This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

InterDyn: Controllable Interactive Dynamics with Video Diffusion Models

Rick Akkerman[1,2*] Haiwen Feng[1*†] Michael J. Black[1] Dimitrios Tzionas[2] Victoria Fernández Abrevaya[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Amsterdam, the Netherlands
{rick.akkerman, haiwen.feng, black, victoria.abrevaya}@tuebingen.mpg.de d.tzionas@uva.nl

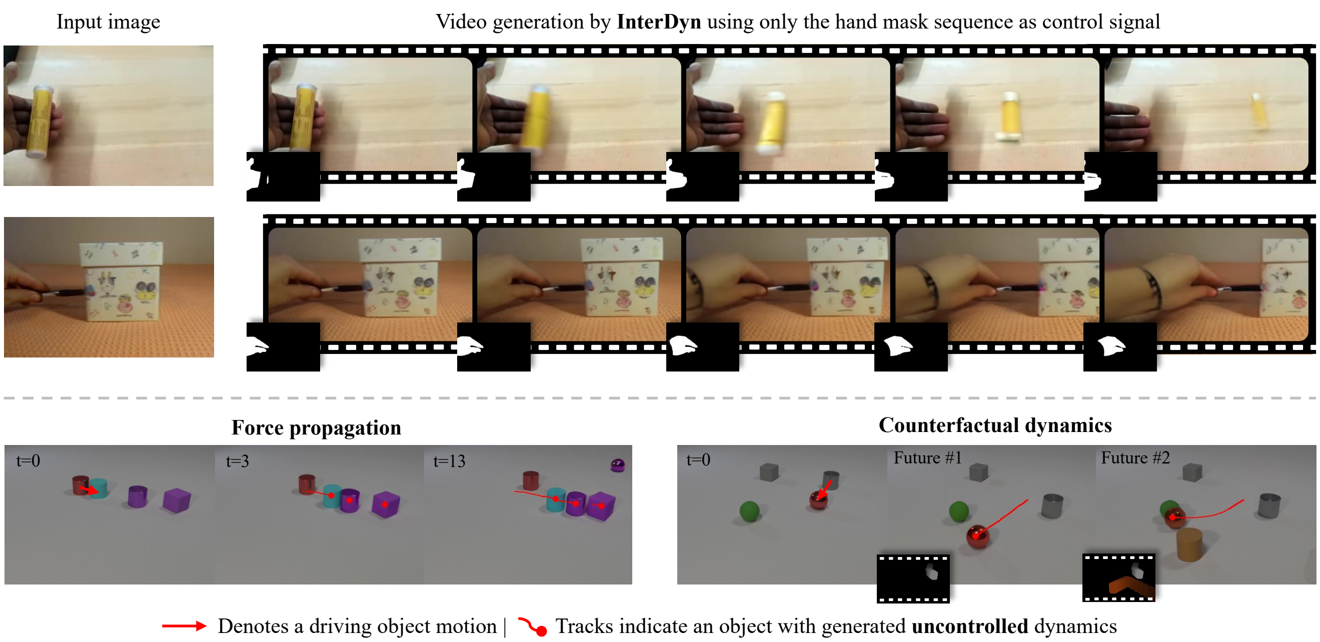


Figure 1. We present **InterDyn**, a framework for synthesizing realistic interactive dynamics without 3D reconstruction and physics simulation. Our core principle is to rely on the implicit physics knowledge embedded in large-scale video generative models. Given an image and a "driving motion", our model generates the consequential scene dynamics. We investigate the generated interactive dynamics in a simple object collision scenario (bottom) and complex in-the-wild human-object interaction (top).

Abstract

Predicting the dynamics of interacting objects is essential for both humans and intelligent systems. However, existing approaches are limited to simplified, toy settings and lack generalizability to complex, real-world environments. Recent advances in generative models have enabled the prediction of state transitions based on interventions, but focus on generating a single future state which neglects the continuous dynamics resulting from the interaction. To address this gap, we propose InterDyn, a novel framework that generates videos of interactive dynamics given an initial frame and a control signal encoding the motion of a driving object or actor. Our key insight is that large video generation models can act as both neural renderers and implicit physics "simulators", having learned interactive dynamics from large-scale video data. To effectively harness this capability, we introduce an interactive control mechanism that conditions the video generation process on the motion of the driving entity. Qualitative results demonstrate that InterDyn generates plausible, temporally consistent videos of complex object interactions while generalizing to unseen objects. Quantitative evaluations show that InterDyn outperforms baselines that focus on static state transitions. This work highlights the potential of leveraging video generative models as implicit physics engines. Code and trained models will be released at: https://interdyn.is.tue.mpg.de/.

*Equal contribution
†Project lead

1. Introduction

Humans have the remarkable ability to predict the future dynamics of observed systems intuitively. With just a single image, we can anticipate and imagine how objects will move over time – not only their motion but also their interactions with the environment and other elements in the scene. Inferring this requires an advanced form of scene-level reasoning beyond merely recognizing the semantics and geometry of static elements; it involves a deep physical and causal understanding of how each object will interact given the environment, object properties, and forces.

There has been a growing interest in developing machine learning systems that emulate similar levels of dynamic understanding given visual observations, such as images or videos. Early work [84] addressed this by first reconstructing a 3D representation from the image, then predicting future states with a physics simulator and finally generating the output video with a rendering engine. This relies heavily on explicit reconstruction and simulation, which is computationally intensive, prone to errors, and may not generalize well. More recent methods [2, 24, 39, 46, 49] leverage keypoint or latent representations within graph relational frameworks; however, they have only been trained and validated in over-simplified, synthetic environments, showing limited generalizability to complex real-world scenarios.

Instead, the advent of powerful generative models [1, 5, 16, 54, 66] opens new avenues for synthesizing interactions under complex scenarios. For example, Sudhakar et al. [70] recently proposed CosHand, a controllable image-to-image model based on Stable Diffusion [66] that infers *state transitions* of an object. The task here is defined as follows: given an image of a hand interacting with an object, alongside a hand mask of the current frame and a mask of the hand at a future frame, generate a modified input image that satisfies the mask, with realistic interactions. The challenge, as in early intuitive physics works, lies in accurately modeling how the objects will change after forces are applied. However, we argue that static state transitions are insufficient for this task, as they fail to capture the continuous dynamic processes inherent to the problem, e.g. see Fig. 2. Investigating interactive dynamics within a two-state setting is highly limiting, since dynamics can extend beyond the period of direct contact – for example, predicting the motion occurring while a person pours water requires a physical understanding that goes beyond the state of the hand at a future frame. The driving force, in this case the hand, may interact with the system only briefly, but the system's subsequent dynamics continue according to physical laws and may even influence other parts via force propagation.

In this paper, we explore *controllable synthesis of interactive dynamics*–generating a video from an input image and a dynamic control signal (e.g. a moving hand mask) to model realistic object dynamics. In particular, we propose

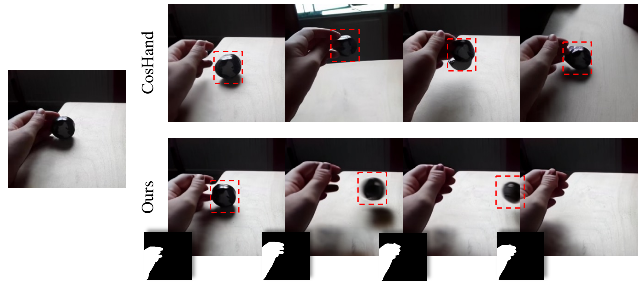


Figure 2. **State transition vs. dynamics.** Methods that generate static state transitions (i.e. predict a future image) such as CosHand [70] struggle to capture the inherent dynamic processes involved in human-object interactions. Here, we show a video sequence where the motion continues beyond the interaction.

InterDyn, a novel framework for synthesizing controllable dynamic interactions that leverages the physical and dynamics "knowledge" of a large video generation model [5]. Unlike prior approaches that rely on explicit physical simulation [84] or are constrained to static state transitions [70], we leverage video generation models to generate dynamic processes implicitly, see Fig. 1. Specifically, we extend Stable Video Diffusion (SVD) [5] with a dynamic control branch and fine-tune it on diverse scenes, enabling synthesis of complex interactions aligned with the control signal.

We start our investigation by fine-tuning InterDyn on a simple synthetic scenario of cubes, cylinders, and spheres: the CLEVRER dataset [97]. To control the motion we add a mask driving signal that manipulates the movement of some (but not all) of the objects in the scene. We then evaluate how the synthesized trajectories of uncontrolled objects change under various interactions, including multiple objects colliding with each other. This multi-object collision setting allows us to "probe" the physical understanding and causal effects of the video diffusion model, and our qualitative experiments show InterDyn's ability for counterfactual future prediction and physical force propagation.

Next, we evaluate how the system performs in a difficult real-world scenario, such as Human-Object Interaction (HOI). Here, the dexterity of hand motions and the diversity of objects vastly increase the complexity of the problem. We fine-tune the model on a commonly used HOI video dataset [21] and compare with the state-of-the-art baseline CosHand [70], as well as two text-control based interactive dynamics generation methods: Seer [23] and DynamiCrafter [89]. We quantify our investigations using standard image and video metrics, as well as a motion fidelity metric based on point tracking. InterDyn surpasses the previous SOTA over 37.5% on LPIPS and 77% on FVD on the Something-Something-v2 (SSV2) dataset [21]. Our experiments also demonstrate diverse and physically plausible generations of interactive dynamics, probing into SVD's "understanding" of physics and dynamics.

In summary, we present InterDyn, a framework that employs video generative models to simulate object dynamics without explicit 3D reconstruction or physical simulation. We demonstrate how the inherent "knowledge" within video generation models can be leveraged to predict complex object interactions and movements over time, implicitly modeling physical and causal dynamics. We perform comprehensive experiments on multi-object collision datasets and hand-object manipulation datasets, demonstrating the effectiveness of our approach.

2. Related Work

Modeling human-object interactions (HOI). Human-object interaction has been widely studied within the context of 3D reconstruction [17, 18, 27, 28, 73, 88, 94], where the goal is to recover realistic geometry of hands and objects. The field of 3D HOI synthesis has also received increasing attention, including the generation of static [40, 44, 71, 106] or dynamic [63, 72, 103, 105] hand poses conditioned on 3D objects, whole-body interactions [90], or more recently, hand-object meshes given textual descriptions [10, 15, 57, 86, 96]. Few works address HOI synthesis in the 2D domain. GANHand [14] predicts 3D hand shape and pose given an RGB image of an object, while AffordanceDiffusion [95] estimates a 2D hand using a diffusion model. Kulal et al. [47] take as input an image of a human and a scene separately and generate a composite image that positions the human with correct affordances. Also relevant is HOIDiffusion [101], in which a texture-less rendering of a 3D hand and object is converted to a realistic image using a text description. Most closely related to us is CosHand [70], which takes as input an RGB image of a hand-object interaction, a hand mask at the current state, and the hand mask of the future state, and generates an RGB image of the future state. Unlike us, they cannot generate post-interaction object dynamics. Importantly, none of these works study *dynamics*, generating instead discrete state transitions that fail to capture the nuanced, temporally coherent behaviors observed in interactions.

Synthesizing causal physical relations from visual input. A growing body of work aims to model and predict physical causal effects from visual inputs such as images or videos. For example, research in intuitive physics seeks to replicate the human-like, non-mathematical understanding of physical events, e.g. by predicting future frames given an input video. Early works like [22, 48] train neural networks to assess the stability of block towers, while [24] leverage prior physical knowledge formalized through partial differential equations (PDEs). Other approaches investigate counterfactual reasoning by leveraging graph neural networks [2, 39]. Wu et al. [82–84] explore the use of an inverse rendering approach, extracting geometry and physical properties from the video which are then coupled with a physics simulator and a rendering engine to generate the future frames. Other works [81] incorporate Interaction Networks [3] to approximate physical systems from video data. These approaches are often limited to simplified, synthetic datasets and struggle to generalize to real-world scenarios.

Recent methods have started to combine language models with physical engines. Liu et al. [51] ground a large language model using a computational physics engine while Gao et al. [19] show that fine-tuning a vision-language model (VLM) on annotated datasets of physical concepts improves its understanding of physical interactions. Closely related to our work is PhysGen [53], which trains an image-to-video model that conditions the video generation on physics parameters (e.g., force or torque). However, the model relies on a dynamics simulator to generate motion, and its application is limited to rigid objects. A related but tangential line of work focuses on identifying and generating the effects of objects on their surroundings. For example, Omnimatte [56] introduces the problem of identifying all parts of a scene influenced by an object, given a video and a mask of the object. Similarly, Lu et al. [55] propose to re-time the motion of different subjects in a scene while maintaining realistic interactions with the environment. ActAnywhere [62] generates videos with plausible human-scene interactions, taking a masked video of a person and a background image as input. These works address the problem of synthesizing realistic interactions within a scene, however, lack fine-grained control.

Controllable video generation. Video generation has advanced significantly in recent years, with diffusion models leading to substantial improvements in unconditional [33, 99], text-based [1, 5, 6, 12, 20, 23, 26, 32, 37, 68, 76, 85, 89, 92, 104] and image-based [1, 5, 20, 25, 78, 89] generation. These advances have raised the question of how to incorporate more nuanced control into video generation. Some text-to-video approaches are trained by "inflating" text-to-image (T2V) models [8, 13, 25, 26, 38, 85], and can thus be integrated with conditional T2V models such as ControlNet [100] or T2V-Adapter [60]. Control can also be achieved by conditioning on trajectories [58, 87, 98] or bounding-boxes [77], fine-tuning on appropriate datasets. VideoComposer [78] incorporates multiple condition types, including text, depth, style, and temporal conditions via motion vectors. Camera motion control has also been explored, with AnimateDiff [79] employing LoRA [34] modules to control camera movement, while MotionCtrl [80] and CameraCtrl [29] directly embed the camera information for more precise control. Additionally, several works target human animation from a pose control signal, such as DreamPose [42], MagicPose [91], and AnimateAnyone [35], but do not account for interactions.

3. Controllable Interactive Dynamics

Video diffusion models such as [5, 54] have demonstrated impressive performance in generating videos from text or images, and have even shown potential in tasks that require 3D understanding when properly fine-tuned [36, 75]. Trained on millions of videos, we hypothesize that these models also possess implicit knowledge of complex interactive dynamics, such as those that appear when humans interact with objects. Out of the box, however, they lack a precise control mechanism, often relying solely on textual inputs or requiring careful selection of the starting frame.

Task. Given an input image, $\boldsymbol{x} \in \mathbb{R}^{1\times H\times W\times 3}$, and a *driving motion* in the form of a pixel-wise corresponding control signal $\boldsymbol{c} \in \mathbb{R}^{N\times H\times W\times 3}$, we task InterDyn with generating a video sequence, $\boldsymbol{y} \in \mathbb{R}^{N\times H\times W\times 3}$, depicting plausible object dynamics. Through this task, we aim to learn the conditional distribution between a driving motion, such as that of a human hand, and the consequent motion of manipulated objects. In other words, the model needs to synthesize plausible object movement and appearance *without any indication* other than the driving motion, while maintaining physical and visual consistency with the input image.

Stable Video Diffusion. We extend Stable Video Diffusion [5] (SVD) to enable controllable interactive dynamics and explore the versatility of this model across a range of scenarios. SVD is a publicly available U-Net-based latent diffusion model [66] that extends Stable Diffusion 2.1 to video generation by interleaving the network with temporal layers. Given a static input image of a scene, SVD denoises a sequence of N frames $\boldsymbol{y} \in \mathbb{R}^{N\times H\times W\times 3}$ to generate a video that follows the initial frame. The input image is fed into the denoising U-Net by concatenating its latent to each of the frames in the noised input, and by supplying its CLIP [64] embedding to the U-Net's cross-attention layers. In addition, SVD is conditioned on the video's FPS and motion ID, where the motion ID represents the amount of motion in the video. We found a motion ID of 40 to align well with our frozen SVD prior.

Control. InterDyn extends SVD with an additional control signal $\boldsymbol{c} \in \mathbb{R}^{N\times H\times W\times 3}$ by integrating a ControlNet-like branch [100]. An overview of our pipeline is presented in Fig. 3. The SVD weights remain frozen to preserve its learned dynamics prior. Following [100], we introduce a trainable copy of the SVD encoder E, connected to SVD's frozen decoder via skip connections, and modulated by zero-initialized convolutions. We use a small CNN, $\mathcal{E}(\cdot)$, to encode the control signal $\boldsymbol{c}$ into the latent space, which is then added to the noisy input latent that is passed to the ControlNet encoder. Similar to SVD, the control branch interleaves convolutional, spatial, and temporal blocks, enabling InterDyn to process the control signal in a temporal-aware manner. This helps InterDyn to be robust to noisy control signals, see Fig. 6. We opt for binary masks as conditioning signal due to their accessibility. However, our method can be extended to incorporate diverse types of signals. We find that for hand-object interactions, the type of conditioning signal does not significantly impact performance, see Appendix A and Tab. S1.

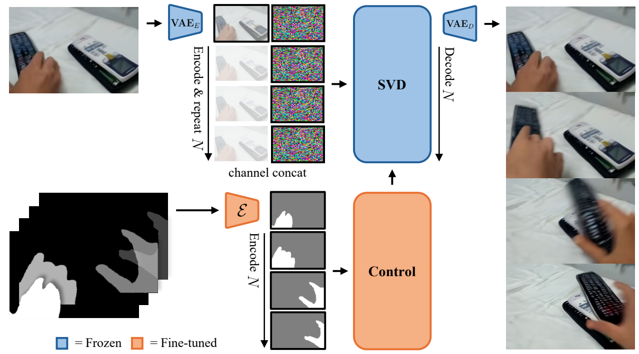
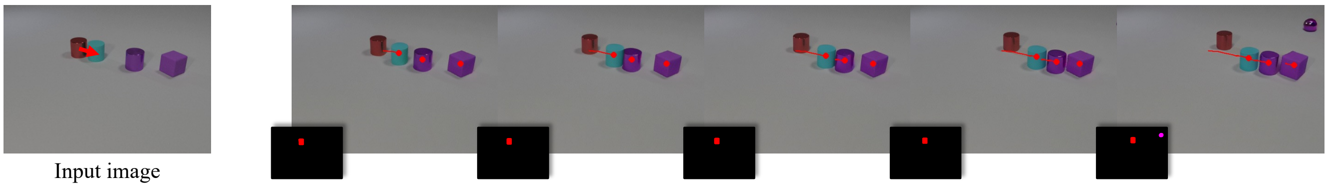


Figure 3. **Overview of InterDyn.** Given an input image depicting a scene, such as a hand holding a remote, and a "driving motion," such as a sequence of binary hand masks, InterDyn generates a video depicting plausible hand and object dynamics. Crucially, InterDyn receives no control signal for the object. Through this setup, we probe and assess the implicit knowledge of large video generation models on complex interactive dynamics. We use Stable Video Diffusion (SVD) as our frozen backbone and fine-tune a separate control signal encoder. Videos are iteratively denoised over T timesteps, starting from Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$.

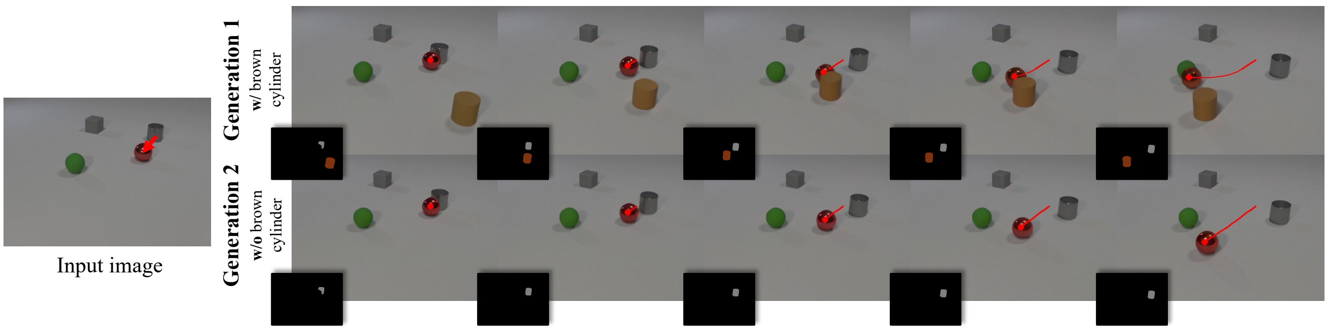
Inference. During inference, we start from an input image, control signal sequence, and randomly sampled Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. Through iteratively applying InterDyn over T denoising timesteps, we generate a video $\boldsymbol{y}$ depicting plausible hand and object dynamics, aligned with the control signal.

4. Experiments

The primary goal of this work is to synthesize scene-level interactive dynamics by leveraging the implicit physical understanding of a pre-trained video generative model. We begin by probing the model's ability to predict physically plausible outcomes within simulated environments, specifically using the CLEVRER dataset [97]. We test force propagation amongst uncontrolled objects and examine counterfactual future prediction by generating videos for one input image with different control signals. Motivated by promising results, we extend our investigation to complex, real-world hand-object interaction scenarios using the Something-Something-v2 (SSV2) dataset [21], conducting comprehensive comparisons with existing baselines that pursue similar objectives. Additionally, we showcase diverse physical examples to demonstrate the capabilities of InterDyn in generating realistic interactive dynamics.



(a) **Force Propagation.** InterDyn generates force propagation to and amongst uncontrolled objects (cyan cylinder, pink cylinder, and pink cube).



(b) **Counterfactual Dynamics.** Two generations of InterDyn on the same input image: (1) w/ brown-cylinder control and (2) w/o brown-cylinder control.

Figure 4. **Qualitative investigation on the CLEVRER dataset.** Given an input image and the "driving" motion of one or two objects, our model predicts the future interactive dynamics of multiple elements in the scene. The driving motion is given in the form of semantic mask sequences. The generated object motions are highlighted with a red-line trajectory. Note that our model can generate videos with force propagation across multiple uncontrolled objects (top) and can generate multiple futures (bottom). **Zoom in** for details.

4.1. Implementation

We initialize InterDyn with the 14-frame image-to-video weights of SVD [5]. During training, we use the Adam optimizer [45] with a learning rate of 1×10^{-5}. We use the EDM framework [43] with noise distribution $\log \sigma \sim \mathcal{N}(0.7, 1.6^2)$. We train on two 80GB H100 GPUs, with a per-GPU batch size of 4. Video length and FPS define the temporal resolution of dynamics; to balance short- and long-range events we subsample videos to 7 FPS. To facilitate classifier-free guidance [31], we randomly drop the input image with a probability of 5%. At inference, we apply the Euler scheduler [43] over 50 denoising timesteps.

4.2. Metrics

We evaluate InterDyn on image quality, spatio-temporal similarity, and motion fidelity. Image quality metrics are computed frame-wise. All metrics are reported excluding the first frame, as it serves as input conditioning.

Image quality. We report the Structural Similarity Index Measure (SSIM) [79], Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [102], Fréchet Inception Distance (FID) [30] and unbiased Kernel Inception Distance (KID) [4].

Spatio-temporal similarity. To assess the spatio-temporal perceptual similarity between the ground truth and the generated video distributions, we use the Fréchet Video Distance (FVD) and unbiased Kernel Video Distance (KVD) proposed in [74]. We use the implementation of [69].

Motion Fidelity. Through InterDyn, we do not have explicit control over object dynamics, which means that the pixel alignment of an object in the generated and ground truth video is only guaranteed in the starting frame. In this case, comparing generated object motion to the ground truth naively might misrepresent the true quality of object motion over time. Therefore, we adapt the Motion Fidelity (MF) metric proposed by Yatim et al. [93], which measures the similarity between point-tracking trajectories.

To compute the metric for any video, we obtain a mask of the object in the starting frame, sample 100 points on the object, and track these throughout both the ground truth and generated video using CoTracker3 [41]. Given the resulting two sets of tracklets $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}, \tilde{\mathcal{T}} = \{\tilde{\tau}_1, \ldots, \tilde{\tau}_m\}$ the motion fidelity metric is defined as:

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

with the correlation between two tracklets $\mathbf{corr}(\tau, \tilde{\tau})$ [50]:

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

where $(v_k^x, v_k^y), (\tilde{v}_k^x, \tilde{v}_k^y)$ are the k^{th} frame displacement of tracklets $\tau, \tilde{\tau}$ respectively. If there are less than 100 points to query on the object due to it being too small, we do not consider the video for the motion fidelity metric.

4.3. Probing Dynamics with Object Collision Events

Here, we fine-tune InterDyn on an object collision dataset to *probe* its ability to generate realistic object interactions. Qualitatively, we review whether InterDyn can produce plausible object motion for uncontrolled objects, given the motion of objects entering the scene. In addition, we examine whether InterDyn can generate counterfactual videos for the same input image, but different control signals.

Dataset. We use CLEVRER [97], which provides 20,000 videos of colliding objects with annotated segmentation masks and metadata on collision events. We construct a control signal for objects entering the scene and aim to use InterDyn to generate the motion of the objects that are already present, upon collision. Stationary objects do not receive any form of control signal. Colored masks help the model distinguish unique objects. The frames are cropped and scaled to 320×448, and we only sample input frames before collisions between objects in the scene, to maximize InterDyn's exposure to interactive dynamics.

Force propagation. InterDyn can generate force propagation dynamics between a controlled object and an uncontrolled object, as well as amongst uncontrolled objects, as illustrated in Fig. 4a. Here, the red cylinder at the top left is the driving force. It collides with the uncontrolled blue cylinder, which then collides with the uncontrolled purple cylinder, in turn striking the uncontrolled purple cube on the far right. Point-tracking trajectories display how collisions alter each object's position. This suggests that InterDyn possesses an implicit understanding of physical interactions, enabling it to generate plausible dynamics.

Counterfactual dynamics. By altering the control signal, InterDyn can simulate counterfactual scenarios for the same input image, as shown in Fig. 4b. In "Generation 1", the gray cylinder (controlled) collides with the stationary red sphere (uncontrolled), causing it to move; it is later struck by the brown cylinder (controlled), altering its path once again. In "Generation 2", removing the brown cylinder lets the red sphere continue along its original trajectory, consistent with expectations. Crucially, there is no control signal for the red sphere throughout the sequence; its motion and trajectory are entirely generated by InterDyn.

Probing InterDyn on CLEVRER highlights its ability to generate interactive dynamics for objects within a simple synthetic environment. We provide additional results in video format on our webpage.

4.4. Generating Human-Object Interactions

For this experiment, we fine-tune InterDyn on a human-action dataset, focused on hand-object interaction. We encode human movement as a sequence of pixel-aligned binary hand masks and task InterDyn to generate a video with hand movements and corresponding object dynamics.

Dataset. We use Something-Something-v2 (SSV2), which provides 220,847 videos of humans performing basic actions with everyday objects. It contains actions such as "pushing [something] from left to right", "squeezing [something]" and "lifting [something] with [something] on it". This dataset allows us to train InterDyn at a larger scale and compare with our baseline CosHand [70]. We train one version of InterDyn at the same resolution as CosHand, 256×256, and a second version at 256×384, which aligns better with the dynamic prior of SVD.

We generate binary hand masks by prompting Segment Anything 2 (SAM2) [65] with hand bounding boxes, provided by the Something-Else dataset [59]. Similar to our baseline CosHand, we exclude the "pretending" class. Additionally, we remove all "[something] is done behind [something]" classes since the object would be out of view of the camera. We remove videos smaller than the target resolution/length, ensure the hand and objects are continuously visible, and crop videos larger than the target resolution while centering the object. We include videos without obvious motion and state transition, such as "holding [something]". Since bounding box annotations are only provided for the train split (79,043 samples for 256×384 and 104,260 for 256×256) and validation split (8,667 samples for 256×384 and 11,229 for 256×256), we report all evaluation metrics on the validation split.

Figure 5. **Qualitative comparison.** A two-state approach such as CosHand [70] struggles with post-interaction object dynamics.

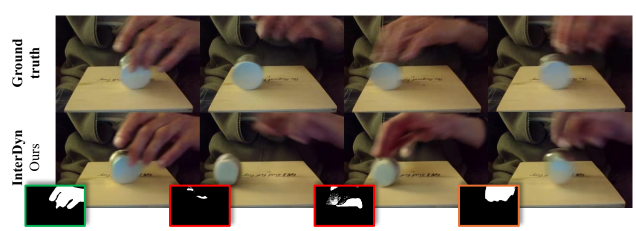


Figure 6. **Robustness to noise.** SAM2 outputs noisy/coarse masks for frames with considerable motion blur (orange/red). Despite this, InterDyn can generate plausible hand and object dynamics.

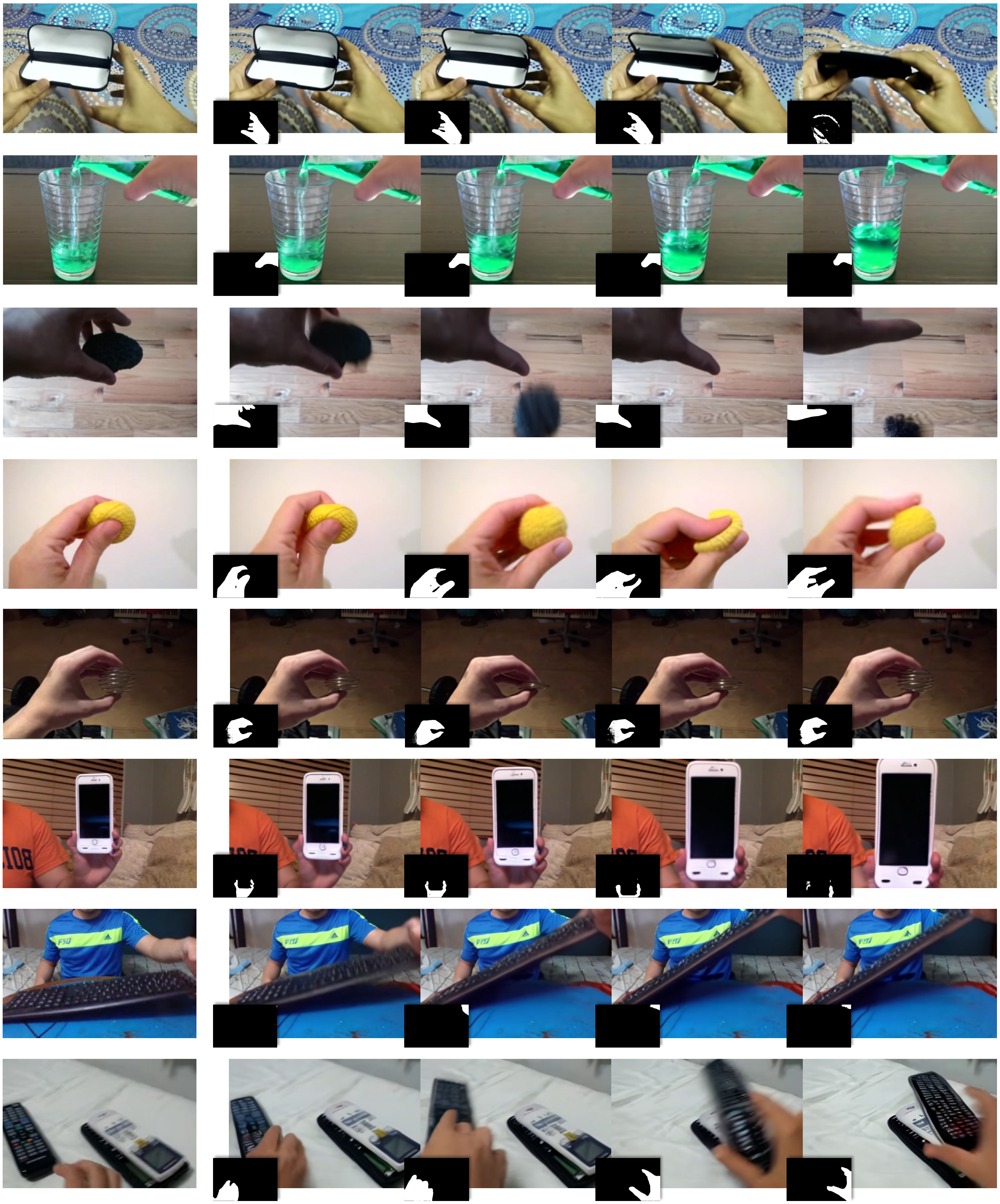


Figure 7. **Diverse generation of interactive dynamics.** We show multiple challenging examples, such as (from top to bottom): interacting with articulated objects, pouring liquid, letting an object fall, squeezing a highly deformable or "collapsible" object, interacting with reflective objects, tilting a ridged object, or stacking objects. 🔍 **Zoom in** for details.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓	KVD ↓	Motion Fidelity ↑ [93]
Seer [23]	0.418	10.71	0.588	33.35	0.030	287.46	81.31	—
DynamiCrafter [89]†	—	—	—	17.48	0.014	204.11	31.81	—
CosHand-Independent [70]	0.615	16.87	0.313	**2.95**	**0.002**	91.18	19.24	0.432
CosHand-Autoregressive [70]	0.531	14.92	0.408	12.66	0.012	90.30	13.68	0.570
Ours 256×256	<u>0.664</u>	<u>18.60</u>	<u>0.260</u>	<u>4.95</u>	<u>0.004</u>	**19.27**	**1.99**	<u>0.633</u>
Ours 256×384	**0.680**	**19.04**	**0.252**	5.07	<u>0.004</u>	<u>22.22</u>	<u>2.09</u>	**0.641**

Table 1. **Quantitative comparison on Something-Something-v2.** We compare against two language-instructed video generation methods, Seer [23] and DynamiCrafter [89] and two video extensions of our baseline CosHand [70]. We report results for InterDyn at two resolutions: 256×256 (matching CosHand) and 256×384 (matching SVD's prior). Methods denoted with † do not use SSV2 as their training dataset.

Baselines. To generate videos using CosHand [70] rather than state transitions, we run CosHand in a frame-by-frame approach (CosHand-Independent) and an auto-regressive approach (CosHand-Autoregressive). In the frame-by-frame variant, each future frame is independently predicted from the initial frame and its corresponding mask:

$$\hat{x}_{t+1} = \text{CosHand}(x_0, h_0, h_{t+1}), \quad \forall t \in [0, 13],$$

where h_0 and h_{t+1} denote masks, and x_0 the initial frame. In the auto-regressive variant, we use CosHand to generate video frames sequentially:

$$\hat{x}_{t+1} = \text{CosHand}(\hat{x}_t, h_t, h_{t+1}), \quad \forall t \in [0, 13],$$

with $\hat{x}_t$ being the previously generated frame and $\hat{x}_0 = x_0$.

We also compare against two language-instructed video generation methods: Seer [23] and DynamiCrafter [89], which we prompt with a first frame and its corresponding SSV2 class label as instruction. Note that DynamiCrafter was trained with a random video frame as conditioning, i.e. not always the first frame. At inference, its generated videos are thus not strictly a continuation of the conditioning, which precludes frame-aligned metrics. We compare InterDyn against these baselines in Tab. 1.

Quantitative analysis. The frame-by-frame variant of CosHand achieves high image quality but struggles with temporal coherence and motion fidelity. Qualitatively, we notice that object locations are inconsistent across frames. In contrast, the auto-regressive variant improves object motion fidelity but suffers from lower frame-wise image quality due to error propagation. Both variants fail in scenarios requiring accurate post-interaction dynamics, such as when objects continue moving after being released from direct hand contact, as shown in Fig. 5.

Our method, InterDyn (256×384), achieves the best overall performance, surpassing our baseline CosHand in spatio-temporal dynamics, motion fidelity, and all but two image quality metrics. We hypothesize that this might be due to two reasons (1) we use a frozen U-Net, while CosHand fine-tunes its model on SSV2, so CosHand might generate frames closer to the SSV2 distribution, and 2) when SVD was trained, it was initialized with SD weights as spatial layers, and then fine-tuned over multiple stages; this might have degraded its spatial prior, and by extension the quality of produced images compared to CosHand.

Fine-tuned on noisy masks and leveraging its temporal-aware control branch, InterDyn can interpret a noisy control sequence; e.g. when SAM2 produces a coarse and noisy hand mask sequence, InterDyn generates detailed hands including individual fingers, see Fig. 6. Though not always consistent, InterDyn is capable of depicting post-interaction dynamics, such as rolling or sliding objects.

Qualitative analysis. We present diverse qualitative results generated by InterDyn in Fig. 7. Row 1 shows how InterDyn generates the articulated motion of an object. Row 2 showcases pouring water into a glass; note how the water level increases over time. Row 3 demonstrates an object being dropped, moving out of frame when falling, and rolling back in frame once hitting the floor, featuring realistic motion blur synthesis. Rows 4 and 5 illustrate how InterDyn handles squeezing interactions—the rubber and the spring are compressed and restored accordingly. Row 6 demonstrates an understanding of physical size and distance to the camera, as the phone moves closer to the viewer. These results highlight the complexity that InterDyn can handle, implying its generalization ability and future potential as an implicit, yet generalized physical simulator and renderer.

5. Conclusion

We introduced InterDyn, a framework that generates videos of interactive dynamics using large video generation models as implicit physics simulators. By incorporating an interactive control mechanism, InterDyn produces plausible, temporally consistent videos of object interactions—including complex human-object interactions—while generalizing to unseen objects. Our evaluations demonstrate that InterDyn effectively captures continuous motion and subsequent dynamics, outperforming baselines that focus on single future states. This work highlights the potential of using video generative models as physics simulators without explicit reconstruction, opening new avenues for future research.

Acknowledgements: We thank Benjamin Pellkofer for IT support, Peter Kulits for discussions, Tomasz Niewiadomski for coding support, and Allan D. Jepson, Artur Grigorev, Soubhik Sanyal, Omid Taheri & Angjoo Kanazawa for feedback. DT is supported by the ERC Starting Grant (project STRIPES, 101165317).

Disclosure: https://files.is.tue.mpg.de/black/CoI_CVPR_2025.txt DT has received a research gift fund from Google.

References

[1] 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 *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 1–11, 2024. 2, 3

[2] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3

[3] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016. 3

[4] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[5] 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:2311.15127*, 2023. 2, 3, 4, 5

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 3

[7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 14

[8] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4D-guided video generation with 2D diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7611–7620, 2024. 3

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 14

[10] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1577–1585, 2024. 3

[11] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 14

[12] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv:2310.19512*, 2023. 3

[13] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-A-Video: Controllable text-to-video generation with diffusion models. *arXiv:2305.13840*, 2023. 3

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5031–5041, 2020. 3

[15] Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 19888–19901, 2024. 3

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 2

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023. 3

[18] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 494–504, 2024. 3

[19] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *International Conference on Robotics and Automation (ICRA)*, pages 12462–12469, 2024. 3

[20] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu Video: Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The something something video database for learning and evaluating visual common sense. In *International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017. 2, 4

[22] Oliver Groth, Fabian B. Fuchs, Ingmar Posner, and Andrea Vedaldi. ShapeStacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision (ECCV)*, pages 702–717, 2018. 3

[23] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 8, 17

[24] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11471–11481, 2020. 2, 3

[25] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2V-Adapter: A general image-to-video adapter for diffusion models. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2024. 3

[26] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv:2307.04725*, 2023. 3

[27] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 3

[28] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *International Conference on 3D Vision (3DV)*, pages 659–668, 2021. 3

[29] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv:2404.02101*, 2024. 3

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6629–6640, 2017. 5

[31] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 5

[32] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, et al. Imagen Video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022. 3

[33] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8633–8646, 2022. 3

[34] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[35] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8153–8163, 2024. 3

[36] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. *arXiv:2409.02095*, 2024. 4

[37] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make It Move: Controllable image-to-video generation with text descriptions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18219–18228, 2022. 3

[38] Zhihao Hu and Dong Xu. VideoControlNet: A motion-guided video-to-video translation framework by using diffusion model with ControlNet. *arXiv:2307.14073*, 2023. 3

[39] Steeven Janny, Fabien Baradel, Natalia Neverova, Madiha Nadri, Greg Mori, and Christian Wolf. Filtered-CoPhy: Unsupervised learning of counterfactual physics in pixel space. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[40] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, pages 11107–11116, 2021. 3

[41] 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 *European Conference on Computer Vision (ECCV)*, pages 18–35, 2024. 5

[42] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. DreamPose: Fashion image-to-video synthesis via stable diffusion. In *International Conference on Computer Vision (ICCV)*, pages 22623–22633, 2023. 3

[43] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:26565–26577, 2022. 5

[44] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, et al. Beyond the Contact: Discovering comprehensive affordance for 3D objects from pre-trained 2D diffusion models. In *European Conference on Computer Vision (ECCV)*, pages 400–419, 2024. 3

[45] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[46] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning (ICML)*, pages 2693–2702, 2018. 2

[47] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 17089–17099, 2023. 3

[48] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International Conference on Machine Learning (ICML)*, pages 430–438, 2016. 3

[49] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[50] Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand, and Edward H. Adelson. Motion magnification. *Transactions on Graphics (TOG)*, 24(3):519–526, 2005. 5

[51] Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's Eye: Grounded language model reasoning through simulation. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[52] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023. 14

[53] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, pages 360–378, 2024. 3

[54] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv:2402.17177*, 2024. 2, 4

[55] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T. Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *Transactions on Graphics (TOG)*, 39(6): 1–14, 2020. 3

[56] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T. Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4507–4515, 2021. 3

[57] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. UGG: Unified generative grasping. In *European Conference on Computer Vision (ECCV)*, pages 414–433, 2024. 3

[58] Wan-Duo Kurt Ma, John P Lewis, and W. Bastiaan Kleijn. TrailBlazer: Trajectory control for diffusion-based video generation. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 1–11, 2023. 3

[59] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-Else: Compositional action recognition with spatial-temporal interaction networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1059, 2020. 6

[60] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3

[61] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *Transactions on Graphics (TOG)*, 38(4): 1–13, 2019. 14

[62] Boxiao Pan, Zhan Xu, Chun-Hao Paul Huang, Krishna Kumar Singh, Yang Zhou, Leonidas J. Guibas, and Jimei Yang. ActAnywhere: Subject-aware video background generation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3

[63] Georgios Paschalidis, Romana Wilschut, Dimitrije Antić, Omid Taheri, and Dimitrios Tzionas. 3D whole-body grasp synthesis with directional controllability. In *International Conference on 3D Vision (3DV)*, 2025. 3

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 4

[65] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 6, 14

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 4, 14

[67] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6), 2022. 14

[68] 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 *International Conference on Learning Representations (ICLR)*, 2023. 3

[69] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3626–3636, 2022. 5

[70] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *European Conference on Computer Vision (ECCV)*, pages 414–430, 2024. 2, 3, 6, 8, 14, 17

[71] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, pages 581–600, 2020. 3

[72] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. GRIP: Generating interaction poses using latent consistency and spatial cues. In *International Conference on 3D Vision (3DV)*, pages 933–943, 2024. 3

[73] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1674, 2022. 3

[74] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *International Conference on Learning Representations Workshops (ICLRw)*, 2019. 5

[75] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, pages 439–457, 2025. 4, 14

[76] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv:2308.06571*, 2023. 3

[77] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 3

[78] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3

[79] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From error visibility to structural similarity. *Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 3, 5

[80] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *Transactions on Graphics (TOG)*, pages 1–11, 2024. 3

[81] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual Interaction Networks: Learning a physics simulator from video. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 4539–4547, 2017. 3

[82] Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 127–135, 2015. 3

[83] Jiajun Wu, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *British Machine Vision Conference (BMVC)*, 2016.

[84] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 153–164, 2017. 2, 3

[85] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 3

[86] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. THOR: Text to human-object interaction diffusion via relation intervention. *arXiv:2403.11208*, 2024. 3

[87] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion control for anything using entity representation. In *European Conference on Computer Vision (ECCV)*, pages 331–348, 2025. 3

[88] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10003–10015, 2024. 3

[89] 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 (ECCV)*, pages 399–417, 2025. 2, 3, 8, 17

[90] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In *International Conference on Computer Vision (ICCV)*, pages 14928–14940, 2023. 3

[91] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1490, 2024. 3

[92] 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. *arXiv:2408.06072*, 2024. 3

[93] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8466–8476, 2024. 5, 8, 14, 17

[94] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *International Conference on Computer Vision (ICCV)*, pages 19717–19728, 2023. 3

[95] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance Diffusion: Synthesizing hand-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 22479–22489, 2023. 3

[96] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1911–1920, 2024. 3

[97] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 4, 6

[98] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv:2308.08089*, 2023. 3

[99] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18456–18466, 2023. 3

[100] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 3, 4

[101] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. HOIDiffusion: Generating realistic 3D hand-object interaction data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8521–8531, 2024. 3

[102] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 5

[103] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 585–594, 2023. 3

[104] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient video generation with latent diffusion models. *arXiv:2211.11018*, 2022. 3

[105] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. GEARS: Local geometry-aware hand-object interaction synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20634–20643, 2024. 3

[106] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *International Conference on Computer Vision (ICCV)*, pages 15741–15751, 2021. 3