ to store observations encountered during the interaction with environments, and prioritize initial observations with high expected learning progress, which is measured by the magnitude of their TD loss. This prioritization mechanism ensures more efficient utilization of the world model by concentrating imagination resources on states that yield more substantial contributions to policy learning.

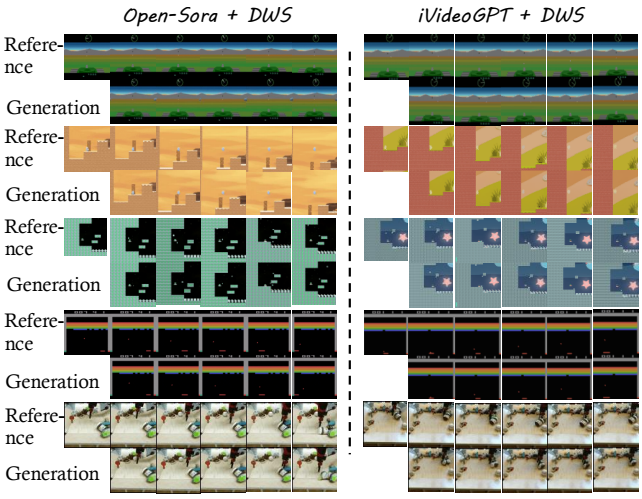


Figure 4: The qualitative results of generated videos for different domains, including games and robotics environments. Given initial observations and conditioned actions, we observe that Open-Sora and iVideoGPT fine-tuned by our proposed method significantly improve dynamic consistency.

Experiments

In this section, we evaluate the video world model fine-tuned by our proposed method from two critical perspectives: (1) action simulation capability, assessed through the quality of action-conditioned video prediction, and (2) the effectiveness in both online and offline model-based reinforcement learning, quantified by the cumulative return across tasks.

Action-Conditioned Simulation

To evaluate the effectiveness of DWS in enhancing action-conditioned video prediction for world simulation, we conduct experiments using three architecturally distinct pre-trained video generative models: Open-Sora (Zheng et al. 2024), Wan2.1 (Wan 2025), and autoregressive transformer-based iVideoGPT (Wu et al. 2024). The details and our setup of these base models can be found in Appendix in He et al. (2025). We leverage the well-established BAIR dataset (Ebert et al. 2017) featuring continuous action spaces to evaluate all three models. Additionally, we evaluate Open-Sora and ivideoGPT on Procgen (Cobbe et al. 2020) and Atari (Bellemare et al. 2013; Machado et al. 2018) datasets incorporating discrete action spaces.

Experiment Setup. The BAIR robot pushing dataset that is about a robotic arm manipulating various objects con-

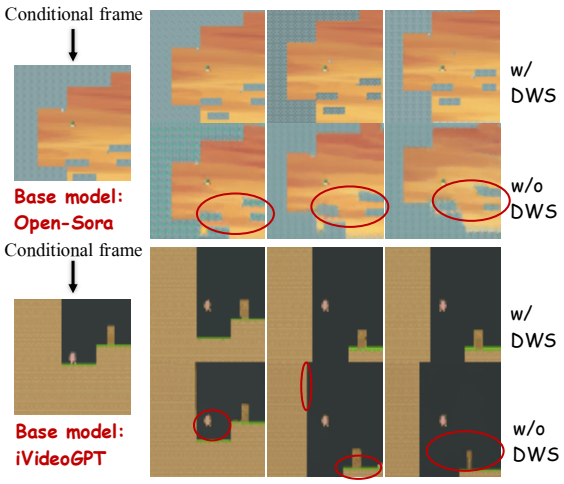


Figure 5: Quantitative comparison between pre-trained video generative models and fine-tuned models by our proposed methods. Base models without DWS can generate inconsistent pixels and fail to match the conditioned actions, as highlighted in red circles.

sists of 43k training and 256 test videos, Following previous works (Yan et al. 2021; Gupta et al. 2022), we predict 15 frames from a single initial frame. For the Procgen dataset, we evaluate DWS on two platformer games, i.e., namely Coinrun and Ninja. For the Atari dataset, we assess performance on two Atari games: Breakout and Battle Zone.

Metrics. We adopt four widely-used metrics to measure the quality of predicted videos across four metrics: 1): FVD (Unterthiner et al. 2018) quantifies the statistical similarity between reference and generated video distributions by computing the Fréchet distance between feature representations extracted from a pre-trained network (Carreira and Zisserman 2017); 2) PSNR (Huynh-Thu and Ghanbari 2008) measures the pixel-wise fidelity between reference and generated frames; 3) SSIM (Huynh-Thu and Ghanbari 2008) evaluates the structural information preservation in generated frames; 4) LPIPS (Zhang et al. 2018) leverages a pre-trained network (Simonyan and Zisserman 2014) to assess image similarity by computing distances in deep feature space.

Qualitative Results Analysis. We qualitatively evaluate two different models, i.e., Open-Sora and iVideoGPT, fine-tuned by DWS using unseen (o.o.d.) initial frames. Fig. 4 showcases examples of video generations across diverse domains given unseen inputs. We observe that both base models fine-tuned by DWS successfully generate high-quality, controllable videos characterized by coherent temporal dynamics and consistent background preservation. Furthermore, the models demonstrate robust generalization capabilities when processing unseen inputs with varying backgrounds and textures, validating the effectiveness of our approach. As evidenced in Figure 5, while the base models tend to generate videos with visual distortions, DWS significantly enhances

BAIR (Ebert et al. 2017)	FVD↓	PSNR↑	SSIM↑	LPIPS↓
VideoGPT (Yan et al. 2021)	103.3	-	-	-
MaskViT (Gupta et al. 2022)	93.7	-	-	-
FitVid (Babaeizadeh et al. 2021)	93.6	-	-	-
MaskViT (Gupta et al. 2022)	70.5	-	-	-
MCVD (Voleti et al. 2022)	89.5	16.9	78.0	-
MAGVIT (Yu et al. 2023)	62.0	19.3	78.7	12.3
Open-Sora (Zheng et al. 2024)	92.1	21.5	84.7	8.6
Open-Sora+**DWS**(Ours)	**81.3**	**22.4**	**87.8**	**6.2**
Wan2.1 (Wan 2025)	29.6	30.2	91.8	6.1
Wan2.1+**DWS**(Ours)	**24.1**	**31.1**	**92.3**	**5.9**
iVideoGPT (Wu et al. 2024)	60.8	24.5	90.2	5.0
iVideoGPT+**DWS** (Ours)	**59.6**	**25.8**	**91.6**	**4.7**

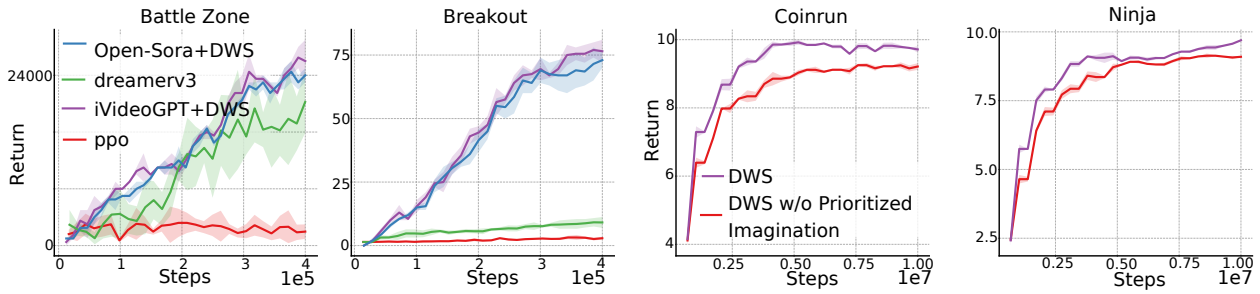
Table 1: Video prediction results on the BAIR robot pushing dataset. LPIPS and SSIM scores are scaled by 100 for convenient display.

the output quality by maintaining object consistency and producing precise visual predictions.

Quantitative Results Analysis. We provide the results on the BAIR dataset in Table 1, and refer to the results on the Atari and Procgen game datasets in the Appendix in He et al. (2025). We have the following key observations: (i) DWS demonstrates superior performance in action-conditioned video prediction across different base models. On the BAIR dataset with a continuous action space, DWS significantly enhances the performance of both the diffusion-based Open-Sora, flow-based Wan2.1, and the autoregressive-based iVideoGPT, highlighting its generalizability and versatility across different architectures. The results show that Open-Sora, a traditional text-to-video model that conditions generation on static, global text prompts, can gain significant improvements in action controllability after DWS fine-tuning. Specifically, we observe an 11.7% reduction in FVD and a 27.9% decrease in LPIPS. Even with iVideoGPT, which is specifically designed for action-conditioned video generation, DWS achieves notable performance improvements. Regarding both Procgen and Atari game datasets with discrete action spaces, DWS consistently improves the base models' performance by a distinct margin, yielding significant enhancements in generated video quality. These enhancements can be attributed to two key components: The action-conditioned module, which efficiently modulates each reaction to its corresponding frame, and the motion-reinforced loss function, which effectively captures frame-to-frame pixel dynamics. (ii) DWS demonstrates its versatility as a universal method that can be efficiently deployed across diverse datasets, ranging from robotics datasets with continuous action spaces to game datasets with discrete action spaces. Furthermore, DWS can be seamlessly integrated into various architectures, where the three popular architectures considered in this work are diffusion-based, flow-based, and autoregressive transformer-based models.

Model-Based Reinforcement Learning

To validate the practical utility of DWS-fine-tuned models for world simulators, we evaluate their performance by utilizing them as world models in MBRL policy learning. Considering



(a) Atari (b) Prioritized Imagination

Figure 6: (a): Averaged return across five random seeds on Atari environments. (b): Prioritized imagination improves the performance of model-based RL.

time efficiency, we take smaller Open-Sora and iVideoGPT as the base models in this task.

Benchmarks and Baselines. We conduct experiments on coinrun and ninja platformer games from the Procgen benchmark, and Breakout and Battle Zone games from the Atari benchmark. We compare our MBRL method with prioritized imagination with the following baselines: 1) PPO (Schulman et al. 2017) is a model-free RL method that is widely used. Our method is built on PPO. 2) Dreamerv3 (Hafner et al. 2023) is a SOTA model-based RL method that employs a recurrent network for dynamic prediction and actor-critic RL for policy learning, which is effective in handling tasks with discrete action spaces. For Procgen task, we additionally include PPG (Cobbe et al. 2021) for comparison, as it represents a competitive algorithm on Procgen.

Results Analysis. The experimental results presented in Fig. 6(a) demonstrate that our DWS-trained world model, when combined with a simple PPO algorithm, significantly outperforms both vanilla PPO and state-of-the-art model-based reinforcement learning methods, i.e., Dreamerv3. In Procgen environments, DWS exhibits substantial performance improvements over model-free approaches such as PPO and PPG. Although the DWS-trained world model requires fine-tuning during MBRL policy training—due to the episodic variations in background and object details inherent to Procgen, it maintains competitive performance compared to existing model-based methods. In Atari environments, DWS demonstrates substantial performance improvements over existing methods, attributed to its world models having acquired comprehensive dynamics knowledge for action simulation. Specifically, in Breakout, DWS achieves a remarkable 7× performance gain compared to SOTA methods. This superior performance, particularly in terms of sample efficiency, can be attributed to our proposed prioritized imagination technique. We validate this contribution through ablation studies conducted on Procgen environments, with results presented in Figure 6(b).

Offline Model-Based Reinforcement Learning

We further explore the potential of leveraging trained video world models to augment offline datasets for policy enhancement. Using CoinRun and Ninja environments as case studies, we first establish baseline datasets by collecting 1M expert trajectories for each environment using a well-trained PPO agent, following Mediratta et al. (2024). For each evaluated environment, we employ *Open-Sora+DWS* to synthesize an additional 1M state-action transitions during training, effectively doubling the size of the original datasets. To validate the effectiveness of this data augmentation approach, we evaluate the performance improvements using two different offline RL algorithms: Conservative Q-Learning (CQL) (Kumar et al. 2020) and Implicit Q-Learning (IQL) (Kostrikov, Nair, and Levine 2022). As shown in Table 2, augmenting offline RL with world model-generated data during training significantly enhances performance across different algorithms and environments. These results demonstrate that the DWS-trained world simulator can generate meaningful state-action-reward transitions that effectively supplement the offline dataset for policy learning.

Tasks	CQL	CQL w/ wm	IQL	IQL w/ wm
Coinrun	8.58 ± 0.29	$\mathbf{8.81 \pm 0.21}$(↑)	8.52 ± 0.26	$\mathbf{8.93 \pm 0.19}$(↑)
Ninja	5.92 ± 0.2	$\mathbf{6.31 \pm 0.23}$(↑)	5.7 ± 0.35	$\mathbf{6.33 \pm 0.17}$(↑)

Table 2: Average return across 3 seeds on Coinrun and Ninja tasks.

Conclusion and Limitation

In this paper, we present DWS, a novel approach that efficiently adapts pre-trained video generative models as world simulators by leveraging their rich prior knowledge learned from large-scale datasets for downstream action-conditioned simulation. Our framework introduces a lightweight action-conditioned module that enables action-frame alignment and can be seamlessly integrated into various model architectures, and a motion-reinforced loss specifically designed to model inter-frame pixel dynamics crucial for accurate world simulation. Extensive experimental results demonstrate that DWS significantly improves both video prediction quality and MBRL performance, leading to meaningful applications. However, DWS is currently limited in modeling videos with extended temporal horizons or high spatial resolutions. There also exists a trade-off between the inference latency and world modeling quality, which is significant for downstream applications. We leave it as future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China 62406266. We thank Yuanfang Peng and Jiwen Yu for the discussion, and the anonymous reviewers for their valuable feedback.

References

Alonso, E.; Jelley, A.; Micheli, V.; Kanervisto, A.; Storkey, A.; Pearce, T.; and Fleuret, F. 2024. Diffusion for World Modeling: Visual Details Matter in Atari. In *Thirty-eighth Conference on Neural Information Processing Systems*.

Babaeizadeh, M.; Saffar, M. T.; Nair, S.; Levine, S.; Finn, C.; and Erhan, D. 2021. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision*.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Bruce, J.; Dennis, M. D.; and et. al. 2024. Genie: Generative Interactive Environments. In *Forty-first International Conference on Machine Learning*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Che, H.; He, X.; Liu, Q.; Jin, C.; and Chen, H. 2025. GameGen-$\mathbb{X}$: Interactive Open-world Game Video Generation. In *The Thirteenth International Conference on Learning Representations*.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; and Tulyakov, S. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cobbe, K.; Hesse, C.; Hilton, J.; and Schulman, J. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, 2048–2056. PMLR.

Cobbe, K. W.; Hilton, J.; Klimov, O.; and Schulman, J. 2021. Phasic policy gradient. In *International Conference on Machine Learning*, 2020–2027. PMLR.

Decart; Quevedo, J.; McIntyre, Q.; Campbell, S.; Chen, X.; and Wachen, R. 2024. Oasis: A Universe in a Transformer.

Ding, Z.; Zhang, A.; Tian, Y.; and Zheng, Q. 2024. Diffusion world model. *arXiv preprint arXiv:2402.03570*.

Ebert, F.; Finn, C.; Lee, A. X.; and Levine, S. 2017. Self-Supervised Visual Planning with Temporal Skip Connections. *CoRL*, 12(16): 23.

Feng, R.; Zhang, H.; Yang, Z.; Xiao, J.; Shu, Z.; Liu, Z.; Zheng, A.; Huang, Y.; Liu, Y.; and Zhang, H. 2024. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*.

Gao, S.; Zhou, S.; Du, Y.; Zhang, J.; and Gan, C. 2025. AdaWorld: Learning Adaptable World Models with Latent Actions. In *International Conference on Machine Learning (ICML)*.

Gupta, A.; Tian, S.; Zhang, Y.; Wu, J.; Martín-Martín, R.; and Fei-Fei, L. 2022. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*.

Ha, D.; and Schmidhuber, J. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020a. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*.

Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J. 2019. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, 2555–2565. PMLR.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2020b. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

He, H.; Zhang, Y.; Lin, L.; Xu, Z.; and Pan, L. 2025. Pre-trained video generative models as world simulators. *arXiv preprint arXiv:2502.07825*.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44: 800–801.

Janner, M.; Fu, J.; Zhang, M.; and Levine, S. 2019. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32.

Kostrikov, I.; Nair, A.; and Levine, S. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191.

Lin, B.; Ge, Y.; Cheng, X.; Li, Z.; Zhu, B.; Wang, S.; He, X.; Ye, Y.; Yuan, S.; Chen, L.; et al. 2024. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*.

Liu, X.; Gong, C.; and qiang liu. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *The Eleventh International Conference on Learning Representations*.

Ma, X.; Wang, Y.; Jia, G.; Chen, X.; Liu, Z.; Li, Y.-F.; Chen, C.; and Qiao, Y. 2024. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*.

Machado, M. C.; Bellemare, M. G.; Talvitie, E.; Veness, J.; Hausknecht, M.; and Bowling, M. 2018. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61: 523–562.

Mediratta, I.; You, Q.; Jiang, M.; and Raileanu, R. 2024. The Generalization Gap in Offline Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Micheli, V.; Alonso, E.; and Fleuret, F. 2023a. Transformers are Sample-Efficient World Models. In *The Eleventh International Conference on Learning Representations*.

Micheli, V.; Alonso, E.; and Fleuret, F. 2023b. Transformers are Sample-Efficient World Models. In *The Eleventh International Conference on Learning Representations*.

Parker-Holder, J.; Ball, P.; and et. al. 2024. Genie 2: A large-scale foundation world model.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Rigter, M.; Gupta, T.; Hilmkil, A.; and Ma, C. 2024. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*.

Robine, J.; Höftmann, M.; Uelwer, T.; and Harmeling, S. 2023. Transformer-based World Models Are Happy With 100k Interactions. In *The Eleventh International Conference on Learning Representations*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casas, D. d. L.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; et al. 2018. Deepmind control suite. *arXiv preprint arXiv:1801.00690*.

Tian, K.; Jiang, Y.; Yuan, Z.; PENG, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Valevski, D.; Leviathan, Y.; Arar, M.; and Fruchter, S. 2024. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wan, T. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wu, J.; Yin, S.; Feng, N.; He, X.; Li, D.; Hao, J.; and Long, M. 2024. iVideoGPT: Interactive VideoGPTs are Scalable World Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiang, J.; Liu, G.; Gu, Y.; Gao, Q.; Ning, Y.; Zha, Y.; Feng, Z.; Tao, T.; Hao, S.; Shi, Y.; et al. 2024. Pandora: Towards General World Model with Natural Language Actions and Video States. *arXiv preprint arXiv:2406.09455*.

Yan, W.; Zhang, Y.; Abbeel, P.; and Srinivas, A. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

Yang, M.; Du, Y.; Ghasemipour, K.; Tompson, J.; Schuurmans, D.; and Abbeel, P. 2023. Learning Interactive Real-World Simulators. *arXiv preprint arXiv:2310.06114*.

Yang, M.; Li, J.; Fang, Z.; Chen, S.; Yu, Y.; Fu, Q.; Yang, W.; and Ye, D. 2024a. Playable Game Generation. *arXiv preprint arXiv:2412.00887*.

Yang, S.; Walker, J. C.; Parker-Holder, J.; Du, Y.; Bruce, J.; Barreto, A.; Abbeel, P.; and Schuurmans, D. 2024b. Position: Video as the New Language for Real-World Decision Making. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 56465–56484. PMLR.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024c. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yu, J.; Qin, Y.; Wang, X.; Wan, P.; Zhang, D.; and Liu, X. 2025. GameFactory: Creating New Games with Generative Interactive Videos. *International Conference on Computer Vision*.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10459–10469.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, W.; Wang, G.; Sun, J.; Yuan, Y.; and Huang, G. 2023. STORM: Efficient Stochastic Transformer based World Models for Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhang, Y.; Bai, C.; Zhao, B.; Yan, J.; Li, X.; and Li, X. 2024. Decentralized Transformers with Centralized Aggregation are Sample-Efficient Multi-Agent World Models. *arXiv preprint arXiv:2406.15836*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-Sora: Democratizing Efficient Video Production for All.

Zhu, F.; Wu, H.; Guo, S.; Liu, Y.; Cheang, C.; and Kong, T. 2024. IRASim: Learning Interactive Real-Robot Action Simulators. *arXiv:2406.12802*.