NIL: No-data Imitation Learning

Mert Albaba[1,2] Chenhao Li[1] Markos Diomataris[1,2] Omid Taheri[2]
Andreas Krause[1] Michael J. Black[2]

[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems

nil.is.tue.mpg.de

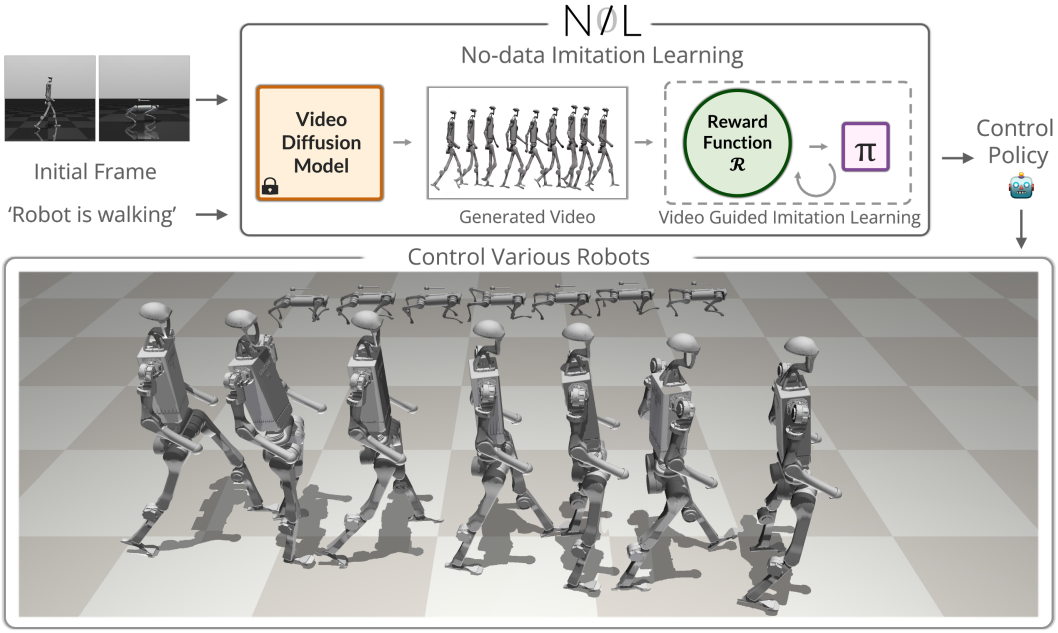


Figure 1. **NIL Overview:** First, from a single frame and a textual prompt, a pre-trained video diffusion model generates a reference video. Reinforcement learning policies are then trained to imitate the generated video and control various robots without using any curated data.

Abstract

Acquiring physically plausible motor skills across diverse and unconventional embodiments, including humanoids and quadrupeds, is essential for advancing character simulation and robotics. Traditional methods, such as reinforcement learning (RL), require extensive reward function engineering. Imitation learning (IL) offers an alternative but relies heavily on curated 3D expert demonstrations, which are scarce and difficult to obtain for non-human morphologies. Video diffusion models, on the other hand, are capable of generating realistic-looking videos of various morphologies, from humans to ants. However, these videos are often not physically plausible, which limits their direct use for skill acquisition. We introduce "No-data Imitation Learning" (NIL): an imitation learning framework that replaces curated expert demonstrations with videos generated by a pretrained video diffusion model. Our key insight is that the physics simulator enforces physical constraints, while the video provides visual guidance. NIL learns 3D motor skills in a physics simulator from 2D-generated videos, with generalization capability to unconventional forms. Specifically, NIL computes a discriminator-free imitation reward that combines (i) a video-embedding similarity between generated and simulated videos using a pretrained video vision transformer, and (ii) an image-based similarity term derived from video segmentation masks. We evaluate NIL on locomotion and whole-body control tasks across unique body configurations. Our experiments show that in humanoid locomotion, NIL matches the performance of state-of-the-art IL baselines trained on motion-capture data; and in whole-body manipulation, it exceeds the performance of RL baselines without requiring any curated data. Our project page, including videos and code, is available at https://nil.is.tue.mpg.de.

1. Introduction

Learning motor skills for multiple and diverse agent morphologies, including robots or animals, is essential to advance robotics and character simulation. However, enabling physically plausible skill acquisition across such a range of morphologies remains a longstanding challenge.

Reinforcement learning (RL) is the standard approach for training skills in a physically plausible manner. RL trains agents within a physical simulator so that the learned behaviors inherently respect physical laws, an important property for both robotics and character simulation. However, RL requires substantial manual effort to engineer reward functions for each specific task-body pair, and poorly specified rewards lead to unintended behavior [3]. Imitation learning (IL) circumvents meticulous reward engineering by learning from expert demonstrations, but it relies on high-quality 3D data with accurate joint positions and velocities. Such high-quality 3D data is difficult to obtain, particularly for non-humanoid robots and animals, where motion-capture data is scarce, expensive, or even impossible to collect.

This data-collection challenge motivates us to explore an alternative to curated 3D demonstrations. Recent advances in video diffusion models [12, 22, 49] offer a new path: generating reference visual demonstrations on demand. These pretrained generative models are capable of generating visually plausible videos across a wide variety of tasks and morphologies. Leveraging such models could eliminate the need for curated 3D data. However, while the generated videos are visually plausible, they are not always physically plausible [30], limiting their direct use for skill acquisition. Furthermore, learning 3D skills from 2D videos is challenging due to 3D ambiguity and the lack of precise action annotations.

To address these challenges, we introduce *No-data Imitation Learning (NIL)*: an imitation learning framework that replaces curated demonstrations with videos generated by a pretrained video diffusion model. NIL uses generated 2D videos purely for visual guidance, relying on a physics simulator to enforce 3D physical constraints. By coupling these components, NIL learns 3D motor skills by training a policy in physical simulation to mimic a single reference 2D video generated by a pretrained video diffusion model (Figure 1 provides an overview).

NIL treats generated videos as supervision by converting them into a dense, discriminator-free reward that combines (i) video-encoder similarity and (ii) segmentation-mask similarity, while the simulator enforces physical plausibility. Specifically, NIL employs video vision transformers [4, 13] to compute a similarity reward between the generated reference video and a video rendered from the agent's physically simulated trajectory. Both videos are embedded into the latent space of a video vision transformer, and the similarity between their encoding becomes a reward signal that encourages the agent to replicate the motion's key aspects. Such holistic video-level similarity is insufficient on its own, as it lacks precise frame-wise spatial guidance. As a fine-grained feedback, NIL also employs an image-based similarity: it segments the agent's body in both the generated video and the rendered simulation trajectory, creates binary masks, and computes frame-aligned Intersection over Union (IoU) as a complementary reward. The video-encoding reward provides temporal and semantic guidance, while the frame-wise mask reward provides spatial guidance, together they yield a stable imitation signal without a discriminator.

Overall, NIL combines pretrained video diffusion models, video vision transformers, and imitation learning to learn skills for diverse morphologies. By directly measuring similarities between video encodings and segmentation masks, NIL generates an effective reward signal that guides learning within a physics simulator, without relying on curated 3D motion-capture data. As our main technical contribution, we introduce the NIL framework, which comprises two key components:

1. **Synthetic Expert Data:** NIL generates expert demonstrations on-the-fly using video diffusion models, conditioned on the agent's initial state and a textual task description. This approach generalizes to any task-body pair by removing any dependency on curated 3D demonstrations.
2. **Video- and Mask-based Reward Signal:** NIL combines video vision transformers and image segmentation to create an informative reward signal from 2D videos for imitation learning. This creates a stable and effective learning signal for imitation learning.

We test our approach on locomotion and whole-body manipulation tasks involving diverse morphologies, including two-legged and four-legged robots, for which obtaining 3D high-quality data is difficult, and reward engineering is challenging. Our results show that NIL matches the performance of imitation learning methods that require curated 3D motion-capture data on locomotion, and outperforms reinforcement-learning baselines on whole-body control tasks, while requiring no curated data.

By leveraging the strengths of video diffusion models and imitation learning, our approach addresses critical bottlenecks in training agents for complex tasks, particularly when high-quality data is scarce or unavailable. This work opens new avenues for research at the intersection of generative modeling and reinforcement learning, with potential applications in robotics and animation. Our code and models will be released for research purposes.

2. Related Work

2.1. Imitation Learning

Imitation learning (IL) has long been an attractive paradigm for training agents to mimic expert behavior. Early approaches like Behavioral Cloning [5] directly map observations to actions but are prone to compounding errors under distributional shifts. Adversarial IL methods such as GAIL [20], InfoGAIL [28], AIRL [18], VDB [32], and AMP [33] address this by adversarially training a discriminator to tell apart trajectories from the expert dataset and trajectories generated by the current policy, and then using the discriminator's output as a learned reward signal, often with additional regularization or motion priors. RILe [2] combines inverse reinforcement learning with adversarial imitation learning to improve performance in higher-dimensional environments.

While adversarial approaches achieve strong performance, the discriminator tends to overfit quickly, leading to instability during training [2]. In contrast, our work bypasses the need for a discriminator by computing a dense reward signal directly from video comparisons. This discriminator-free objective sidesteps the adversarial training instabilities often observed in adversarial IL methods.

2.2. Video Diffusion Models

If we had access to limitless, high-quality, 3D training data, imitation learning would work well. Since capturing such data is challenging, is it possible to generate it? Denoising Diffusion Probabilistic Models (DDPMs) [21] enable high-quality image synthesis, and video diffusion models [22] extend these ideas to the temporal domain, generating coherent video sequences. Early text-to-video models such as Make-A-Video [37] and Imagen [36] demonstrate that cascaded diffusion processes capture complex motion patterns from large-scale datasets. Subsequent models like Stable Video Diffusion [12] and I2VGen [49] improve video quality and conditional controllability, and works such as DynamiCrafter [44] explore architectures tailored to dynamic scene synthesis. More recent approaches such as Lumiere [7], MagicVideo-V2 [42], and CogVideoX [46] further improve global coherence, aesthetic quality, and video length. Together with proprietary systems such as Kling and Pika, these advances provide increasingly visually plausible reference videos.

Despite their impressive visual performance, these models sometimes output 2D results that are visually convincing but physically implausible [30], posing a challenge to using them for imitation learning.

2.3. Video Encoders

To exploit generated 2D data as supervision, we require video encoders that provide meaningful spatio-temporal features bridging 2D observations and 3D behavioral understanding. Vision transformer architectures such as ViViT [4] and TimeSFormer [8] effectively capture dynamic patterns in video data. Masked autoencoding techniques [17] and transformer architectures such as VidTr [50] and DistInit [19] further validate the potential of transformer-based models for video understanding. Subsequent work, including VideoMAE-v2 [41] and VideoMamba [27], advances video representation learning via dual masking strategies and efficient state-space models. More recently, video foundation models such as InternVideo2 [43], VideoPrism [51], and VideoLLaMA3 [47] scale video encoders and achieve state-of-the-art performance on video understanding tasks.

These advances in video understanding provide robust representations that are essential for comparing generated and simulated behaviors in our framework.

2.4. Learning from Generated Data

There is a growing body of work that leverages generated or weakly supervised data to reduce reliance on curated expert demonstrations. Early explorations in one-shot imitation learning [16] and meta-imitation learning [26] show that agents can learn effectively from sparse or unstructured data. DexMV [34] and video-language planning methods [14] incorporate video data directly into the policy or planner. Complementary efforts in imitation learning from human videos [45] and zero-shot robotic manipulation using pretrained image-editing diffusion models [11] further highlight the potential of harnessing generated data. Methods such as MimicPlay [40] repurposes human video play data and UniPi [15] uses a video generator as a planner. More recent approaches such as Track2Act [10] use internet videos to generate motion plans that are mapped to robot actions, while Gen2Act [9] combines generated human videos with a small set of action-labeled robot trajectories to train a policy. RoboDreamer [52] uses generated videos to train compositional world models, and AVDC [25] combines a text-to-video diffusion model with pretrained flow networks to infer actions from videos. Dreamitate [29] fine-tunes a video diffusion model for robotic control.

Despite this progress, most prior methods either rely on at least some curated data for policy training, or treat generated videos as open-loop plans or world models rather than integrating them as a dense imitation reward signal inside a physics simulator. In contrast, NIL uses pretrained video diffusion models purely as a source of visual reference-demonstrations for discriminator-free imitation learning. To the best of our knowledge, NIL is the first framework to demonstrate that only generated videos are sufficient to learn physically plausible whole-body locomotion and loco-manipulation skills across diverse robot morphologies in simulation.

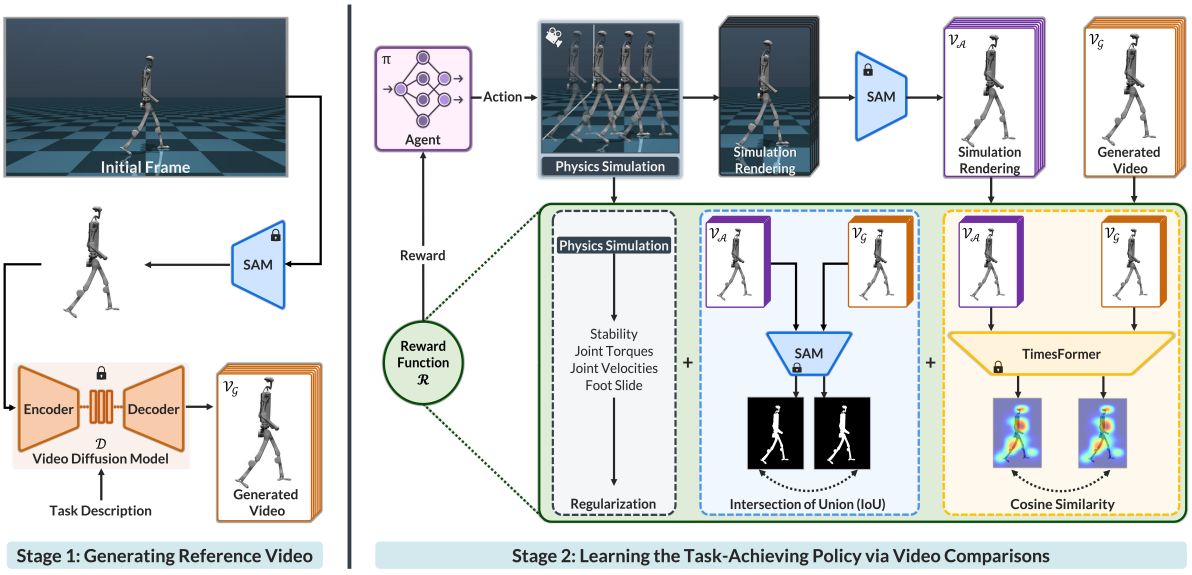


Figure 2. **NIL:** No-data Imitation Learning consists of two stages. *Stage 1 (Video generation):* Render the agent's initial frame, remove the background, and generate a reference video with a pretrained video diffusion model conditioned on the initial frame and a textual task description. *Stage 2 (Policy learning):* Train a reinforcement learning agent in a physics simulator to imitate the generated video via a reward comprising (1) video-encoder similarity, (2) segmentation-mask IoU, and (3) regularization for smooth behavior.

3. Method

3.1. Overview

No-data Imitation Learning (NIL) learns physically plausible 3D motor skills from 2D videos generated by a video diffusion model. Given a skill s_i and an embodiment b_j, the goal is to learn a policy π_{s_i,b_j} that enables a simulated agent e_{b_j} to perform s_i. NIL comprises two stages (Fig. 2):

1. **Video generation:** A reference video F_{s_i,b_j} is generated by a video diffusion model D, conditioned on the initial 2D simulation frame e_0 and a textual prompt p_{s_i,b_j}.
2. **Policy learning:** A dense imitation reward compares the generated video F_{s_i,b_j} with a rendered simulation video E_{s_i,b_j} and, together with smoothness regularization, guides the optimization of π_{s_i,b_j}.

3.2. Stage 1: Video Generation

The video generation module uses a frozen, pretrained video diffusion model, D, to generate a 2D video of the agent performing the skill s_i. The inputs are (i) the initial frame e_0 of embodiment b_j rendered from the physical simulation at a fixed starting position and (ii) a textual prompt (p_{s_i,b_j}) describing the task p_{s_i,b_j} = "The b_j agent is s_i, camera follows the agent." We use a fixed camera setup; the *camera follows the agent* in both the generated video and the renderings from the physical simulation. D outputs a color video

$$D(p_{s_i,b_j},\, e_0) = F_{s_i,b_j} \in \mathbb{R}^{n\times H\times W\times 3} \tag{1}$$

$$F_{s_i,b_j} = \{f_0^{(s_i,b_j)}, f_1^{(s_i,b_j)}, \ldots, f_{n-1}^{(s_i,b_j)}\}, \tag{2}$$

where n is the number of frames, and $H \times W$ is the spatial resolution.

3.3. Stage 2: Learning the Task-Achieving Policy

Video Similarity Pipeline: The video-similarity metric computes a reward signal by comparing the generated video F_{s_i,b_j} with the rendered simulation video $E_{s_i,b_j} = \{e_0^{(s_i,b_j)}, e_1^{(s_i,b_j)}, \ldots, e_{n-1}^{(s_i,b_j)}\}$, where n is the length of the rendered agent video. The objective is to extract meaningful learning signals from the 2D-generated video to guide the acquisition of 3D motor skills. The reward is computed in three steps: (a) segmentation and masking; (b) video encoding; and (c) similarity computation.

3.3.1. Segmentation and Masking

We segment the agent in both F_{s_i,b_j} and E_{s_i,b_j}. For F_{s_i,b_j}, Segment Anything Model 2 (SAM) [35] yields masks $M^F = \{M_0^F, \ldots, M_{n-1}^F\}$, prompted with initial frame mask which is provided by the simulator; for E_{s_i,b_j}, the simulator provides masks as $M^E = \{M_0^E, \ldots, M_{n-1}^E\}$. Masked frames are denoted as $f_t^{M,(s_i,b_j)}$, $e_t^{M,(s_i,b_j)}$. The segmented videos are thus represented as F_{s_i,b_j}^M and E_{s_i,b_j}^M.

3.3.2. Video Encoding

To capture spatiotemporal dynamics for both the generated and rendered videos, we employ a pretrained video encoder T. For each time step t, we construct n_T-frame clips from masked videos as:

$$C_t^{(s_i,b_j)} = \{f_{t-n_T+1}^{M,(s_i,b_j)}, \ldots, f_t^{M,(s_i,b_j)}\} \tag{3}$$

where n_T is the number of frames that the video encoder expects. We left-pad the videos with initial frames when $t < n_T$ to ensure $|C_t| = n_T$. Each clip is passed through T to obtain the embedding

$$z_t^{F,(s_i,b_j)} = T\Big(C_t^{(s_i,b_j)}\Big). \tag{4}$$

An analogous procedure yields $z_t^{E,(s_i,b_j)}$ for $E^M_{s_i,b_j}$. Embeddings are derived from the last hidden states of the video-encoder, yielding a compact motion representation.

3.3.3. Reward Function

At each frame t, we combine a video-embedding similarity, a mask-based image similarity, and regularization as the reward.

a) Video Similarity: The video similarity at time step t is defined as the cosine similarity between the corresponding embeddings of the generated and rendered videos:

$$S_{v,t} = \frac{z_t^{F,(s_i,b_j)} \cdot z_t^{E,(s_i,b_j)}}{||z_t^{F,(s_i,b_j)}|| \cdot ||z_t^{E,(s_i,b_j)}||}. \tag{5}$$

The cosine similarity ranges between -1 and 1, with higher values indicating greater similarity between the encodings.

b) Image-Based Similarity: We compute the Intersection over Union (IoU) between the binary masks $M_t^F, M_t^E \in \{0,1\}^{H\times W}$ of the generated and rendered videos as:

$$S_{M,t} = \frac{\sum_{k,l} M_t^F(k,l) \cdot M_t^E(k,l)}{\sum_{k,l} M_t^F(k,l) + M_t^E(k,l) - M_t^F(k,l) \cdot M_t^E(k,l)} \tag{6}$$

The IoU score ranges between 0 and 1, with higher values indicating greater similarity between the masks.

c) Regularization: To ensure smooth behavior, we introduce an aggregated regularization term, $\mathcal{P}_t \leq 0$:

$$\mathcal{P}_t = P_{J,t} + P_{A,t} + P_{V,t} + P_{F,t} + P_{S,t},$$

where $P_{J,t}$ penalizes joint torques, $P_{A,t}$ penalizes action deltas, $P_{V,t}$ penalizes angular velocities, $P_{F,t}$ penalizes foot slip and $P_{S,t}$ penalizes torso tilt. These regularization components are standard in robotic control frameworks and ensure that the policy adheres to physical constraints.

d) Combined Reward: The overall reward at each time step t is computed as a weighted sum of the video similarity, the image-based similarity, and the aggregated penalty:

$$R_t = \zeta\, S_{v,t} + \beta\, S_{M,t} + \eta\, \mathcal{P}_t,$$

where ζ, β, and η are scalar weights that balance the contributions of each term. This composite reward effectively aligns the rendered simulation with the generated video while promoting smooth behavior.

3.4. Policy Learning

NIL learns a policy, π_{s_i,b_j}, by maximizing the expected discounted return under the imitation reward. In contrast to state-of-the-art imitation learning approaches combining discriminators with reinforcement learning, we directly maximize the imitation reward using entropy-regularized reinforcement learning. This change eliminates the need for adversarial training and simplifies the learning process.

At each time step t, the agent receives an observation $o_t \in \mathcal{O}$, which comprises joint positions and velocities, and selects an action $a_t \in \mathcal{A}$ (i.e., the torques to be applied to the joints) according to the policy $\pi_{s_i,b_j}(a_t|o_t)$. The environment then provides an imitation reward, defined as:

$$R_t = \zeta\, S_{v,t} + \beta\, S_{M,t} + \eta\, \mathcal{P}_t,$$

where $S_{v,t}$ is the video similarity, $S_{M,t}$ is the image-based similarity, $\mathcal{P}_t$ is the aggregated penalty (see Sec. 3.3.3), and ζ, β, and η are scalar weights. The overall objective is to maximize the expected cumulative discounted reward:

$$J(\pi_{s_i,b_j}) = \mathbb{E}_{\pi_{s_i,b_j}}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right],$$

where $\gamma \in [0,1)$ is representing the discount factor.

In the entropy-regularized reinforcement learning, the policy is optimized by maximizing a soft value function that includes an entropy term to encourage exploration:

$$\max_{\pi} \mathbb{E}_{(o,a)\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(R_t + \alpha\, \mathcal{H}\big(\pi(\cdot|o_t)\big)\right)\right],$$

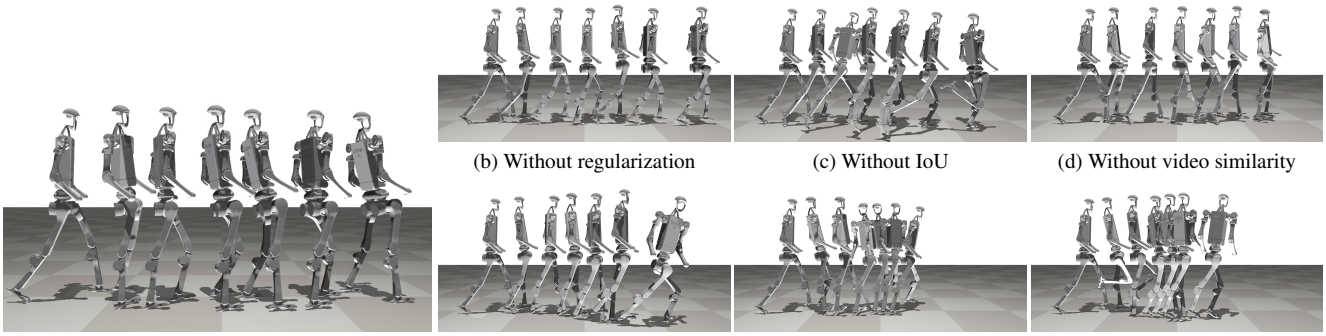
where $\mathcal{H}\big(\pi(\cdot|o_t)\big)$ denotes the entropy of the policy at state o_t, and α is a temperature parameter controlling the trade-off between reward maximization and exploration.

NIL effectively leverages the dense imitation reward signal, derived from the similarity metric, and reproduces expert motion patterns in the reference generated video F_{s_i,b_j}. We employ BRO [31] in our experiments, but our entropy-regularized RL formulation remains generic.

Temporal Alignment: The videos rendered by the physical simulation are 100Hz, which are higher than the generated videos, which are typically between 24-30Hz. Therefore, we upsample generated videos 4x using RIFE [23] then use them for imitation learning.

3.5. NIL

NIL provides a discriminator-free route to 3D skill acquisition from generated 2D videos by pairing a perceptual video-encoder similarity with a frame-wise mask IoU inside a physics simulator. The video generator supplies *visual guidance*; the simulator enforces *physical plausibility*; and the dense reward guides the learning. The result is a single recipe that applies across diverse and unconventional morphologies without curated 3D demonstrations.



(a) All components (b) Without regularization (c) Without IoU (d) Without video similarity (e) Only regularization (f) Only IoU (g) Only video similarity

Figure 3. **Reward Components:** Ablation of the reward function. **(a) All components:** All components are employed, and agent learns to walk well. **(b) Without regularization:** The resulting motion is jittery. **(c) Without IoU:** The learned behavior is distorted slightly. **(d) Without video similarity:** The walking is slower, and jittery. **(e) Only regularization:** Agent fails to walk straight, and employs suboptimal large leg movements. **(f) Only IoU:** Agent fails to walk forward continuously. **(g) Only video similarity:** Agent walks in a jittery way, and stops midway while walking.

4. Experiments

In this section, we evaluate the performance of No-data Imitation Learning (NIL) for 3D motor skill acquisition using generated 2D videos.

Experimental Settings: We use a fixed video encoder (TimeSformer [8] pretrained on Kinetics-400 [24]), a fixed clip length ($n_T = 8$) and input resolution for T (224x224), a 100 Hz control frequency, and identical observation/action spaces across methods. We use fixed $\zeta, \beta, \eta = 1$ for all embodiments. We run each configuration with multiple seeds and report the mean. Higher environment reward indicates faster, more stable locomotion (details in the Supplementary Materials).

Baselines: Since NIL is the only method that relies solely on generated data (without any curated or collected data of the same embodiment), we compare NIL against both upper and lower baselines. For locomotion experiments, we use imitation learning methods as baselines, and all baselines are trained using *motion-capture data* from [1] that is adapted to the simulation domain, with perfect joint correspondence. As upper baselines, we employ AMP [33], GAIfO [39], and as lower baselines, we consider Behavioral Cloning from Observations (BCO [38]). For loco-manipulation, since we do not have motion-capture data to imitate, we use reinforcement learning as the upper baseline and employ an off-policy RL method, BRO [31], as the baseline. We provide details regarding metrics in the Supplementary Materials.

We perform three ablation studies to analyze NIL:

- **Reward Component Ablation:** We analyze the impact of individual reward components on the performance.
- **Diffusion Model Comparison:** We compare several pretrained video diffusion models to determine which one provides the most effective reference demonstrations for imitation learning.
- **Improving Diffusion Models:** We assess how incremental advancements in video diffusion models affect the quality of the learned behaviors.

Then, we evaluate the performance of NIL in challenging robotic control tasks:

- **Locomotion across various embodiments:** Learning to walk with four different robot embodiments, each of which has different unique configurations and challenges.
- **Whole-body loco-manipulation**: Learning to sit, hang on a highbar, balance on a board purely from generated videos.

We present additional studies on the video selection protocol, open-source video diffusion models, sensitivity to the camera, and robustness to frame interpolation in the Supplementary Material.

4.1. Reward Components

To understand the contribution of each reward term, we train NIL on a walking task using the Unitree H1 humanoid robot. We evaluate performance using the environment reward, namely, the speed and stability of the learned policy.

Table 1 presents quantitative results, and Fig. 3 shows qualitative demonstrations. First, we analyze how the lack of individual components affects NIL. Overall, regularization helps NIL to smooth the learned motions, while both image-based and video-based similarity scores help the agent to understand the essentials of walking.

Second, we evaluate whether isolated components of the reward function enable imitating motions in generated reference videos. With only video similarity, NIL achieves a reasonable performance, albeit failing to generate visually plausible motions. In contrast, using only regularization or IoU rewards results in poor-performing policies.

Table 1. **Reward Ablation**: We analyze effects of each reward function component on the performance of NIL.

	Env. Reward ↑
NIL (all components)	396.1
without regularization	382.4
without IoU score	381.4
without video similarity	387.3
only regularization	363.6
only IoU Score	328.4
only video similarity	369.6
Expert	400

4.2. Diffusion Models for Imitation Learning

We evaluate the impact of different video diffusion models and systematically compare various diffusion models for their usability on imitation learning. We consider five open- and closed-source video diffusion models: Kling AI, Pika, Runway Gen-3, OpenAI Sora, and Stable Video Diffusion (SVD) [12]. For each model, we generate reference videos for the Unitree H1 walking task.

Quantitatively (see Fig. 4), Kling, despite exhibiting intermittent instabilities, yields the most visually plausible outputs and the highest NIL performance. Interestingly, even though Pika has shown limitations in physical plausibility [6], it still leads to high imitation scores. We hypothesize that the visual plausibility of the reference video is the most crucial property for NIL, as NIL is designed to refine physically implausible motions within the simulator.

We further analyze the correlation between the visual plausibility and the performance by comparing generated videos with a video of the simulated motion-capture-data trajectory. We measure the plausibility of the generated video as its perceptual similarity to the reference motion-capture video, which is only used for comparisons, and calculate the LPIPS score [48] between these two videos. As Fig. 4 shows, there is a positive correlation between the visual plausibility of generated videos and NIL's performance.

	LPIPS ↓	NIL Perf. ↑
SVD	0.070	366.5
Sora	0.204	370.8
Runway	0.115	383.7
Pika	0.081	385.9
Kling 1.0	0.081	388.7
Kling 1.6	0.046	396.1

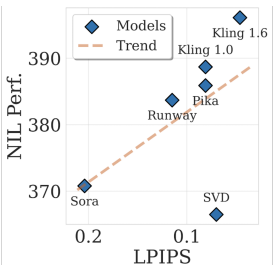
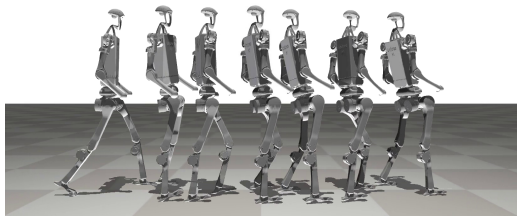


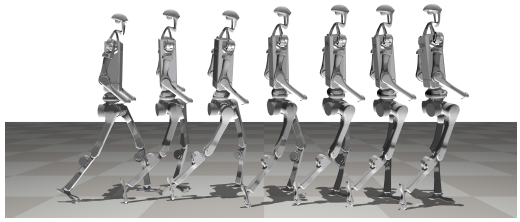
Figure 4. **Visual Plausibility:** We evaluate the correlation between the visual plausibility and the performance of NIL.

4.3. Improvements in Video Diffusion Models

To examine the sensitivity of NIL to advancements in video diffusion models, we compare two versions of Kling: v1.0 and v1.6. Both versions are used to generate reference videos for the Unitree H1 walking task. While the quantitative metrics are similar for both versions, qualitative results (see Fig. 5) reveal that the newer Kling v1.6 produces significantly more natural gaits. In contrast, the reference video from Kling v1.0 leads to an unbalanced gait with asymmetric leg movements.



(a) NIL trained on Kling v1.6



(b) NIL trained on Kling v1.0

Figure 5. **Different versions of video diffusion models:** (a) NIL learns to walk using the newest Kling version for reference video generation; (b) Reference video generated by the older version of Kling results in walking with an unbalanced gait.

This experiment underscores that even with small improvements in video diffusion models, the performance of NIL gets better. Therefore, better video diffusion models would enable NIL to learn more challenging tasks without using any collected/curated data.

4.4. Continuous Control of Various Robots

We test NIL on locomotion tasks across multiple robotic embodiments: three humanoid platforms (Unitree H1, Talos, and Unitree G1) and a quadruped (Unitree A1). For each robot, NIL is trained using a single reference video generated by Kling AI (Pika for Unitree A1), and we compare its performance against upper baselines (AMP, GAIfO) as well as a lower baseline (BCO), all of which are trained with 25 motion-capture trajectories from LocoMujoco [1]. Table 2 presents quantitative results. For the Unitree H1 and Unitree A1, NIL matches the performance of AMP, and it also produces a more natural and balanced walking gait. In contrast, for Unitree G1, even though NIL obtains competitive scores, AMP generates visually more natural and stable locomotion. With the Talos platform, both NIL and AMP

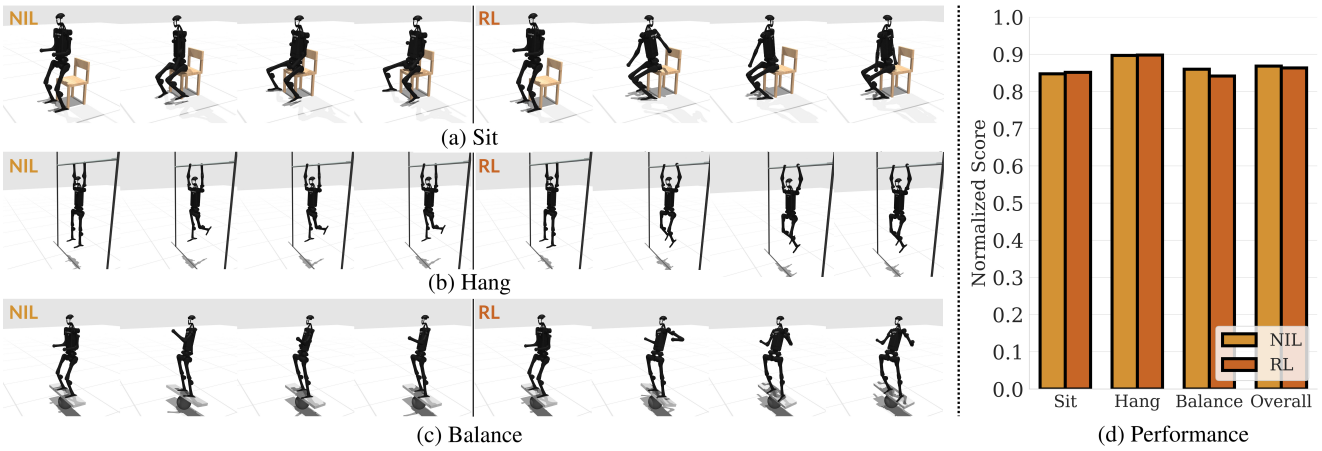


Figure 6. **Whole-body Manipulation:** NIL learns to (a) sit, (b) hang on a high-bar and (c) balance on a board from a single generated reference video. NIL matches the performance of RL in terms of normalized environment rewards (d).

Table 2. **Robotic Control:** We evaluate NIL on challenging robotic locomotion tasks across multiple robots.

	Environment Reward ↑			
	Unitree H1	Talos	Unitree G1	Unitree A1
NIL (ours)	396.1	352.8	356.9	290.3
AMP	393.5	231.1	393.4	286.9
GAIfO	347.8	204.4	353.1	260.8
BCO	72.0	26.6	21.2	30.3
Expert	400	400	400	300

face significant challenges due to the robot's complex morphology; however, NIL performs better and learns to move forward, albeit with less natural motion than desired.

4.5. Loco-Manipulation

Finally, we test NIL on whole-body manipulation tasks: sitting on a chair, hanging on a high-bar, and balancing on a board with the Unitree H1 robot. We compare its performance to the upper reinforcement learning baseline. The RL policy is trained using the reward function for these tasks defined in Supplementary Materials using BRO [31]. As shown in Fig. 6, NIL matches the performance of the RL baseline, learning from a single generated video. Both NIL and the RL baseline achieve 100% success rate; therefore we report the normalized environment reward in Fig. 6.

4.6. Summary of Experiments

Overall, our experimental results show that NIL, by leveraging generated data and a discriminator-free imitation reward, effectively learns task-achieving policies across diverse robotic platforms. The ablation studies underscore the importance of the reward components, while the diffusion model comparison highlights that visually plausible generation, even if not physically perfect, is important for effective imitation. We also present that improvements in video diffusion models enable better performance of NIL. These findings demonstrate the potential of NIL as a promising alternative to conventional, data-intensive imitation learning approaches. Future improvements in the video diffusion model could enable NIL to achieve more complex tasks, with different embodiments.

5. Discussion and Future Directions

We introduce NIL as a first step towards eliminating the dependency on curated expert data in imitation learning. By leveraging video diffusion models to generate expert demonstrations on-demand, NIL not only reduces the dependency on platform-specific data collection but also achieves competitive performance across diverse robotic platforms. One of the key insights from our study is that the performance of NIL is closely tied to the quality of the generated videos. As improvements in video diffusion models continue to emerge, NIL naturally benefits from these advancements, leading to more realistic behaviors.

Looking forward, several directions can further enhance the capabilities of NIL. First, integrating NIL as a pretraining step offers an exciting opportunity; the policies learned in a data-free manner can be fine-tuned using a small amount of curated data to boost performance, especially for complex morphologies where current methods face challenges. Also, extending NIL to more challenging tasks such as object interaction is an exciting direction.

In summary, NIL's performance is expected to improve with the rapid advancements in video diffusion and world models. We believe this work lays the foundation for future research at the intersection of generative modeling and imitation learning, providing a new approach to robot learning.

Acknowledgments

The authors thank Berna Kabadayi, Yarden As, Haiwen Feng and Shashank Tripathi for their discussions on the project, Benjamin Pellkofer for IT support, the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for supporting Mert Albaba, the Max Planck ETH Center for Learning Systems (CLS) for supporting Mert Albaba and Markos Diomataris, and the ETH AI Center for supporting Chenhao Li. While MJB was a co-founder and Chief Scientist at Meshcapade, his contributions were performed at, and funded by, the MPG.

References

[1] Firas Al-Hafez, Guoping Zhao, Jan Peters, and Davide Tateo. Locomujoco: A comprehensive imitation learning benchmark for locomotion. *arXiv preprint arXiv:2311.02496*, 2023. 6, 7

[2] Mert Albaba, Sammy Christen, Thomas Langarek, Christoph Gebhardt, Otmar Hilliges, and Michael J Black. Rile: Reinforced imitation learning. *arXiv preprint arXiv:2406.08472*, 2025. 3

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 2

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2, 3

[5] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995. 3

[6] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 7

[7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*. PMLR, 2021. 3, 6

[9] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 3

[10] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pages 306–324. Springer, 2024. 3

[11] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 3

[12] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3, 7

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[14] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 3

[15] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[16] Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. *Advances in neural information processing systems*, 30, 2017. 3

[17] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 3

[18] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018. 3

[19] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 852–861, 2019. 3

[20] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016. 3

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2, 3

[23] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 5

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[25] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[26] Jiayi Li, Tao Lu, Xiaoge Cao, Yinghao Cai, and Shuo Wang. Meta-imitation learning by watching video demonstrations. In *International Conference on Learning Representations*, 2021. 3

[27] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024. 3

[28] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. *Advances in neural information processing systems*, 30, 2017. 3

[29] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *Conference on Robot Learning*, pages 3943–3960. PMLR, 2025. 3

[30] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–958, 2026. 2, 3

[31] Michal Nauman, Mateusz Ostaszewski, Krzysztof Jankowski, Piotr Miłoś, and Marek Cygan. Bigger, regularized, optimistic: scaling for compute and sample efficient continuous control. *Advances in Neural Information Processing Systems*, 37:113038–113071, 2025. 5, 6, 8

[32] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *International Conference on Learning Representations*, 2018. 3

[33] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 3, 6

[34] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 3

[35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[38] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957, 2018. 6

[39] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018. 6

[40] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning*, pages 201–221. PMLR, 2023. 3

[41] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3

[42] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 3

[43] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 3

[44] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 3

[45] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021. 3

[46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[47] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[49] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 3

[50] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13577–13587, 2021. 3

[51] Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. In *International Conference on Machine Learning*, pages 60785–60811. PMLR, 2024. 3

[52] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning*, pages 61885–61896. PMLR, 2024. 3