DEXSIM: REAL-TIME DEXTEROUS SIMULATION WITH UNIFIED CAUSAL VIDEO DIFFUSION

Adam Lee
UC Berkeley
alee00@berkeley.edu

ABSTRACT

Recent progress of video diffusion models have enabled extensive simulation of the physical world. While simulation with hand object interaction has been less explored. We propose DexSIM, a dexterous simulation framework for simulating dexterous manipulation in real-time. While previous works utilizing video diffusion and 3D reconstruction focus on navigation, dexterous manipulation has been limited while it has extensive applications for creating interactive experiences with the simulated world and for generating synthetic data for robotics. Existing methods lack real-time interactivity and long-term spatial consistency and memory. We propose a 2-stage training framework for DexSIM. First we train a bi-directional video diffusion model by jointly embedding the hand action trajectory and video in a unified feature space. We utilize gaussian heatmap hand encoding for more accurate hand representation. Then we conduct a roll-out based autoregressive training with updated spatial cache as attention sink for spatial memory, which improves long-term consistency and 3D aware dexterous manipulation simulation. DexSIM outperforms the baseline on pixel and semantic similarity, motion fidelity, and hand projection accuracy. It also allows new applications such as hand motion transfer and runs at 15.24 FPS real-time interactivity.

1 INTRODUCTION

Diffusion and flow models have allowed realistic generation of images and video Labs et al. (2025); Wan et al. (2025); Brooks et al. (2024). With recent scaling and advancement of video diffusion has especially led to emergent abilities of simulating the physical world and interactions Brooks et al. (2024). Among generative simulation tasks dexterous manipulation simulation is an intuitive and natural extension of how we interact in the real world and thus allowing video diffusion models to simulate dexterous simulation can lead to natural interactions with the simulated world intuitively in creative applications. With long-term consistent and 3D aware simulations, dexterous generative simulations can also be used for training vision action models in robotics Huang et al. (2026). While recent works have explored low latency real-time interactive causal video diffusion models for gaming He et al. (2025), motion control Shin et al. (2025); Zhao et al. (2025), and live avatar Huang et al. (2025b), there hasn't been previous work for real-time interactive video generation for dexterous manipulation simulation.

Dexterous manipulation task aims to predict future states acting as a world model based on current state, given an action signal. As opposed to animation Wang et al. (2025), hand inpainting tasks Chen et al. (2025), the model must learn the dynamics its actions have on the environment and the more precise knowledge it has on the environment, input image, it achieves more physically realistic generations. This also differs from navigation tasks He et al. (2025); HunyuanWorld (2025) which have much less changes in the environment and the dynamics and properties of interactions with different objects is not required. While previous work has focused on 2D signals Akkerman et al. (2025) and bi-directional approaches Kim et al. (2025), we propose a new framework for causal generation with spatial cache to embody spatial awareness and memory. For manipulation, action conditions should be more precise, while 2D signals are limited to have ambiguity. While real time interactive has not yet explored for manipulation tasks which greatly hinders usage and adoption in commercial applications. We first train a bi-directional model that is most powerful. Then convert to

an autoregressive model through self roll out with iterative training from teacher forcing to roll-out based Distribution Matching Distillation (DMD) Yin et al. (2024) training.

Prior works in dexterous manipulation Akkerman et al. (2025); Kim et al. (2025); Goswami et al. (2025) face two significant challenges: lack of spatial awareness and consistency and real-time interaction. Spatial consistency is required for long-term consistency and accurate simulations since dexterous action conditions in 3D space can be ambiguous leading to inaccurate generation of hand object interactions. Also low-latency and real-time interactivity is required for creative applications such as augmented reality, whereas prior works lack causal generation or real-time interactive generation.

To address these challenges we propose DexSIM, a 2-stage real-time causal generative simulation framework for dexterous manipulation. Given an input image DexSIM can generate video temporally autoregressively given streaming input of hand trajectories.

DexSIM is a real-time interactive causal video diffusion model with unified feature representation and spatial cache mechanism allowing real-time interactivity for dexterous manipulation simulation. First, we train a bi-directional video diffusion model that generates the entire frame sequence by jointly embedding the hand trajectory, initial 3D point and noise in the same space. Then we train our model with roll-out using DMD while including a updated spatial memory and initial scene attention sink to generate long-term consistent spatial aware interactive generation. Causal distillation results in a low-latency DexSIM with greater throughput with less number of function evaluations (NFE) resulting in a distilled 4-step model for real-time interactive dexterous manipulation simulation. We rigorously evaluate DexSIM comparing to baseline on appearance similarity, and projected hand distance.

Our contribution can be summarized as follows:

- We propose DexSIM, a real-time interactive dexterous manipulation model based on a two-stage training framework with a unified feature space and gaussian heatmap hand encoding for more accurate hand representation.
- DexSIM generates real-time interactive simulations with spatial awareness and memory with an updatable spatial cache, resulting in a 15.24 fps interactive dexterous manipulation simulation.
- We rigorously evaluate DexSIM which outperforms the baseline on appearance quality, hand pose accuracy, and motion smoothness and allows new applications such as hand motion transfer.

2 Related Works

Causal Video Distillation Flow matching is an ode solution to project from distribution p to q. Starting with image generation, video generation is enabled with an additional temporal dimension recent methods scaling DiT based architectures Brooks et al. (2024); Wan et al. (2025); Kong et al. (2024). Causal generation can be naively generated with teacher forcing but severly lacks long-term consistency. To address this challenge, diffusion forcing proposes progressively different levels of noise and history conditioning for video generation. Distribution Matching Distillation (DMD) Yin et al. (2024) initially proposed to utilize score matching from variational score distillation (VSD) Wang et al. (2023) for step-distillation has been a crucial component for causal video distillation Yin et al. (2025). Self-forcing Huang et al. (2025a) applies self-rollout to decrease train-test distribution gap.

Interactive Video Generation LongLive conducts causal video generation with prompts but lacks specific detailed conditions. Drag simulation Zhao et al. (2025); Shin et al. (2025) has also been applied to synthesize motion from dragging but it lacks accurate 3D conditioning mostly relying on ambiguous 2D signals and lacks interactive capabilities with the 3D environment. Avatar lip syncing and simulation also focus on generating appearance wise realistic human pose and motion while interactions with surrounding objects and scene are limited. Interactive causal video generation has been applied to different applications in recent works. Including gaming primarily focused on scene navigation He et al. (2025); Kong et al. (2024). While these models are impressive and can conduct 3D tasks such as novel view synthesis one-shot, it lacks the ability to physically interact with the environment where the dynamics and forces are applied to different surrounding objects.

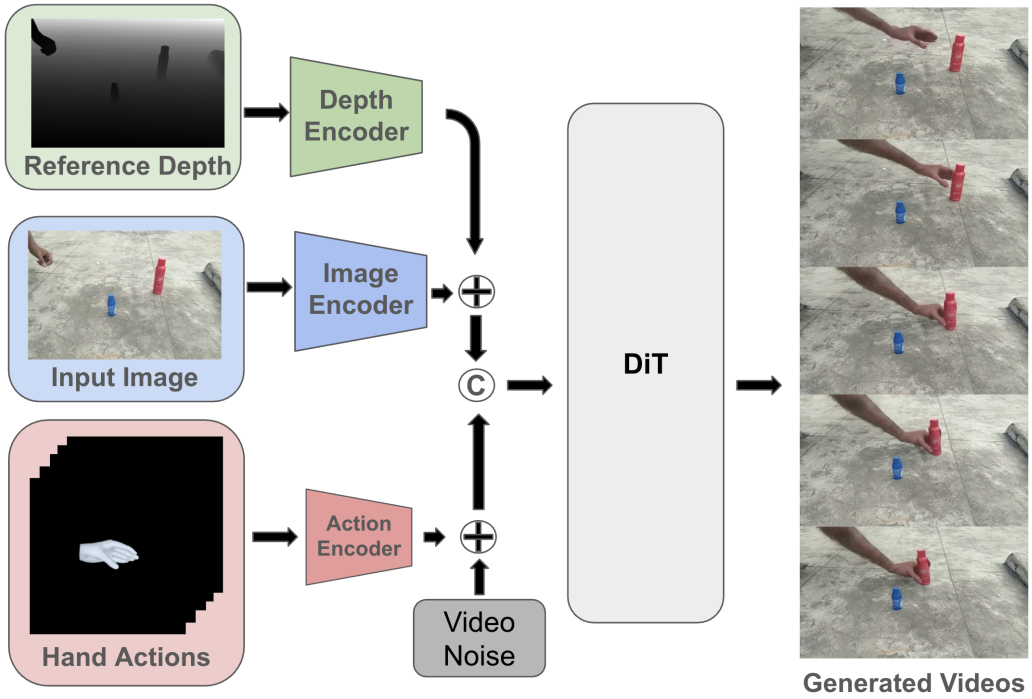

Figure 1: Overall architecture of DexSIM

Dexterous Manipulation InterDyn Akkerman et al. (2025) generates video with bi-directionally based on 2D segment conditioning trained on controlnet. DWM Kim et al. (2025) generates video bi-directionally trains a video diffusion model to learn from hand warped video. DexWM Goswami et al. (2025) is trained from hand pose difference as action signals to a causal video generation model. It has 3D signal for hand action but lacks 3D awareness of scene and memory. Also lacks real-time interactivity and long-term consistency due to train test time gap addressed by DMD and self-forcing. EgoEdit Li et al. (2025) performs egocentric hand streaming input video can interact with objects based on text input but lacks ability to interact with given 3D, generates hand interacting augmented reality. Since in robotics tasks and for 3D scene interaction, accurate 3D action conditioning and spatial awareness is required for the model to learn which objects are simulated it's dynamics magnitude. While other works have proposed non-egocentric simulation words, PlayerOne Tu et al. (2025) the realistic interaction is not as much of a priority as egocentric simulation and serves as a novel view synthesis task. EgoEdit generates egocentric simulation based on text context which also has limited explicit action conditioning and the focus is less on accurate interaction with the given environment but generating synthetic objects and assets for interaction that matches the text prompt. Interactivity requires better spatial consistency and awareness. Since it interacts with the 3D world and objects from the given input. Hand inpainting Chen et al. (2025) and animation generation Wang et al. (2025) are similar in 2D key point conditioning, these tasks focus on accurate pose generation without learning to simulate the dynamics between the pose and the environment.

3 Preliminaries

3.1 Video Diffusion

Diffusion models learn a reverse-time denoising process that transforms a simple noise distribution into the data distribution. Let $\mathbf{x}_0 \sim p_d$ denote a data sample. A forward noising process q length T has a variance schedule $\{\beta_t\}_{t=1}^{T}$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The forward process admits the closed form

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{1}$$

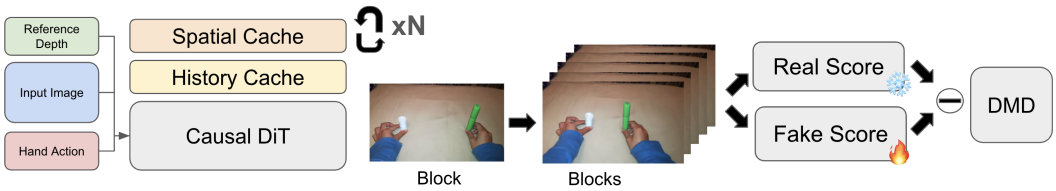

Figure 2: Rollout training with Spatial Memory

equivalently,

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; t \in \{1, \dots, T\}. \tag{2}$$

A neural network ϵ_θ is trained to predict a denoising target from a noisy sample $\mathbf{x}_t$ and timestep t, and a conditioning c. Using the common noise-prediction parameterization, the objective is

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t\sim p(t),\, \mathbf{x}_0 \sim p_d,\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[w(t)\,\big\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c) - \boldsymbol{\epsilon}\big\|_2^2\Big], \tag{3}$$

Modern implementations conduct the diffusion process in the latent space with a diffusion transformer (DiT) Peebles & Xie (2023) backbone. And the denoised latent is decoded to pixels through a 3D VAE.

3.2 Distribution Matching Distillation

Distribution Matching Distillation (DMD) utilizes the score difference between teacher and fake model given the noised latent of a fully denoised latent from the generator.

$$\nabla\mathcal{L}_{\text{DMD}} = -\mathbb{E}_t\left(\int \Big(s_{\text{real}}\big(F(G_\theta(z), t), t\big) - s_{\text{fake}}\big(F(G_\theta(z), t), t\big)\Big)\,\frac{dG_\theta(z)}{d\theta}\,dz\right). \tag{4}$$

The generator can be a causal model while the teacher is bi-directional, more expressive bigger model and the fake model updates more frequently to match generator distribution. As a distribution matching objective VSD was originally proposed for 3D distillation but DMD is utilized for step distillation while recent works Yin et al. (2025); Huang et al. (2025a) have adapted it for causal video distillation. Other distillation and joint distillation of different properties such as cfg, step, and causal video is possible and remains a future work to inspect the optimal settings for joint distillation.

4 Method

DexSIM predicts the future frames given an input image and hand action conditions, and depth as a world model. It has to learn to model the hand and naturally interact with the objects in the given input image and environment using pre-trained priors and depth information. We focus on egocentric views to model how people interact with objects in the real world. In the following sections we describe our architecture and the 2-stage training process with spatial cache.

4.1 Task formulation

Given a state S, we aim to predict the future states $\{s_t\}_{t=1}^T$ conditioned on actions $\{c_t\}_{t=1}^T$ with $c_t \in \mathcal{C}$. We represent each state by a latent image encoding $s_t := \text{enc}(I_t)$. For the first-stage bi-directional training stage, DexSIM predicts all future states immediately:

$$p(s_{1:T} \mid s_0,\, c_{1:T},\, d_0)\,, \tag{5}$$

where d_0 is the depth latent of the initial state. And DexSIM predicts all future states $s_{1:T}$ at once. Causal DexSIM from the second stage training aims to predict the next state autoregressively resulting in the following formulation:

$$p(s_t \mid s_{<t},\, c_{\leq t},\, d)\,. \tag{6}$$

Here c_t is implemented as sliding window attention with n most recent frames and first frame initial sink. d represents the updatable spatial cache which is described in the following section.

4.2 ARCHITECTURE

We employ a bi-directional video diffusion model as our base model and train by jointly embedding the hand trajectory and initial 3D map to condition the generation.

We adopt a hand action representation framework inspired by the hand inpainting domain. As mentioned in Chen et al. (2025) 3D hand representations such as Mano are more expensive to gather and also less robust to occlusion. To provide spatial signals we include the initial depth map in the image encoder.

Unified feature representation Inspired by animation literature Wang et al. (2025) we adopt a unified feature representation As hand action signal is crucial and requires long term consistency. We adopt a DiT architecture with a unified feature space with the condition and noise. While previous work Akkerman et al. (2025) and Controlnet Zhang et al. (2023) employ a dual model for conditioning or dual network for joint prediction facilities feature alignment with less parameters. Memory optimization is crucial for 2nd stage causal distillation and DMD.

While previous methods implicitly learn relevant state of the environment including current pose and the spatial distance, we incorporate the references with separate encoders. The target action is encoded and its feature is fused with the noise latent.

Hand action and state conditioning We use mediapipe landmark keypoints for both hands resulting in keypoints $p \in \mathbb{R}^{42\times 2}$ and project it to gaussian heatmap following Chen et al. (2025):

$$\mathcal{H} = \mathcal{N}(x \mid p, \sigma^2), \tag{7}$$

resulting in 42 channel hand representation $\mathbf{h}_t \in \mathbb{R}^{42\times h\times w}$ which we encode with a convolutional encoder resulting in a hand action latent $\mathbf{l}_t \in \mathbb{R}^{c\times h'\times w'}$. The size of latent channel c is set to 1536. With hand positioning being a crucial signal we follow Wang et al. (2025) and add the hand pose to the initial noise. This hand representation is much more expressive in representing hand pose than previous work Akkerman et al. (2025) relying on 2D masks. We adopt a depth map of the initial frame and project it to the joint space as the latent space. This state acts as a guidance to infer spatial dynamics and action correspondences complementing the 2D hand representations.

To increase accurate hand pose projection and generation we refer to the hand inpainting literature Chen et al. (2025) and employ a temporal hand trajectory projection, projection 22 keypoints from both hands into noise token space.

Our model is trained on the latent space $\mathbf{z}_0 \in \mathbb{R}^{f\times c\times h\times w}$ with standard diffusion loss with depth c_h and hand c_h conditions:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, \mathbf{c}_d, \mathbf{c}_h)\|_2^2\right]. \tag{8}$$

During inference, hand from source streaming video is projected to source image and an initial 3D map is utilized for spatial awareness while simulating from hand trajectories that require inductive bias to accurately simulate surrounding object manipulation.

4.3 ROLL-OUT TRAINING WITH SPATIAL CACHE

The main challenges with autoregressive generation is long term consistency which is more challenging for dexterous interaction generation as they require spatial understanding to generate realistic simulations and accurate interactions with the environment. Whereas, for dexterous manipulation simulation, real-time interactivity has been the bottleneck for usage in creative applications and robotics tasks requiring real-time guidance. To address these challenges we first train the model with teacher forcing with causal attention. Then train with self roll out based DMD loss where the bi-directional DexSIM is used as the teacher model.

We adopt a spatial cache that acts as an attention sink along with the frame sink used in causal video generation Yang et al. (2025); Huang et al. (2025b). The spatial cache is updated with new blocks with cross-attention. We encode the initial depth of the scene to project to generate key and value

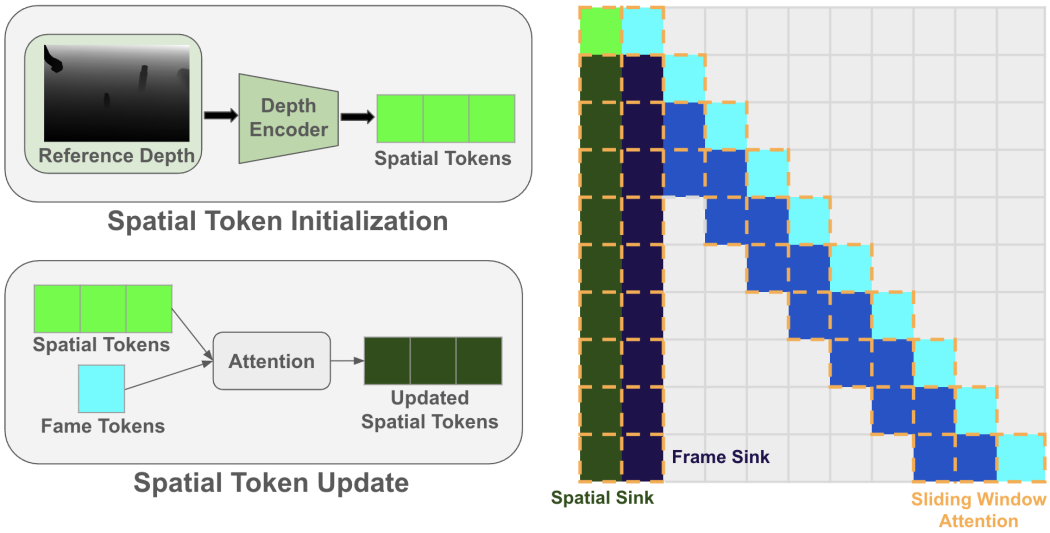

Figure 3: Spatial Cache

Method	SSIM ↑	PSNR ↑	LPIPS ↓	Motion Fidelity ↑	PCK@20 ↑
InterDyn	0.506	11.64	0.388	0.576	61
DexSIM Bi-directional	0.573	15.92	0.322	0.613	71
DexSIM Causal	0.526	12.21	0.362	0.594	65

Table 1: Comparison with baseline . ↑ higher is better, ↓ lower is better.

tokens. The tokens are subsequently updated after a new block is generated from the causal DiT with cross attention. This updated spatial guidance allows the model to have long term spatial awareness and memory.

We design an updatable spatial cache which is updated with frame-wise autoregressive video generation during inference. Initialized by the 3D map used as the training condition for the bi-directional model $d \in \mathbb{R}^{1\times h\times w}$, we project the depth map using a convolutional encoder to match the embedding space of the DiT latents resulting in a depth latent $d' \in \mathbb{R}^{n\times c}$ where n is the number of tokens and c is the channel. In practice, we use 256 tokens and 1536 channels. They are updated by the initial frame latent tokens with cross-attention and for each subsequently generated frame latent blocks, we update the spatial cache tokens for more accurate 3D consistent generation. This update allows the depth conditioning to modify based on the new frames and it's changes from the previous frames.

We adopt the initial 3D map as condition through attention sink along with updatable spatial memory to preserve long-term 3D consistency and awareness along with providing with an initial spatial reference point. Resulting in a total 512 additional tokens that are used as conditions for causal frame generation. Through both updatable spatial memory and initial scene attention sink, DexSIM is capable of depth aware dexterous manipulation. Autoregressive generation allows interactive and real time interactions with the conditioned environment.

5 Experiments

5.1 Experimental settings

We initialize our model with a Wan 1.3B base model Wan et al. (2025). The the action encoders are implemented with convolutional layers. The experiments are done in 8 H100 GPUs. Number of frames is set as 81 with spatial resolution resized to 768 x 512 with 5k steps for stage 1 and 2k steps

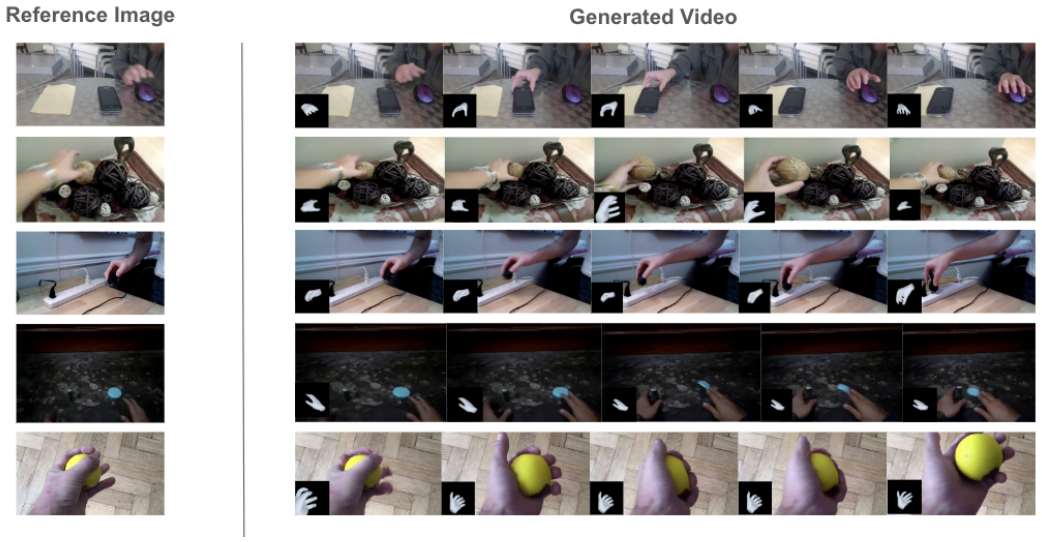

Figure 4: Qualitative Results

for stage 2. Per GPU batch size is 4 for stage 1 and 1 for stage 2. We follow self-forcing to reduce memory footprint by using adopting gradient accumulation and stop gradient for each step. We train both stages with LoRA Hu et al. with rank 64.

We employ hand object interaction dataset something-something-v2 (ssv2) to train our model. To filter for quality we filter out frames that do not have hand action signals and to mitigate only sample video chunks with a continuous hand signal where a hand signal not present for 12 consecutive frames is discarded and a new chunk is initialized when they hand is visible again.

Metrics While appearance and motion smoothness rely on scaling model, data and compute, DexSIM focuses on hand pose accuracy and motion fidelity of generated simulations. For evaluation we use the test set of the ssv2 dataset which contains hand object interaction videos at scale the majority at a static setting.

We compare with the baseline InterDyn which similarly generates videos of forced conditioned based on a 2D mask. We acquire the segmentation by getting the keypoints and using them as a prompt for SAM to generate the condition masks. We utilize both perceptual and pixel level similarity metrics as well as PCK@20 which evaluates the amount of keypoints that are within 20 pixels of target. We use a Cotracker Karaev et al. (2024) based motion fidelity metric by Yatim et al. (2024) to measure motion fidelity. We follow a similar implementation to Akkerman et al. (2025) and track the pixels of the hand masks from the initial image with two tracklets from the ground truth and the comparison model $T = \{\tau_1, \ldots, \tau_n\}$ and $\tilde{T} = \{\tilde{\tau}_1, \ldots, \tilde{\tau}_m\}$..

$$\frac{1}{m} \sum_{\tilde{\tau} \in \tilde{T}} \max_{\tau \in T} \mathbf{corr}(\tau, \tilde{\tau}) \; + \; \frac{1}{n} \sum_{\tau \in T} \max_{\tilde{\tau} \in \tilde{T}} \mathbf{corr}(\tau, \tilde{\tau}), \tag{9}$$

where the correlation $corr$ is defined as the following:

$$\mathbf{corr}(\tau, \tilde{\tau}) = \frac{1}{F} \sum_{k=1}^{F} \frac{v_k^x \cdot \tilde{v}_k^x \; + \; v_k^y \cdot \tilde{v}_k^y}{\sqrt{(v_k^x)^2 + (v_k^y)^2} \; \cdot \; \sqrt{(\tilde{v}_k^x)^2 + (\tilde{v}_k^y)^2}}, \tag{10}$$

5.2 Baseline

We compare with InterDyn as the alternative hand pose conditioned interaction simulation model and demonstrate our model significantly outperforms in pixel similarity, motion fidelity, and hand projection accuracy. InterDyn employs a controlnet to condition a video diffusion model on 2D masks to learn force propagation and dynamics. While employing a 2D masks allow a more general

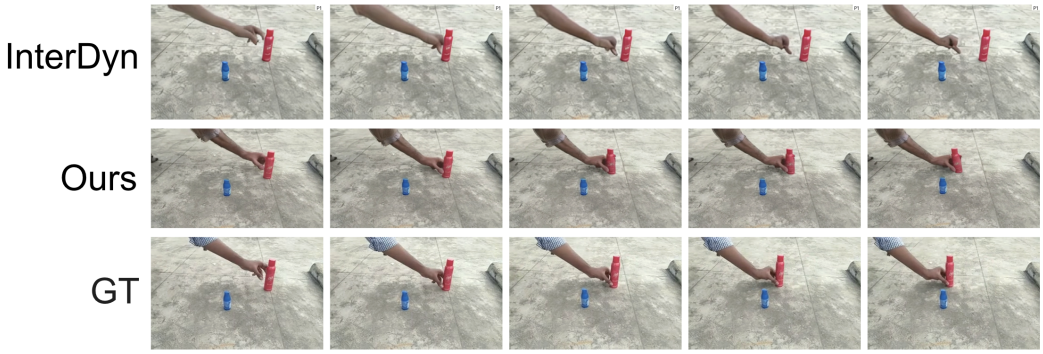

Figure 5: Qualitative Comparisons

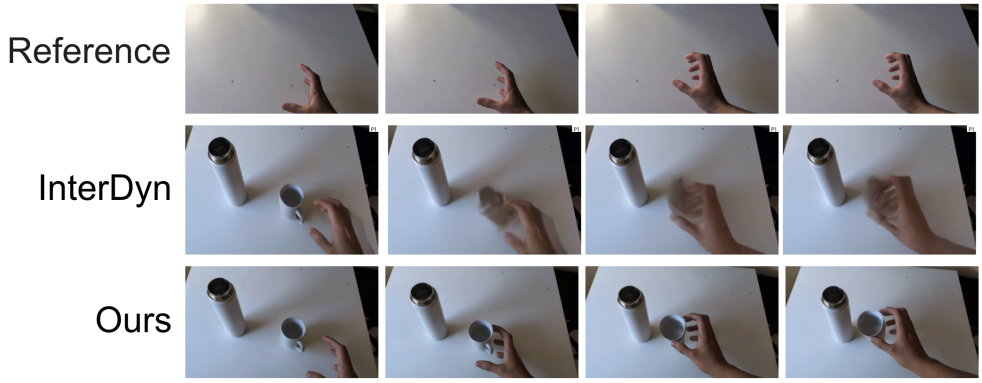

Figure 6: Qualitative Comparisons with Reference

approach for force dynamics simulation, it lacks explicit hand conditioning and spatial awareness of the environment.

5.3 Quantitative Comparisons

DexSIM outperforms InterDyn on visual fidelity, motion smoothness and hand projection accuracy. PCK@20 determines the percentage of keypoints that are within 20 pixels of the ground truth. With spatial cache and depth signals, the object maintains better consistency throughout the generated frames as can be seen in Fig. 5. And the explicitly conditioned hand pose allow demonstrate much improved hand projection accuracy compared to the baseline.

Notably there is a quality and latency trade-off between the bi-directional model and causal versions of DexSIM. While the bi-directional model learns directly from the ground truth, the causal model's performance can be attributed to two factors: 1) multi-stage training runs and 2) inference time error accumulation. The causal model underperforms the bi-directional model on frame similarity, motion and hand projection metrics as the causal distillation training process relies on the bi-directional model as the teacher model and during inference, error is accumulated overtime as a frame-wise autoregressive model where as the bi-directional model generates all frames at once. However, the causal model is able to generate with 4 NFEs with an order of magnitude faster FPS in 15.24 compared to bi-directional model's 50 NFEs and 1.44 FPS along with much faster end-to-end time to first frame.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	PCK@20 ↑
DexSIM	0.526	12.21	0.362	65
w/o Spatial Cache	0.506	11.58	0.374	65

Table 2: Spatial cache ablation comparison.

Methods	**NFE**	**FPS ↑**	**TTFF ↓**
DexSIM Bi-directional	50	1.44	57.24
DexSIM Causal	4	15.24	2.12

Table 3: Runtime and latency comparison. TFF indicates the end-to-end time-to-first-frame.

5.4 Qualitative Comparisons

Fig. 5 demonstrates the qualitative results of DexSIM and it's ability to conduct specially accurate dexterous simulations compared to InterDyn. DexSIM is able to accurately target the object and generate realistic simulations and interactions while following the hand action guidance signal. 2D signals such as segment masks have innate ambiguity as conditions passed to the diffusion models therefore, InterDyn fails at grasping objects with 2D mask conditioning more frequently.

Fig. 6 we show motion transfer as an application of DexSIM where a motion from a source can be used for a different target video while manipulating and interacting with the object. Reference image and motion transfer is much more challenging with depth ambiguity of the environment and the hand signal leading to blurry generations by the baseline.

5.5 Ablation

Spatial cache We conduct ablation experiments to investigate the significance of spatial cache for spatial memory and long-term consistent dexterous interaction generation. Spatial cache allows more accurate manipulation with objects with additional depth conditioning resulting in better performances for pixel and perceptual similarity where as we found the hand projection accuracy is similar due to the hand projection in 2D space not requiring spatial information.

Causal Distillation Distilling with DMD with bi-directional DexSIM as teacher to generate frames autoregressively results in a low-latency causal DexSIM with greater throughput with less number of function evaluations (NFE) resulting in a distilled 4-step model for real-time interactive dexterous manipulation simulation. And the FPS is significantly increased with end-to-end time-to-first-frame is greatly reduced through causal distillation.

6 Conclusion

Here we present DexSIM a real-time interactive dexterous manipulation simulation model train from a 2-stage training process with a unified feature representation and roll-out training with spatial cache to provide spatial conditions for more accurate hand-object interaction. While DexSIM focuses on improvements in spatial aware and memory generation, the scale of dataset and model size limits the model's performance in overall realism and appearance. Scaling compute with stronger base model and model and dataset size will significantly increase appearance realism and motion smoothness. Joint video and step distillation could be employed for reducing latency and denoising steps. And data free distillation to learn the teacher distribution better without overfitting on the training-data. Finally the spatial cache could be updated with more intricate techniques such as optical flow to preserve the depth information more semantically accurately.

References

Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12467–12479, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17448–17460, 2025.

Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models can leverage human videos for dexterous manipulation. *arXiv preprint arXiv:2512.13644*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025a.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b.

Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proc. arXiv:2410.11831*, 2024.

Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models. *arXiv preprint arXiv:2512.17907*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Runjia Li, Moayed Haji-Ali, Ashkan Mirzaei, Chaoyang Wang, Arpit Sahni, Ivan Skorokhodov, Aliaksandr Siarohin, Tomas Jakab, Junlin Han, Sergey Tulyakov, et al. Egoedit: Dataset, real-time streaming model, and benchmark for egocentric video editing. *arXiv preprint arXiv:2512.06065*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. *arXiv preprint arXiv:2506.09995*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 68(10):1–14, 2025.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8466–8476, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Kesen Zhao, Jiaxin Shi, Beier Zhu, Junbao Zhou, Xiaolong Shen, Yuan Zhou, Qianru Sun, and Hanwang Zhang. Real-time motion-controllable autoregressive video diffusion. *arXiv preprint arXiv:2510.08131*, 2025.