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
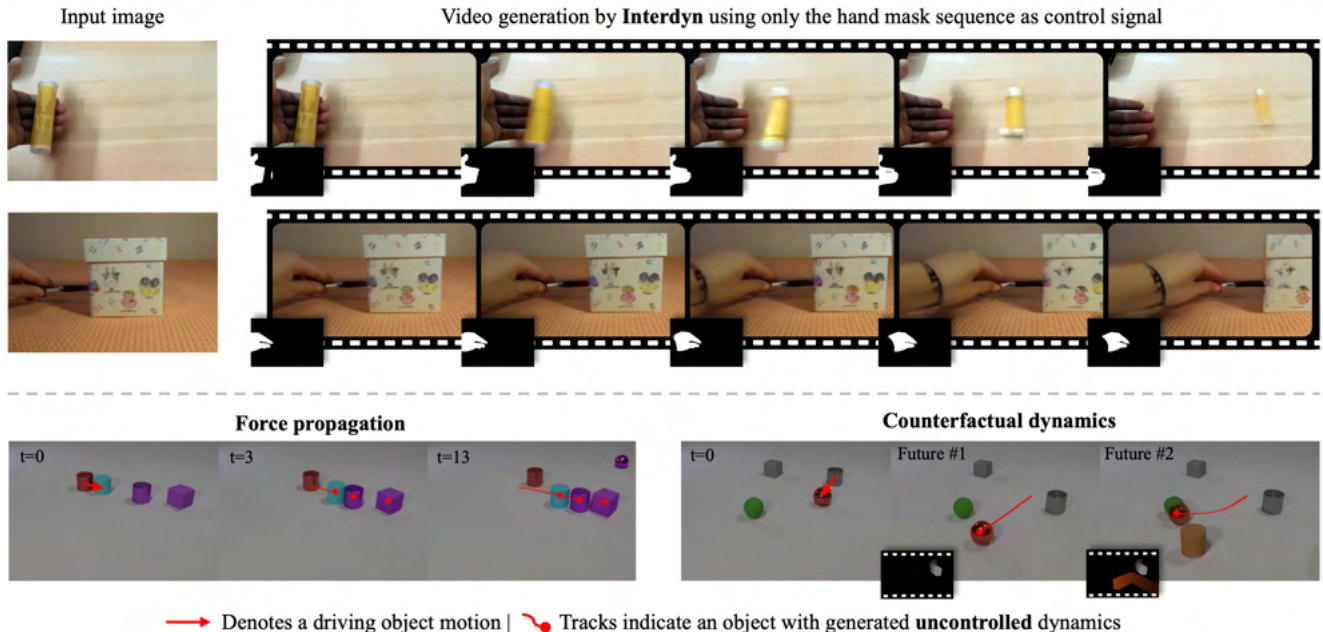[2] University of Amsterdam, the Netherlands

Figure 1. **InterDyn.** Given an image and a "driving motion" in the form of a mask sequence, our model generates scene dynamics that are physically plausible without 3D reconstruction and physics simulation. We investigate the generated "interactive dynamics" under simulated settings and in-the-wild HOI scenarios.

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 motion and subsequent 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 foundation models can act as both neural renderers and implicit physics "simulators" by learning 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

1

1. Introduction

Humans have the remarkable ability to intuitively predict the future dynamics of observed systems. 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 [81] 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 in the reconstruction, and may not generalize well. More recent methods [2, 22, 36, 43, 46] 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, 4, 15, 50, 62] opens new avenues for synthesizing interactions under complex scenarios. For example, Sudhakar et al. [66] recently proposed CosHand, a controllable image-to-image model based on Stable Diffusion [62] 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 –see e.g. 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

Figure 2. **State transition vs. dynamics.** Methods that generate static state transitions (i.e. predict a future image) such as CosHand [66] 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.

model realistic object dynamics. In particular, we propose InterDyn, a novel framework for synthesizing controllable dynamic interactions that leverages the physical and dynamics "knowledge" of a large video foundation model [4]. Unlike prior approaches that rely on explicit physical simulation [81] or are constrained to static state transitions [66], we leverage video foundation models to generate dynamic processes implicitly. Specifically, we extend Stable Video Diffusion (SVD) [4] with a dynamic control branch and fine-tune it on diverse scenes, enabling the 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 [94]. 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 have shown InterDyn's ability for counterfactual future prediction and physical force propagation.

Further, we evaluate how the system performs in a highly complex, 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 [20] and compare it with the state-of-the-art baseline CosHand [66] as well as other recent baselines. In both scenarios, 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 dataset [20]. Our experiments also demonstrate diverse physical plausible generations of interactive dynamics, probing into SVD's "understanding" of physics and dynamics.

In summary, we present InterDyn, the first 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 foundation 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 interactions have been widely studied within the context of 3D reconstruction [16, 17, 25, 26, 69, 85, 91], 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 [37, 41, 67, 104] or dynamic [59, 68, 101, 103] hand poses conditioned on 3D objects, whole-body interactions [87], or more recently, hand-object meshes given textual descriptions [9, 14, 53, 83, 93]. Few works address HOI synthesis in the 2D domain. GANHand [13] predicts 3D hand shape and pose given an RGB image of an object, while AffordanceDiffusion [92] estimates a 2D hand instead, by leveraging a diffusion model. Kulal et al. [44] 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 [99], 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 [66], which takes as input an RGB image of a hand-object interaction, hand mask at the current state, and hand mask of the future state, and generates an RGB image of the future state. Unlike us, they cannot generate post-interaction object dynamics and are more heavily dependent on hand mask quality. 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 [21, 45] train neural networks to assess the stability of block towers, while [22] leverage prior physical knowledge formalized through partial differential equations (PDEs). Other approaches investigate counterfactual reasoning by leveraging graph neural networks [2, 36]. Wu et al. [79–81] 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 [78] 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. [48] ground a large language model using a computational physics engine while Gao et al. [18] 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 [49], 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 [52] 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. [51] propose to re-time the motion of different subjects in a scene while maintaining realistic interactions with the environment. ActAnywhere [58] 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 [30, 96], text-based [1, 4, 5, 11, 19, 24, 29, 34, 64, 73, 82, 89, 102] and image-based [1, 4, 19, 23, 75, 86] 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 [7, 12, 23, 24, 35, 82], and can thus be integrated with conditional T2V models such as ControlNet [98] or T2V-Adapter [56]. Control can also be achieved by conditioning on trajectories [54, 84, 95] or bounding-boxes [74], fine-tuning on appropriate datasets. VideoComposer [75] incorporates multiple condition types, including text, depth, style, and temporal conditions via motion vectors. Camera motion control has also been explored, with AnimateDiff [76] employing LoRA [31] modules to control camera movement, while MotionCtrl [77] and CameraCtrl [27] directly embed the camera information for more precise control. Additionally, several works target human animation from a pose control signal, such as DreamPose [39], MagicPose [88], and AnimateAnyone [32]; however, they do not account for interactions.

3. Controllable Interactive Dynamics

Video diffusion models such as [4, 50] 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 [33, 71]. 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.

In this work, we extend Stable Video Diffusion [4] (SVD) to enable controllable interactive dynamics and explore the versatility of these models across a range of scenarios. SVD is a publicly available latent diffusion model [62] that extends Stable Diffusion 2.1 to video generation by interleaving the network with temporal layers. Given a static input image of a scene as conditioning, SVD denoises a sequence of T frames $y \in \mathbb{R}^{T \times H \times W \times 3}$ to generate a video that follows the initial frame. The conditioning 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 [60] embedding to the U-Net's cross-attention layers. In addition, SVD is conditioned on the FPS and motion ID of the video, where the motion ID represents the amount of motion present in the sequence.

Given an input image, $x \in \mathbb{R}^{H \times W \times 3}$, and a *driving motion* in the form of a pixel-wise corresponding control signal $c \in \mathbb{R}^{T \times H \times W}$, we task InterDyn with generating a video sequence, $y \in \mathbb{R}^{T \times H \times W \times 3}$, 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 our task, 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.

InterDyn extends SVD with an additional control signal $c \in \mathbb{R}^{T \times H \times W \times 3}$ by integrating a ControlNet-like branch [97]. The SVD weights remain frozen to preserve its learned dynamics prior. Following [97], we introduce a trainable copy of the SVD network connected to the original one via zero-convolutions. We use a small CNN, $\mathcal{E}(\cdot)$, to encode the control signal c into the latent space, which is then added to the noisy input latent that is passed to the ControlNet encoder. Importantly, similar to SVD, the control branch interleaves convolutional, spatial, and temporal blocks, enabling InterDyn to be temporally-aware when processing the control signal. This allows InterDyn to be robust to noisy control signals, see Figure 6.

We set the motion ID to a constant 40 during training and inference. To facilitate classifier-free guidance [28], we randomly drop the input image with a probability of 5%. At inference, we start from Gaussian noise and use InterDyn to generate a video with plausible object dynamics. We choose masks as conditioning signals due to their accessibility. However, our method can potentially be extended to incorporate diverse types of signals. Our experiments demonstrate that the choice of conditioning signal does not significantly impact performance, see Appendix B.

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 [94]. We test counterfactual future predictions based on different interactions and generate dynamics resulting from force propagation. Motivated by these promising results, we extend our investigation to complex, real-world hand-object interaction scenarios using the Something-Something-v2 dataset [20], 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.

4.1. Training details

We initialize InterDyn with the 14-frame publicly available [72] image-to-video weights of SVD [4]. We use the Adam optimizer [42] with a learning rate of 1×10^{-5}. For all experiments, we use the EDM framework [40] with a noise distribution defined by $\log \sigma \sim \mathcal{N}(0.7, 1.6^2)$. We train on two NVIDIA H100 GPUs with a per-GPU batch size of 4. Dynamic events vary in duration; with brief actions such as dropping an object spanning just a couple of frames and longer actions such as moving an object spanning several seconds. Here, the context length and FPS together define the temporal resolution of dynamics. To strike a balance between capturing short- and long-range events, we subsample videos to 7 FPS for all datasets.

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. To assess per-frame image quality we report the Structural Similarity Index Measure (SSIM) [76], Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [100].
Spatio-temporal similarity. To assess the spatio-temporal perceptual similarity between the ground truth and the generated video distributions, we employ the Fréchet Video Distance (FVD) [70], as implemented in [65].
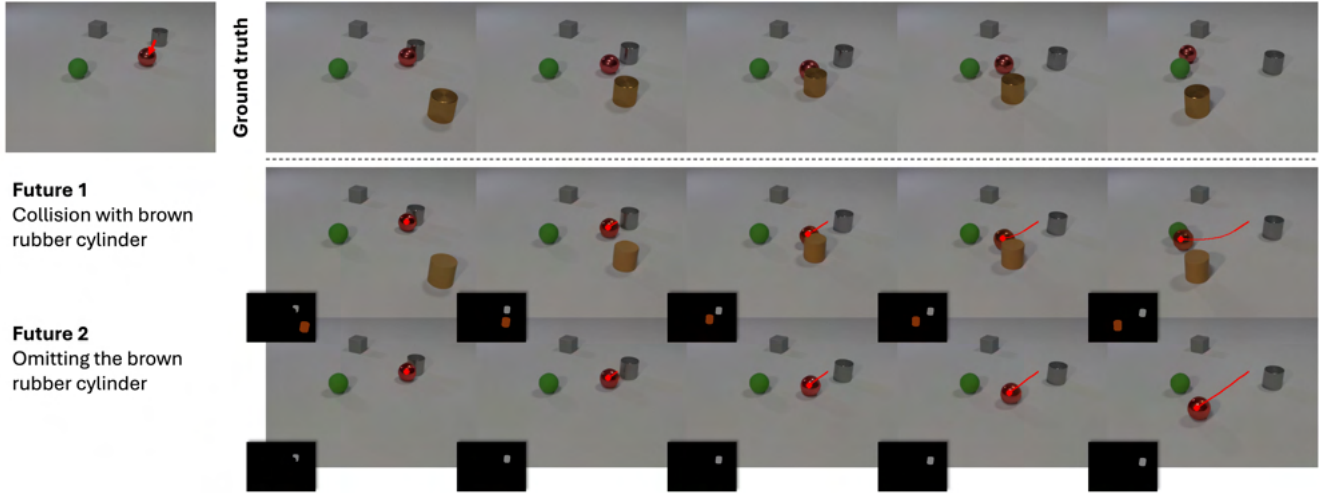
(a) **Counterfactual Dynamics.** Top: ground-truth collision sequence. Bottom: two different futures generated by InterDyn(1) w/ brown-rubber-cylinder and (2) w/o brown-rubber-cylinder. Note how the trajectory (shown as a red line) of the uncontrolled red-metal-sphere changes.
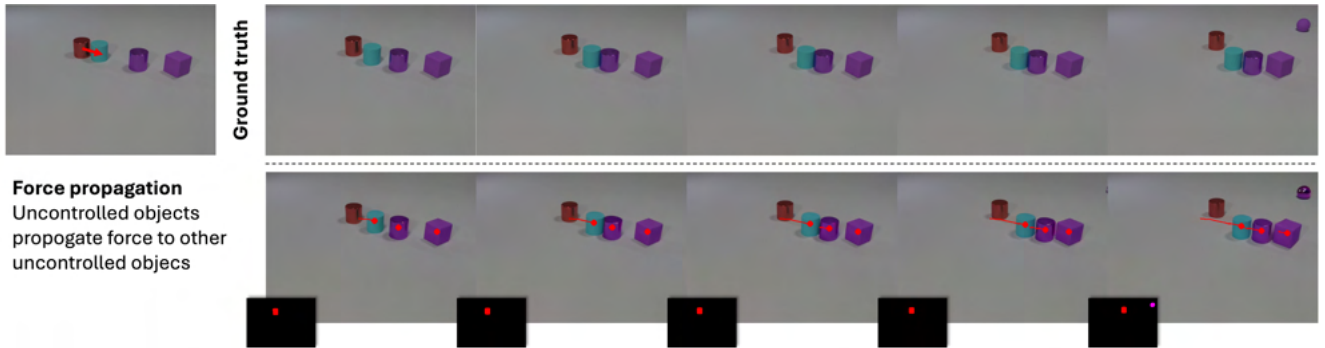
(b) **Force Propagation.** Top: ground-truth collision sequence. Bottom: InterDyn can generate plausible force propagation to and amongst uncontrolled objects (cyan-rubber-cylinder, pink-metal-cylinder, and pink-rubber-cube).

Figure 3. **Physical investigation on the CLEVRER dataset.** Given a 3D scene and the "driving" motion of one or two objects, our model predicts the future interaction dynamics of multiple elements in the scene. The driving motion is given in the form of semantic mask sequences. The predicted motions are highlighted with a red-line trajectory. Note that our model can predict multiple futures (top), or even force propagation across multiple uncontrolled objects (bottom). **Q Zoom in** for details.

Motion Fidelity. Since we do not explicitly control object dynamics, pixel alignment between the object in the ground truth and the generated video is only guaranteed in the starting frame. In light of this, we assess the quality of object motion by adapting the motion fidelity metric proposed by Yatim et al. [90], which measures the similarity between point-tracking trajectories.

Using the object bounding box from the Something-Else dataset as an input prompt to SAM2 [61], we obtain a mask of the object in the starting frame of the video. Using the mask, we sample 100 points on the object and track these throughout both the ground truth and generated video using CoTracker3 [38]. 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 measures the correlation their and is defined as:

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

where the correlation between two tracklets $\mathbf{corr}(\tau, \tilde{\tau})$ [47] is defined as:

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

Here $(v_k^x, v_k^y), (\tilde{v}_k^x, \tilde{v}_k^y)$ are the kth 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

We investigate the ability of SVD to generate interactive dynamics using the synthetic CLEVRER dataset [94], which includes 20,000 videos of colliding objects with annotated segmentation masks, 3D locations, and collision events. We sample videos where stationary objects interact with moving objects entering the scene. Segmentation masks are used to construct control signals for moving objects, and differently colored masks help the model distinguish unique objects. Frames are randomly cropped and scaled to 320×448, and starting frames are sampled before collisions to maximize exposure to dynamics. Our primary goal is to assess whether InterDyn can generate future frames that are physically plausible in response to introduced interactions. Specifically, we examine whether the model can produce meaningful object movements for uncontrolled objects in the scene, given the movement of objects entering the scene—for instance, determining how an object would move if struck by another.

Beyond direct interactions, we focus on two more challenging physical scenarios: **Counterfactual Dynamics**: We investigate whether the model can predict plausible motion trajectories involving multiple interactions at different time steps. This involves generating realistic motions under counterfactual scenarios, where different interaction configurations lead to different causal effects—much like a physics simulator. This tests how multiple interactions influence a single object. **Force Propagation**: We examine whether the model can generate subsequent object dynamics through force propagation. This means assessing whether objects beyond our direct control interact with each other in physically plausible ways. Here, we are interested in how one interaction influences multiple objects. We conduct both experiment settings on the CLEVRER test set.

As illustrated in the "Future 1" row in Fig. 3a, the gray cylinder (controlled object 1) collides with the stationary red sphere, causing it to move. Along its new trajectory, the red sphere is then struck by the brown cylinder (controlled object 2), altering its path once again. Compared to the top row, which shows the ground truth physics-simulated renderings, InterDyn synthesizes dynamics that are physically plausible without explicit knowledge of the objects' mass or stiffness. For the counterfactual experiment shown in the "Future 2" row, we remove the brown cylinder and observe how dynamics change after this. Without this second interaction, the red sphere continues along its original trajectory, as indicated by the red point-tracking path—consistent with physical expectations. Importantly, there is no control signal for the red sphere throughout the sequence; its motion is entirely generated by InterDyn.

InterDyn can generate force propagation dynamics, where uncontrolled objects interact with each other, as illustrated in Fig. 3b. The sole driving force is the red cylin-

der at the top left. This controlled object moves towards and collides with the uncontrolled blue cylinder, setting it into motion. The blue cylinder then collides with the uncontrolled purple cylinder, which subsequently impacts the uncontrolled purple cube on the far right. The point-tracking trajectories display the position changes of all three objects in a manner consistent with physical laws. This experiment suggests that InterDyn possesses an implicit understanding of physical interactions, enabling it to generate plausible dynamic sequences. We provide directions to (more) results in video format in Appendix A.

While the CLEVRER experiments highlight the fundamental capabilities of InterDyn in generating interactive dynamics, we are also interested in evaluating its performance in complex, real-world scenarios—the environments for which the video model was originally trained.

4.4. Generating Human-Object Interactions

We find such scenarios in the Something-Something-v2 dataset [20]. Originally proposed for human action recognition and video understanding, this dataset provides 220,847 videos of humans performing basic actions with everyday objects. It contains actions like "pushing [something] from left to right", "squeezing [something]" and "lifting [something] with [something] on it". The Something-Else dataset [55] provides bounding box annotations for 180,049 of these videos, which we use to extract hand masks with SAM2. This dataset allows us to train InterDyn at a larger scale and to compare with our closest competitor CosHand [66]. We train one version at the same resolution as CosHand, 256×256, and a second version at 256×384, which aligns better with the video resolution SVD has been trained on. We provide directions to (more) results in video format in Appendix A.

4.4.1. Baselines

We compare InterDyn with two CosHand variants: a frame-by-frame approach and an auto-regressive approach. The results can be found in Tab. 1. In the first approach, we predict each frame $\hat{x}_{t+1}$ in the sequence from the input frame and its corresponding mask:

$$\hat{x}_{t+1} = \text{CosHand}(x_0, h_0, h_{t+1}), \quad \forall t \in [1, 13], \quad (3)$$

where h_0, h_t denote a mask, and $\hat{x}_0$ is the initial frame. In the second approach, we use CosHand to generate videos auto-regressively in the following manner:

$$\hat{x}_{t+1} = \text{CosHand}(\hat{x}_t, h_t, h_{t+1}). \quad \forall t \in [1, 13], \quad (4)$$

where $\hat{x}_t$ is the generated frame from the previous time step. To avoid error propagation from the VAE decoding-encoding process, $\hat{x}_{t+1}$ remains in latent space.
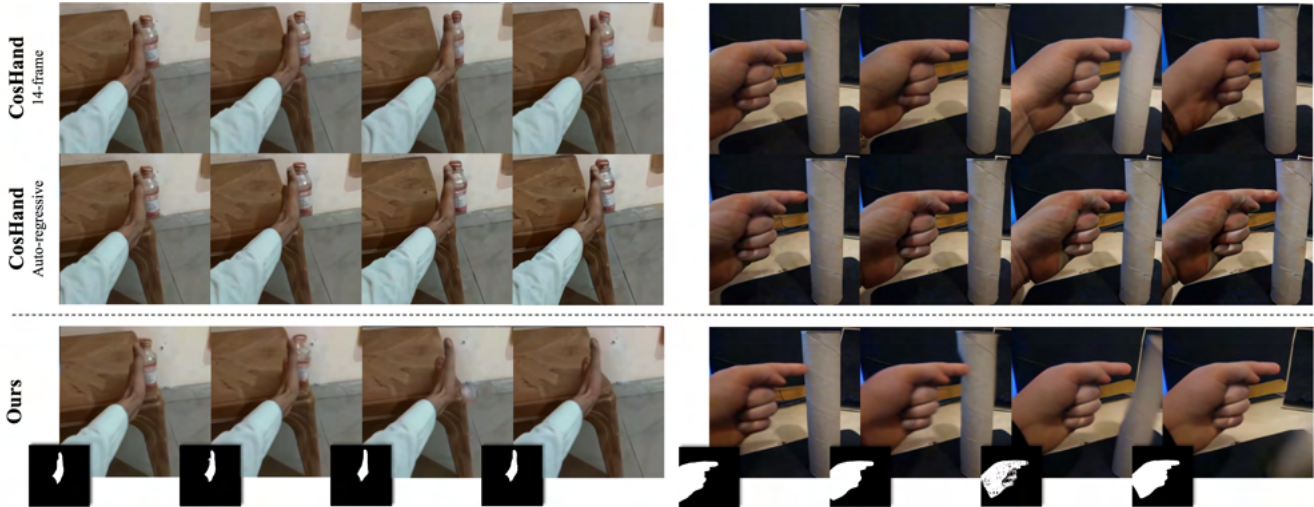
Figure 4. **Qualitative comparison with two temporal variants of CosHand [66].** A two-state approach such as CosHand struggles with post-interaction object dynamics. 🔍 **Zoom in** for details.

	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	Motion Fidelity ↑
CosHand 14-frame	0.615	16.87	0.313	91.18	0.432
CosHand Auto-regressive	0.531	14.92	0.408	90.30	0.570
Ours (256×256)	0.664	18.60	0.260	**19.27**	0.633
Ours (256×384)	**0.680**	**19.04**	**0.252**	22.22	**0.641**

Table 1. **Comparisons on the Something-Something-v2 dataset (video data).** We compare against two video extensions of CosHand: (1) by generating each frame from the initial image (`CosHand 14-frame`), (2) by auto-regressively generating the frames (`CosHand Auto-regressive`). All videos are generated with 14 frames. We provide two resolutions for InterDyn: 256×256 (same resolution as CosHand) and 256×384 (more compatible with SVD). We evaluate in terms of SSIM, PSNR, LPIPS, FVD, and a tracking-based metric (Motion Fidelity) [90].

Analysis We observe that CosHand in the frame-by-frame setting achieves high image quality, as indicated by the SSIM, PSNR, and LPIPS metrics, but struggles with temporal coherence and motion fidelity. Qualitatively, we notice that the object location may appear inconsistently across frames due to the lack of temporal context. In contrast, the auto-regressive version improves object motion fidelity over the frame-by-frame approach but suffers from lower framewise image quality due to error propagation. Importantly, it fails in scenarios requiring accurate post-interaction dynamics, such as when objects continue moving after being released from direct hand contact as shown in Fig. 4.

Our method, InterDyn (256×384), achieves the best overall performance, surpassing CosHand in image quality, spatio-temporal dynamics, and motion fidelity. It generates realistic post-interaction dynamics—such as rolling or sliding objects—and effectively handles noisy control signals. InterDyn (256×384) aligns better with the frozen U-Net prior, outperforming the InterDyn (256×256) variant.

Different from its two-states-based competitors, InterDyn interprets the control signals at the sequential motion level, rather than simply pixel-aligning with the provided hand masks. For instance, when SAM2 produces noisy mask sequences due to significant motion blur (as shown in Fig. 6), leading to incomplete or coarse hand shapes, InterDyn remains robust during inference. By leveraging its temporal prior and temporal control branch to handle noisy outliers, InterDyn can translate coarse and noisy masks into detailed hands –including individual fingers–, holistically reasoning about the **temporal control signal**. This sets it apart from methods that rely on precise input mask fidelity. We have conducted experiments on different variants for the control signal, see Appendix B.

Static comparison. For completeness, we also compare against CosHand [66] on a per-frame basis. For this, we extract the per-frame metrics for the second frame from InterDyn's and CosHand's video generation, and compare against the baselines evaluated in [66]. Quantitative results are shown in Tab. 2.

4.4.2. Diverse physical generation with InterDyn

We present several examples of the diverse interactive dynamics generated by InterDyn in Figure 5. 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
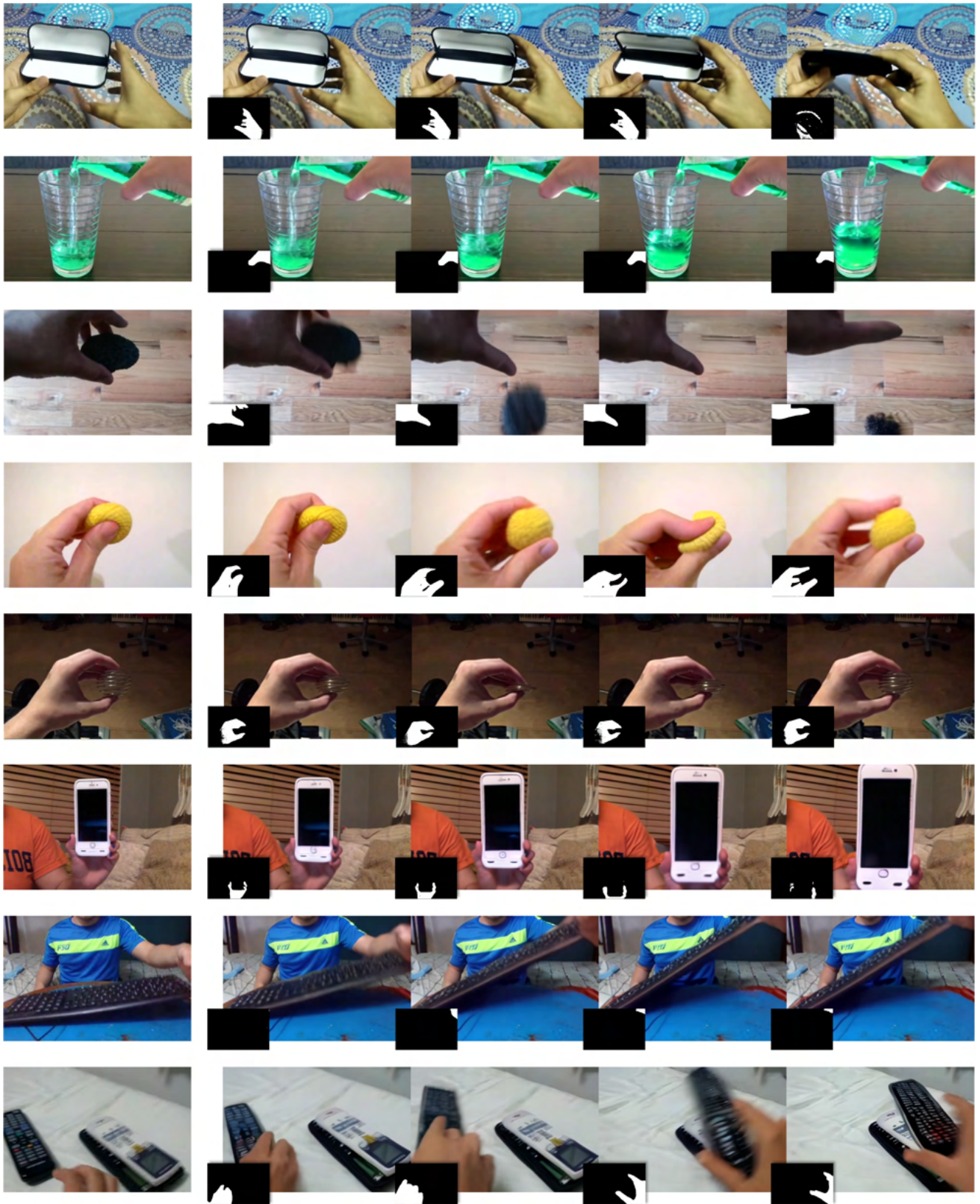
Figure 5. **Diverse generation of interactive dynamics.** We show multiple challenging examples, such as (from top to bottom): interacting with articulated objects, pouring liquid, letting an object fall, squeezing a highly deformable or "collapsible" object, interacting with reflective objects, tilting a ridged object, or stacking objects. 🔍 **Zoom in** for details.
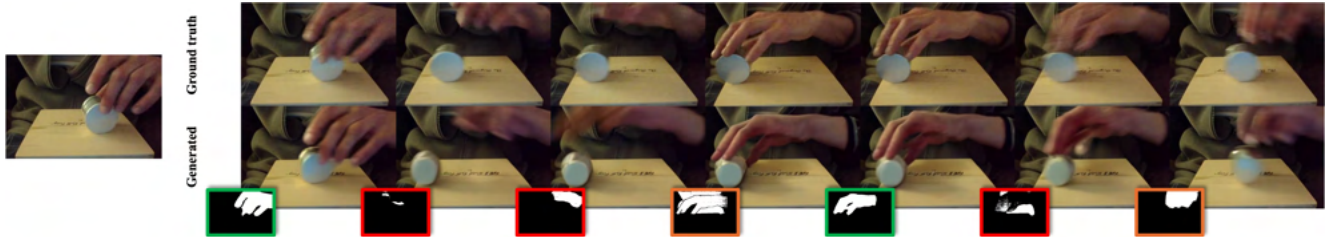
Figure 6. **Robustness to noise.** InterDyn can generate plausible hand and object dynamics (green) with incomplete or coarse control signals (red) extracted by SAM2 from frames with considerable motion blur. 🔍 **Zoom in** for details.

Method	SSIM ↑	PSNR ↑	LPIPS ↓
MCVD [66, 71]	0.231	8.75	0.307
UCG [62, 66]	0.340	12.08	0.124
IPix2Pix [6, 66]	0.289	9.53	0.296
TCG [62, 66]	0.234	9.05	0.221
CosHand [66]	0.414	13.72	**0.116**
CosHand (our test set)	0.698	20.55	0.194
Ours (256×256)	<u>0.785</u>	<u>23.93</u>	0.127
Ours (256×384)	**0.796**	**24.37**	<u>0.122</u>

Table 2. **Comparisons on the Something-Something-v2 dataset (static data).** We compare against CosHand and other static baselines on the task of generating a single future frame. The top five rows were taken from [66]. Since CosHand does not provide the exact validation split, we also evaluate their method on our own validation set. For a fair comparison, we train two methods: one with the same resolution as in CosHand (256×256), and another one on the native resolution of SVD (256×384).

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 is capable of handling, 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 by leveraging large video foundation 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 for physics simulation without explicit reconstruction, opening new avenues for future research.

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. *arXiv preprint arXiv:2401.12945*, 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. *Advances in neural information processing systems*, 29, 2016. 3

[4] 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] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 9

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7611–7620, 2024. 3

[8] 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. 1

[9] 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

[10] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 1

[11] 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

[12] 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 preprint arXiv:2305.13840*, 2023. 3

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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 *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024. 3

[19] 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

[20] 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, 6

[21] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the european conference on computer vision (eccv)*, pages 702–717, 2018. 3

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

[23] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, Haibin Huang, and Chongyang Ma. I2v-adapter: A general image-to-video adapter for diffusion models. In *SIGGRAPH 2024 Conference Papers*, 2024. 3

[24] 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

[25] Yana Hasson, Gul 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

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

[27] 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 preprint arXiv:2404.02101*, 2024. 3

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[29] 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 preprint arXiv:2210.02303*, 2022. 3

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:8633–8646, 2022. 3

[31] 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*, 2022. 3

[32] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[33] 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

[34] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 3

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

[36] 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

[37] 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

[38] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv:2410.11831*, 2024. 5

[39] 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

[40] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 4

[41] 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)*, 2024. 3

[42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[43] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. *arXiv:1802.04687*, 2018. 2

[44] 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

[45] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International conference on machine learning*, pages 430–438. PMLR, 2016. 3

[46] 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)*, 33, 2020. 2

[47] 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

[48] 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

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

[50] 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

[51] 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

[52] 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

[53] 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)*, 2024. 3

[54] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2023. 3

[55] 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

[56] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3

[57] 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 sin-

gle depth camera. *Transactions on Graphics (TOG)*, 38(4): 1–13, 2019. 1

[58] 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. *arXiv preprint arXiv:2401.10822*, 2024. 3

[59] Georgios Paschalidis, Romana Wilschut, Dimitrije Antić, Omid Taheri, and Dimitrios Tzionas. 3d wholebody grasp synthesis with directional controllability. *arXiv:2408.16770*, 2024. 3

[60] 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*, pages 8748–8763. PMLR, 2021. 4

[61] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 5

[62] 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, 9

[63] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6), 2022. 1

[64] 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. *arXiv:2209.14792*, 2022. 3

[65] 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. 4

[66] 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)*, 2024. 2, 3, 6, 7, 9, 1

[67] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 3

[68] 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. *arXiv:2308.11617*, 2023. 3

[69] 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

[70] 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. 4

[71] 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, 9

[72] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 4

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

[74] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv:2402.01566*, 2024. 3

[75] 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

[76] 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, 4

[77] 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

[78] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30, 2017. 3

[79] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 28, 2015. 3

[80] Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, page 7, 2016.

[81] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual deanimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[82] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[83] 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

[84] 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. Springer, 2025. 3

[85] 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

[86] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3

[87] 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

[88] 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

[89] 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 preprint arXiv:2408.06072*, 2024. 3

[90] 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, 7

[91] 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

[92] 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

[93] 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)*, 2024. 3

[94] 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

[95] 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

[96] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18456–18466, 2023. 3

[97] 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*, pages 3836–3847, 2023. 4

[98] 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

[99] 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

[100] 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. 4

[101] 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

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

[103] 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

[104] 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

13

InterDyn: Controllable Interactive Dynamics with Video Diffusion Models
Supplementary Material

	Occl.	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	Motion Fidel. ↑
Mask	✗	**0.829**	24.08	0.123	39.94	0.666
Joints	✗	0.827	24.00	0.124	40.02	**0.673**
Mesh	✗	0.828	24.14	0.122	41.99	0.663
Mask	✓	**0.829**	24.15	0.122	**37.64**	0.675
Joints	✓	0.827	24.05	0.124	44.07	0.665
Mesh	✓	**0.829**	**24.15**	**0.121**	40.11	0.675

Table 3. **Control-signal evaluation on the DexYCB [10] dataset, 256×256 resolution.** Using binary masks does not incur significant performance loss compared to joints or meshes.

A. Additional Results

We provide more qualitative results in Fig. 9 and as videos at: https://interdyn.is.tue.mpg.de/.

B. Control Signal

In this work, we use binary masks of a hand as the control signal. However, other types of signals such as skeletons or meshes could provide richer controllability, since they inherently encode pseudo-3D information and capture correspondences across frames. Nevertheless, collecting large datasets with these types of annotations is significantly more challenging compared to binary masks, for which high-performance image segmentation models like SAM2 are readily available.

To test the impact of our choice, we ran an experiment on the DexYCB dataset [10]. Originally proposed for 3D pose estimation, DexYCB provides 8000 videos of hands grasping an object with accompanying ground-truth hand and object poses [10]. To represent 3D hand pose, DexYCB uses the parametric human hand model MANO [63], which we render as a binary mask, as joints in a similar style as OpenPose [8], and as a colored mesh using the colormap from [57], see Fig. 7.

We noticed that when generating hand masks with SAM2 on the Something-Something-v2 dataset, the resulting masks have the object's contour cut out when held in the hand. This inadvertently provides InterDyn with a small amount of information about the shape of the object that is being manipulated (a limitation shared with CosHand [66]). For this experiment, to ensure a fair comparison between control signals, we train separate versions of InterDyn that explicitly incorporate or leave out occlusions of all objects in the scene, see Fig. 7.

We train 12 different versions of InterDyn on the DexYCB dataset, each version being a combination of the following: resolution (256×256 or 256×384); control signal (mask/joints/mesh); and control signal occlusion (yes/no). In this experiment, as Tab. 3 and Tab. 4 show, we
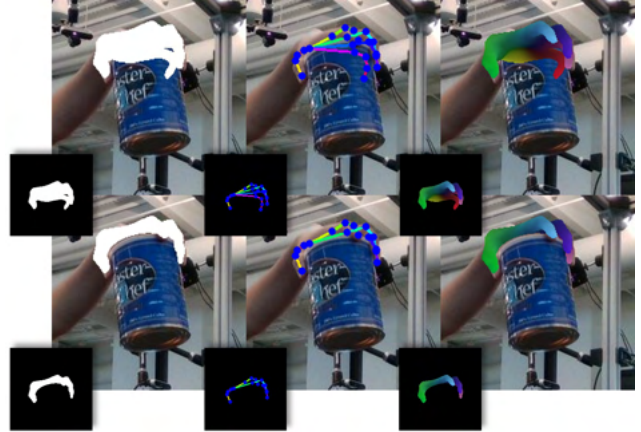

Figure 7. **Evaluated control signals.** From left to right: binary mask, joints in the style of OpenPose [8], and colored mesh [57]. Top: w/o occlusions. Bottom: with occlusions. 🔍 **Zoom in** for details.

	Occl.	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	Motion Fidel. ↑
Mask	✗	0.847	24.75	0.121	41.99	0.670
Joints	✗	0.846	24.72	0.122	41.17	0.676
Mesh	✗	0.847	24.83	0.121	42.26	0.665
Mask	✓	0.847	24.79	0.121	41.18	0.672
Joints	✓	0.846	24.69	0.122	41.41	0.676
Mesh	✓	**0.848**	**24.86**	**0.119**	**38.83**	**0.680**

Table 4. **Control-signal evaluation on the DexYCB [10] dataset, 256×384 resolution.** Using binary masks does not incur significant performance loss compared to joints or meshes.

find that the control signal has almost no effect on the image quality, spatiotemporal dynamics, and motion fidelity. This highlights the potential of using simple and easily accessible control signals to generate high-quality video outputs.

C. Limitations

Fig. 2 illustrates the current limitations of InterDyn on challenging scenarios. The first row shows that object consistency is not guaranteed, with the yellow toy changing shape when the tower falls over. The second row shows that when the object is not already in the frame, InterDyn sometimes generates an object that is not identifiable as belonging to a specific class. The third row shows a scenario in which it is particularly difficult to estimate the depth of the scene, causing InterDyn to drop the pen while it is already on top of the surface. Lastly, the fourth row shows a scenario in which a coin should be buried in flour, however, the object in our generation simply disappears.
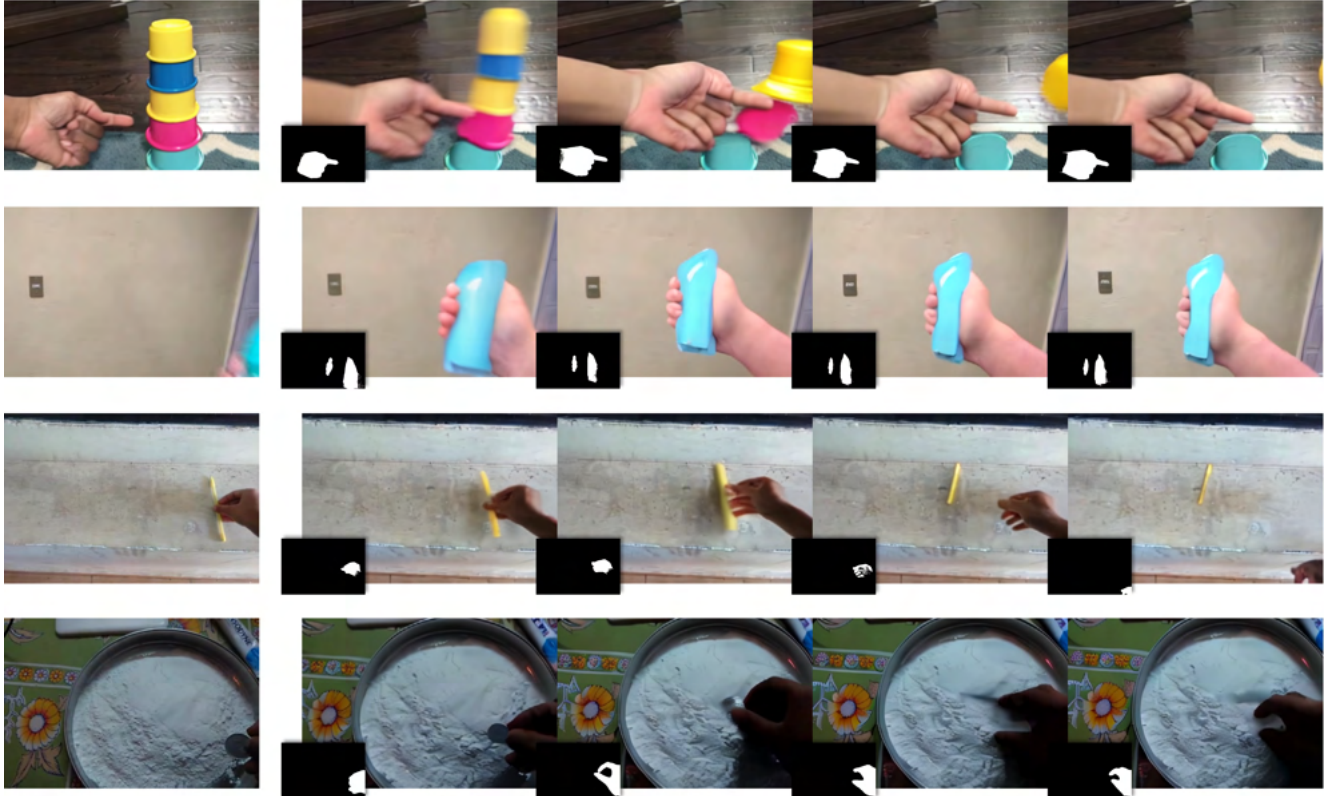
Figure 8. **Limitations of InterDyn.** We show challenging scenarios in which InterDyn underperforms, such as (from top to bottom): object consistency in highly dynamic scenarios, no object in first frame, depth ambiguity and burying an object. **Q Zoom in** for details.

D. Video Class Analysis

As mentioned in Sec. 4.4, the Something-Something-v2 dataset includes videos of human-object interactions, such as "pushing [something] from left to right". These phrases serve as class labels, enabling us to analyze the video classes where InterDyn (256×384) performs best and worst, in terms of motion fidelity. Table 5 presents the 20 best and 20 worst-performing classes based on motion fidelity, alongside the number of videos for each class in the validation set and the average motion fidelity score.

First of all, we notice that InterDyn generally performs less effectively on underrepresented classes within the dataset, while at the same time, many of these underrepresented classes involve complex dynamics, such as spinning, burying, or folding objects. From Tab. 5, we observe that InterDyn (256×384) performs best on simpler motions, particularly translations relative to the camera. This includes actions like moving an object up, down, or across, as well as dropping objects, moving objects toward or away from the camera, and picking up objects. Conversely, InterDyn struggles with classes involving highly complex dynamics, such as non-translational object-object interactions (e.g., throwing one object at another, burying an object, or poking structures made up of multiple com-

ponents like stacks or toy towers). It also underperforms on spinning objects or their parts (e.g. a fidget spinner or a ceiling fan). We believe this is not an inherent limitation of InterDyn but rather the result of a combination of factors: the dataset's frame rate (FPS), the presence of motion blur, and the under-representation of these more unique classes.

E. Ethical Considerations

Video generation models, such as InterDyn, could be misused for creating deepfake videos that spread misinformation or compromise trust. Such applications pose a large risk for society and highlight the importance of developing ethical guidelines that can serve as safeguards when developing video generation models as well as when applying them to downstream tasks. On the positive side, advances such as InterDyn could be used to enhance storytelling in entertainment or aid in deepfake detection by identifying motion-dynamic inconsistencies. By advancing our understanding of interactive dynamics responsibly, we can amplify their positive societal impact while mitigating risks.
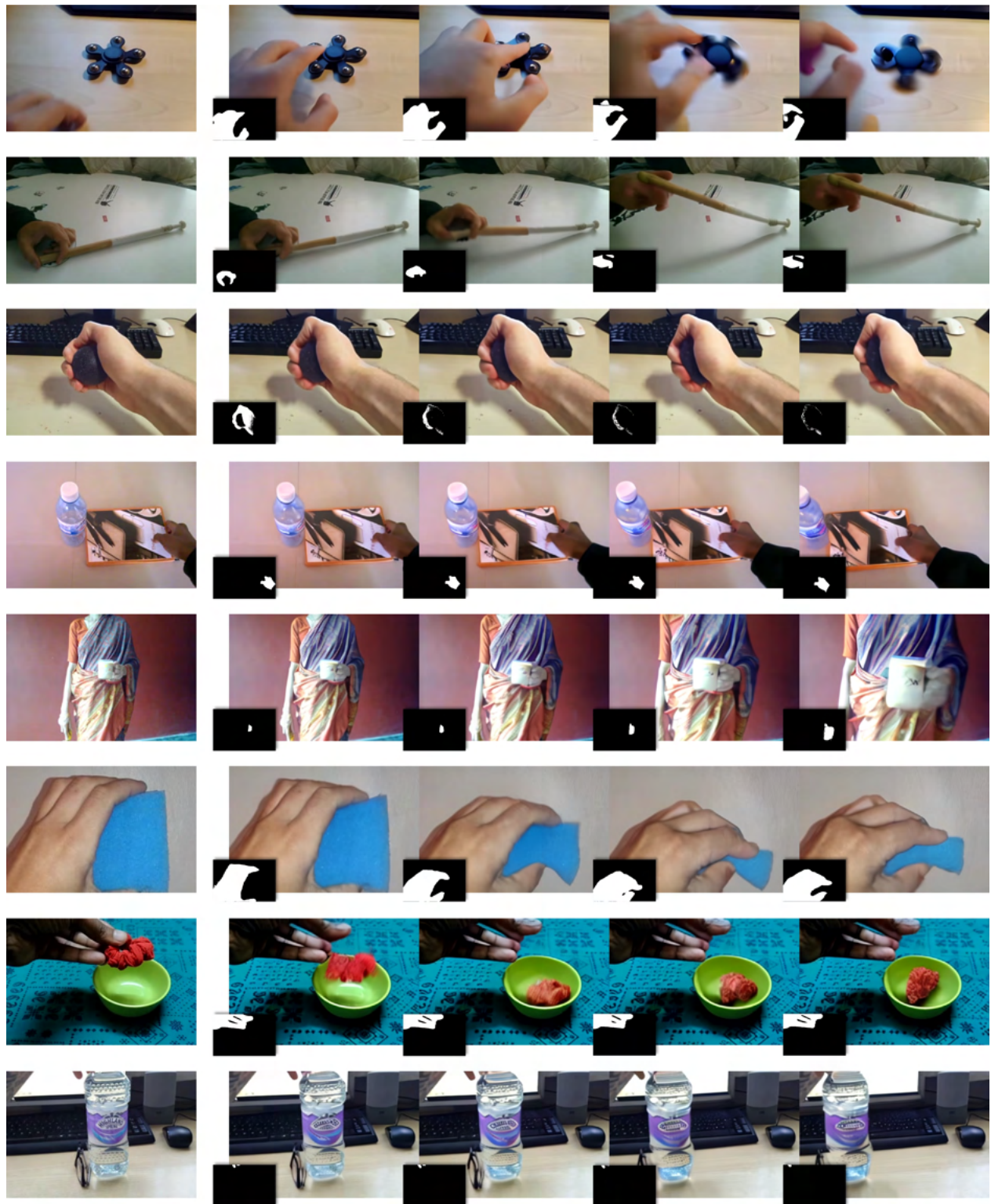
2

Figure 9. **Additional results of InterDyn on Something-Something-v2.** We show multiple challenging examples, such as (from top to bottom): spinning a fidget spinner, tilting a sleek ridged object, squeezing a ball despite receiving an incomplete control signal, hand object-object interaction, zooming in, squeezing a sponge, dropping a hairband, or hand object-object interaction despite receiving a sparse control signal. **Q Zoom in** for details.

Order	Label	Count	Avg motion fidel.
1	Moving something down	182	0.86
2	Pulling something from right to left	57	0.84
3	Moving something up	197	0.82
4	Pulling something from left to right	83	0.82
5	Holding something over something	165	0.80
6	Holding something	103	0.80
7	Moving something across a surface without it falling down	26	0.79
8	Pushing something from left to right	123	0.79
9	Holding something in front of something	138	0.78
10	Pushing something from right to left	122	0.77
11	Putting something on a surface	85	0.77
12	Moving something across a surface until it falls down	28	0.77
13	Lifting something with something on it	369	0.77
14	Squeezing something	216	0.77
15	Lifting something up completely without letting it drop down	66	0.75
16	Throwing something in the air and letting it fall	6	0.75
17	Moving something closer to something	105	0.75
18	Holding something next to something	135	0.75
19	Putting something that can't roll onto a slanted surface, so it stays where it is	15	0.75
20	Trying to bend something unbendable so nothing happens	74	0.74
…	…	…	…
114	Spinning something so it continues spinning	51	0.47
115	Poking something so that it falls over	42	0.46
116	Pulling something out of something	33	0.46
117	Folding something	187	0.46
118	Poking something so it slightly moves	71	0.45
119	Spinning something that quickly stops spinning	47	0.45
120	Taking something out of something	66	0.45
121	Unfolding something	122	0.44
122	Putting something, something, and something on the table	60	0.44
123	Piling something up	27	0.43
124	Something being deflected from something	10	0.41
125	Poking something so lightly that it doesn't or almost doesn't move	83	0.41
126	Burying something in something	4	0.41
127	Showing something next to something	19	0.40
128	Pushing something so it spins	23	0.39
129	Poking something so that it spins around	7	0.39
130	Putting number of something onto something	5	0.37
131	Poking a stack of something so the stack collapses	8	0.34
132	Showing something on top of something	14	0.34
133	Wiping something off of something	9	0.32

Table 5. **Motion fidelity for different action classes on the Something-Something-v2 dataset.** The table shows the top and bottom 25 categories in terms of motion fidelity, together with the number of samples for that category in the validation set. The top categories contain many translation dynamics with respect to the camera, such as moving something up or from left to right. The bottom categories contain very complex dynamics such as spinning, burying, or showing an object from behind something.