

---

# SViMo: Synchronized Diffusion for Video and Motion Generation in Hand-object Interaction Scenarios

---

Lingwei Dang<sup>\*1</sup>, Ruizhi Shao<sup>\*2</sup>, Hongwen Zhang<sup>3</sup>,  
Wei MIN<sup>4</sup>, Yebin Liu<sup>2</sup>, Qingyao Wu<sup>†1</sup>

<sup>1</sup>School of Software Engineering, South China University of Technology

<sup>2</sup>Department of Automation, Tsinghua University

<sup>3</sup>School of Artificial Intelligence, Beijing Normal University

<sup>4</sup>Shadow AI

## Abstract

Hand-Object Interaction (HOI) generation has significant application potential. However, current 3D HOI motion generation approaches heavily rely on predefined 3D object models and lab-captured motion data, limiting generalization capabilities. Meanwhile, HOI video generation methods prioritize pixel-level visual fidelity, often sacrificing physical plausibility. Recognizing that visual appearance and motion patterns share fundamental physical laws in the real world, we propose a novel framework that combines visual priors and dynamic constraints within a synchronized diffusion process to generate the HOI video and motion simultaneously. To integrate the heterogeneous semantics, appearance, and motion features, our method implements tri-modal adaptive modulation for feature aligning, coupled with 3D full-attention for modeling inter- and intra-modal dependencies. Furthermore, we introduce a vision-aware 3D interaction diffusion model that generates explicit 3D interaction sequences directly from the synchronized diffusion outputs, then feeds them back to establish a closed-loop feedback cycle. This architecture eliminates dependencies on predefined object models or explicit pose guidance while significantly enhancing video-motion consistency. Experimental results demonstrate our method’s superiority over state-of-the-art approaches in generating high-fidelity, dynamically plausible HOI sequences, with notable generalization capabilities in unseen real-world scenarios. Project page at [https://droliven.github.io/SViMo\\_project](https://droliven.github.io/SViMo_project).

## 1 Introduction

Human-object or hand-object interaction (HOI) generation serves critical applications across gaming, animation, digital human creation, and robotic action retargeting [16, 37, 53, 58]. Some studies [7, 13, 72] construct high-precision 3D interaction datasets through laboratory-based multi-view camera arrays and motion capture systems, then train diffusion-based motion generators. These object-centric approaches typically predict parametric human/hand motions and the corresponding object pose sequences given object meshes and initial configurations. However, existing datasets [39, 71, 38] collected in laboratory environments lack diversity in object types and interaction patterns, constraining model generalization and resulting in ambiguous object boundaries, implausible or inconsistent actions (Fig. 1, left). Moreover, the reliance on precise 3D object models fundamentally limits their zero-shot generation capabilities.

---

<sup>\*</sup>Equal contributions. Email: levondang@163.com, jia1saurus@gmail.com.

<sup>†</sup>Corresponding Author. Email: qyw@scut.edu.cn.

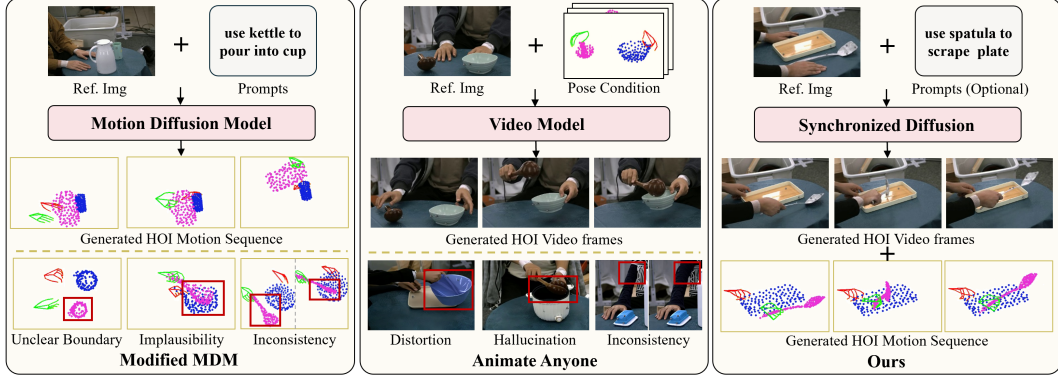


Figure 1: Different HOI generation methods. Approaches like MDM [57] rely on limited mocap data without visual guidance, resulting in blurred boundaries and compromised plausibility and consistency. Methods like Animate Anyone [21] leverage large-scale visual priors but exhibit distortions and inconsistencies because of inadequate physical awareness. Our method marries visual priors with 3D motion constraints and eliminates dependency on pre-defined object models or pose guidance.

Recent advances in large video foundation models based on Diffusion Transformers (DiT) [44] (*e.g.*, Sora [5], CogVideo [20, 69], HunyuanVideo [26]), have shown impressive capabilities in modeling physical dynamics through large-scale video training. These models can generate interaction videos with high visual fidelity end-to-end from text or reference images. However, their pixel-level generation approaches often struggle to produce accurate and coherent hand-object interactions due to limited explicit modeling of motion dynamics and physical constraints. To address this, some methods extend image-based diffusion models (*e.g.*, SVD [3]) by adding pose-guided pipelines [66, 21, 75, 65]. These approaches combine pose conditions and appearance features to improve human-object interaction generation. While effective, they require pose sequences or externally estimated motion trajectories as inputs, preventing full end-to-end text/image-conditioned generation. Additionally, their single-frame generation leads to poor temporal coherence, causing flickering and identity inconsistencies (Fig. 1, middle).

These challenges reveal a longstanding methodological contradiction: motion generation systems excel at physical constraint modeling but suffer from limited data scales. In contrast, video generation models leverage massive visual priors but lack motion plausibility. We argue that this division stems from neglecting the co-evolution mechanism between visual appearance and motion patterns: they not only share the same foundation of physical dynamics but also could leverage similar diffusion processes. Based on this insight, we propose SViMo, a **S**ynchronized **V**ideo-**M**otion diffusion framework that enables synchronous HOI video generation and motion synthesis within a unified architecture (Fig. 1, right). The core innovation lies in extending a pretrained image-to-video foundation model into the multimodal joint generation framework through the scalable DiT architecture. To better integrate the heterogeneous features of text semantics, visual appearance, and motion dynamics, we introduce the triple modality adaptive modulation to align feature scales and employ a 3D full-attention mechanism to learn their synergistic and complementary dependencies. Additionally, it is difficult for video foundation models to learn explicit 3D interaction motions directly. To bridge the representation gap and reduce optimization complexity, we project 3D motions onto 2D image planes, constructing “rendered motion videos” as SViMo’s motion representation.

To further enhance the video-motion consistency, we design a Vision-aware 3D Interaction Diffusion model (VID). This model generates explicit 3D hand poses and object point clouds using denoised latent codes from the synchronized diffusion, which are then reinjected into the SViMo as interaction guidance and gradient constraints. Unlike methods requiring pre-specified action series, our approach integrates video synthesis and 3D interaction generation within an end-to-end denoising pipeline. This creates a closed-loop feedback mechanism where motion guidances refine video generation while video latents update motion results, enabling synergistic co-evolution of both modalities.

In summary, our contributions are threefold:

- A novel synchronized diffusion model for joint HOI video and motion denoising, effectively integrating large-scale visual priors with motion dynamic constraints.
- A vision-aware 3D interaction diffusion that generates explicit 3D interaction sequences, forming a closed-loop optimization pipeline and enhancing video-motion consistency.
- Our method generates HOI video and motion synchronously without requiring pre-defined poses or object models. Experiment results demonstrate superior visual quality, motion plausibility, and generalization capability to unseen real-world data.

## 2 Related Work

**3D interaction synthesis** relies on high-precision motion capture datasets, some of which focus on human action conditioned on static objects [18, 73], while others simultaneously capture interactions of both human and dynamic objects [8, 39, 67, 71, 38, 54, 14]. Building upon these datasets, existing 3D interaction generation methods employ diffusion models to either introduce intermediate contact maps or milestones for modeling kinematic features [7, 13, 29, 46, 45, 72, 62, 28, 27, 30, 36, 31, 37] or leverage physical simulations to ensure physical dynamics plausibility [64, 60, 4, 63, 41]. However, due to the limited availability of 3D interaction data, the generalization capability of these approaches remains constrained. In contrast, our method leverages large-scale visual priors, operates conveniently without requiring 3D object models, and demonstrates promising generalization potential.

**Interaction Video Generation.** The success of image generation models [19, 50–52] has inspired significant advances in video generation. Several approaches [3, 66, 43, 21, 75, 65, 61] extend 2D image denoising U-Nets to video by incorporating temporal attention layers, while improving controllability via pose guidance and reference networks. Alternatively, native large video models [5, 20, 69, 26, 25] directly generate videos using spatiotemporal DiT architectures [44]. A separate line of work aims to generate both video and interactive motion dynamics [9, 11]. For instance, VideoJAM [9] produces video frames alongside 2D optical flow to improve visual fidelity and motion coherence. In contrast, our method generates videos with explicit 3D motion representation. This enables direct modeling of object geometry, scale, and spatial relationships, leading to superior physical plausibility particularly in complex scenarios involving occlusions or intricate interactions. Notably, Vlogger [11] employs a two-stage pipeline: it first generates 3D human poses using one network, then synthesizes video via a separate model conditioned on the predicted motions. In contrast, we unify video and motion generation within a single, synchronized diffusion framework. Since human-object interaction videos and their corresponding motion dynamics inherently obey the same physical laws, joint modeling allows the system to learn intrinsic co-evolutionary patterns, thereby enhancing the quality and consistency of both modalities.

**Multimodal Generative Models.** Driven by advancements in vision-language models [10, 34, 59, 2, 68], researchers have explored versatile generative models that align with the multimodal essence of the physical world. Some works [23, 70] cascade single-modality generators into asynchronous pipelines, yet suffer from complex workflows and noise accumulation. Others develop end-to-end multimodal joint generation frameworks that rely on massive aligned multimodal data [1, 32, 17, 76], indirect bridging mechanisms [55], or intricate hierarchical attention strategy [35] to ensure cross-modal synchronization. Differently, our method extends a native large video model into synchronized video-motion generation systems, aligning heterogeneous modality features through multimodal adaptive modulation and closed-loop feedback strategies.

## 3 Methodology

We define the HOI video and motion generation task as follows: Given a reference image frame  $I \in \mathbb{R}^{H \times W \times 3}$  and a textual prompt  $P$ , generate the future video  $V \in \mathbb{R}^{N \times H \times W \times 3}$  and 3D motion sequence  $M = \{(\mathbf{h}_i, \mathbf{o}_i)\}_{i=1}^N$  with  $N$  time steps. Here  $\mathbf{h}_i \in \mathbb{R}^{J \times 3}$  and  $\mathbf{o}_i \in \mathbb{R}^{K \times 3}$  are the hand joints trajectories and object point clouds, each has  $J$  and  $K$  nodes respectively. The following sections detail each component of our approach. See the **Appendix** for training/inference pseudo-code and additional details.

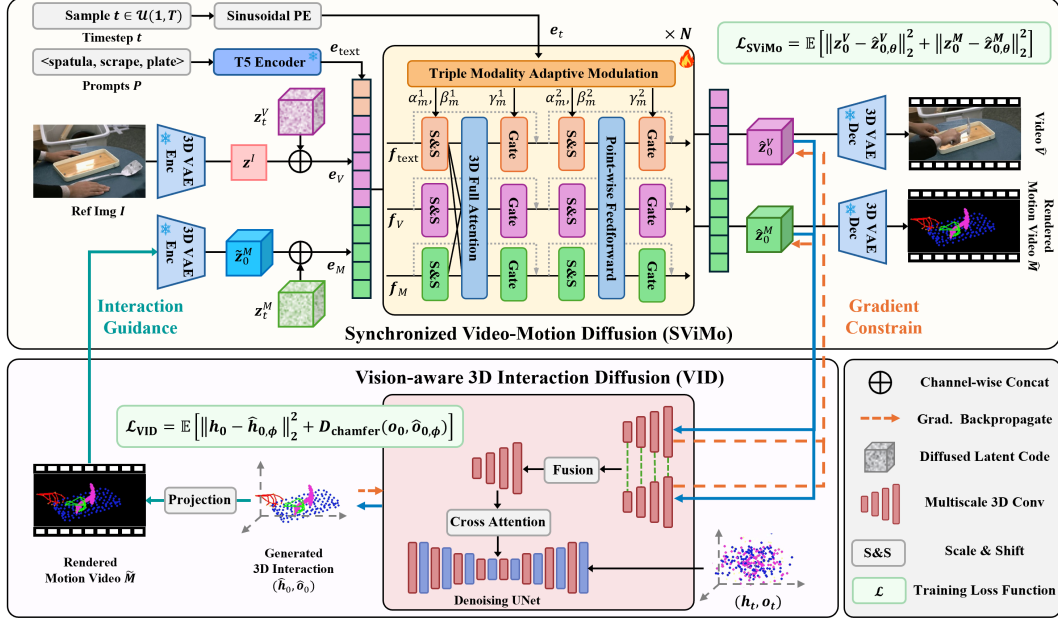


Figure 2: Our method comprises: (1) A synchronous diffusion model that jointly generates HOI videos and motions (Sec. 3.3). (2) A vision-aware interaction diffusion model that generates 3D hand pose trajectories and point clouds from the former’s outputs (Sec. 3.4), then feeds them back into the synchronized denoising process to establish closed-loop optimization (Sec. 3.5).

### 3.1 Preliminary: Basic Large Video Generation Model

The core of large video generation models is diffusion and its reverse denoising process [19]. The diffusion process adds Gaussian noise to video latent codes  $z$  step-by-step using  $q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t}z_{t-1}, \beta_t \mathbf{I})$ , where  $\beta_t \in [0, 1]$  increases monotonically with time step  $t$ , and  $\alpha_t = 1 - \beta_t$ . Repeated application of this formula yields the marginal distribution:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad \bar{\alpha}_t = \prod_{1}^t (1 - \beta_t). \quad (1)$$

The above distribution converges to an isotropic Normal distribution when  $t$  grows large.

Correspondingly, the reverse denoising generator  $\mathcal{G}$  with trainable weights  $\theta$  predicts the clean latent code  $\hat{z}_{0,\theta}$ , and is optimized by:

$$\mathcal{L}_{x_0\text{-prediction}} = \mathbb{E}_{\mathbf{z}, \mathbf{c}, t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} [\|\mathbf{z}_0 - \mathcal{G}_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2]. \quad (2)$$

During the inference phase, the trained model is utilized to iteratively denoise from pure Gaussian noise step by step, following the procedure:

$$\begin{cases} \hat{z}_{0,\theta} = \mathcal{G}_\theta(\mathbf{z}_t, \mathbf{c}, t), \\ \mathbf{z}_{t-1} \sim \mathcal{N}\left(\frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{z}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t}\hat{z}_{0,\theta}, \frac{1 - \bar{\alpha}_{t-1}(1 - \alpha_t)}{1 - \bar{\alpha}_t}\mathbf{I}\right). \end{cases} \quad (3)$$

Finally, we obtain the clean denoised latent code  $\hat{z}_0^*$ , which is then decoded to the raw video space.

### 3.2 Framework Overview

Based on the insight that visual appearance and motion dynamics share inherent physical laws of the real world, we propose an end-to-end framework that unifies HOI video and 3D motion generation by integrating visual priors with dynamic constraints. Our framework consists of two key components. The first is SViMo, which focuses on video generation (Fig. 2, top; Sec. 3.3). It extends a pre-trained image-to-video foundation model into a joint video-motion generation architecture. During joint



denoising, SViMo dynamically aligns visual appearance with high-level motion signals, improving both visual quality and motion plausibility. Notably, the motion representations in SViMo refer not to explicit 3D hand trajectories or object point cloud sequences. Because there is a significant gap between 3D motion data and 2D video representations, directly modeling explicit 3D motions would disrupt the pre-trained foundation model’s visual priors and degrade performance. Instead, we project explicit 3D interactions onto the 2D image plane to create “rendered motion videos” as intermediate representations for SViMo. The second component is VID, which generates explicit 3D motions (Fig. 2, bottom; Sec. 3.4). It maps the output of SViMo, both the generated video and the 2D rendered motion video, into target 3D interactions, *i.e.*, hand-object trajectories and object point clouds. Additionally, the collaboration between SViMo and VID forms a closed-loop feedback pipeline, ensuring consistency between generated videos and 3D motions (Sec. 3.5).

### 3.3 Synchronized Video-Motion Diffusion

The SViMo learns to predict the video and motion, given the time step, text prompt, and reference image:  $(\hat{z}_0^V, \hat{z}_0^M) = \mathcal{G}_\theta((z_t^V, z_t^M), (\mathbf{P}, \mathbf{I}), t)$ . The entire video-motion generation process comprises feature embedding, multimodal feature modulation and fusion, and joint denoising.

**Feature Embedding.** Time step  $t$  is sampled from a uniform distribution and then added sinusoidal position encoding, followed by a simple two-layer linear mapping to obtain the time embedding  $e_t \in \mathbb{R}^{d_{\text{time}}}$ . For the text prompt, we use a frozen pre-trained language model, Google T5 [48], to extract text embedding  $e_{\text{text}} \in \mathbb{R}^{L \times d_p}$ , and then calculate the semantics features  $\mathbf{f}_{\text{text}}$  with linear projection, where  $L$  is the max length of textual tokens. Thirdly, for the original reference image  $\mathbf{I}$ , to ensure a certain level of robustness, we first add random noise yielding  $\mathbf{I}_{\text{noised}}$ . Then, we compute the compressed latent code  $\mathbf{z}_I$  through a 3D video VAE,  $\mathbf{z}_I = \mathcal{E}(\mathbf{I}_{\text{noised}}) \in \mathbb{R}^{\frac{H}{r_h} \times \frac{W}{r_w} \times d_{\text{VAE}}}$ . Additionally, for the target video  $\mathbf{V}$ , we also map it to the latent space with the same VAE. Then the noised latent code  $\mathbf{z}_t^V \in \mathbb{R}^{(\frac{N-1}{rn}+1) \times \frac{H}{r_h} \times \frac{W}{r_w} \times d_{\text{VAE}}}$  are obtained based on the forward diffusion process in Eq. 1. To better align it with the image condition signal during the video denoising process, we repeat the image feature  $\mathbf{z}_I$  along the temporal axis, and concatenate them along the channel-axis to get video feature embedding  $\mathbf{e}_V \in \mathbb{R}^{(\frac{N-1}{rn}+1) \times \frac{H}{r_h} \times \frac{W}{r_w} \times (2 \times d_{\text{VAE}})} = \mathbf{z}_t^V \oplus \mathbf{z}_I$ . Then the visual feature is patchified through a 2D stride convolution, resulting in  $\mathbf{f}_V \in \mathbb{R}^{[\frac{N-1}{rn}+1] \times \frac{H}{2 \times r_h} \times \frac{W}{2 \times r_w} \times d}$ . The rendered motion video is encoded through the same VAE and diffused to obtain  $\mathbf{z}_t^M$ . In contrast, to get motion embedding  $\mathbf{e}_M$ , the diffused latent code here are concatenated with interaction guidance  $\hat{\mathbf{z}}_0^M$  provided by VID (Sec. 3.4), rather than being combined with  $\mathbf{z}_I$ . After that, the motion feature  $\mathbf{f}_M$  is obtained in the same way as that of  $\mathbf{f}_V$ .

**Multimodal Feature Modulation and Fusion.** In our SViMo framework, the DiT token sequence comprises three distinct modalities: text tokens  $\mathbf{f}_{\text{text}}$ , video tokens  $\mathbf{f}_V$  and motion tokens  $\mathbf{f}_M$ , which differ significantly in both feature spaces and numerical scales. To bridge these disparities while preserving modality-specific characteristics, we adopt a triple modality adaptive modulation method that learns modulation parameters from the timestep signal  $e_t$ . These parameters determine the scaling, shifting, and gating operations of each modality’s features separately. Additionally, a 3D full-attention mechanism is employed to capture intra- and inter-modal relationships. Take the processing of text features as an example, the DiT Block  $\mathcal{B}$  proceeds as follows:

$$\mathcal{B}(\cdot) = \left\{ \begin{array}{l} \{\alpha_{\text{text}}^i, \beta_{\text{text}}^i, \gamma_{\text{text}}^i\}_{i=1}^2 = \text{MLP}(e_t), \\ \mathbf{f}'_{\text{text}} = \text{LN}(\mathbf{f}_{\text{text}}) \odot (1 + \alpha_{\text{text}}^1) + \beta_{\text{text}}^1, \\ \mathbf{f}'_{\text{text}} = \mathbf{f}'_{\text{text}} + \gamma_{\text{text}}^1 \odot \cap_{\text{text}} (3\text{DFA}(\cup(\mathbf{f}'_{\text{text}}, \mathbf{f}'_V, \mathbf{f}'_M))), \\ \mathbf{f}''_{\text{text}} = \text{LN}(\mathbf{f}'_{\text{text}}) \odot (1 + \alpha_{\text{text}}^2) + \beta_{\text{text}}^2, \\ \mathbf{f}''_{\text{text}} = \mathbf{f}''_{\text{text}} + \gamma_{\text{text}}^2 \odot \cap_{\text{text}} (\text{FFD}(\cup(\mathbf{f}'_{\text{text}}, \mathbf{f}'_V, \mathbf{f}'_M))), \end{array} \right\} \text{ same for } \mathbf{f}_V \text{ and } \mathbf{f}_M, \quad (4)$$

where “3DFA” and “FFD” are unified multi-head 3D full-attention and feedforward layers,  $\cup$  and  $\cap$  denote token concatenation and segmentation along the sequence dimension, respectively.

**Video-Motion Joint Denoising.** The video and motion features output by the final DiT block then go through an MLP to reconstruct the VAE latent codes, yielding  $\hat{\mathbf{z}}_0^V$  for video and  $\hat{\mathbf{z}}_0^M$  for motion. Finally, the SViMo is optimized according to Eq. 2:

$$\mathcal{L}_{\text{SViMo}} = \mathbb{E}_{(\mathbf{z}^V, \mathbf{z}^M), (\mathbf{P}, \mathbf{I}), t, \epsilon} [\|\mathbf{z}_0^V - \mathcal{G}_\theta(\mathbf{z}_t^V, (\mathbf{P}, \mathbf{I}), t)\|_2^2 + \|\mathbf{z}_0^M - \mathcal{G}_\theta(\mathbf{z}_t^M, (\mathbf{P}, \mathbf{I}), t)\|_2^2]. \quad (5)$$

### 3.4 Vision-aware 3D Interaction Diffusion

The VID  $\mathcal{M}_\phi$  generates the explicit 3D hand pose trajectories  $\hat{\mathbf{h}}_0$  and object point cloud sequences  $\hat{\mathbf{o}}_0$  given latent codes of videos  $\mathbf{z}_t^V$  and motions (rendered motion video)  $\mathbf{z}_t^M$  at any time. The framework operates as follows: First, a dual-stream 3D convolutional module encodes multi-scale spatiotemporal features from both video and motion codes. Then they are fused and subsequently injected into the 3D interaction denoising modules through cross-attention mechanisms to synthesize 3D HOI trajectory sequences. Following the  $\mathbf{x}_0$ -prediction formula in Eq. 2, the model is optimized through the following loss function:

$$\begin{cases} (\hat{\mathbf{h}}_{0,\phi}, \hat{\mathbf{o}}_{0,\phi}) = \mathcal{M}_\phi [(\mathbf{h}_t, \mathbf{o}_t), (\mathbf{z}_t^V, \mathbf{z}_t^M), t], \\ \mathcal{L}_{\text{VID}} = \mathbb{E}_{(\mathbf{h}, \mathbf{o}), (\hat{\mathbf{z}}_\theta^V, \hat{\mathbf{z}}_\theta^M), t, \epsilon} [\text{MSE}(\mathbf{h}_0, \hat{\mathbf{h}}_{0,\phi}) + \text{D}_{\text{chamfer}}(\mathbf{o}_0, \hat{\mathbf{o}}_{0,\phi})]. \end{cases} \quad (6)$$

### 3.5 Close-loop Feedback and Training Objectives

**Close-loop Feedback.** To enhance the mutual promotion and co-evolution between SViMo and VID, and improve the consistency between video and 3D motions, we design a closed-loop feedback mechanism. This mechanism includes two pathways: interaction guidance and gradient constraint. Specifically, for the straightforward interaction guidance strategy (Eq. 7, row 1st), we first generate the 3D interaction from the video and motion inputs of SViMo:  $(\tilde{\mathbf{h}}_0, \tilde{\mathbf{o}}_0) = \mathcal{M}_{\text{no-grad}}((\mathbf{h}_t, \mathbf{o}_t), (\mathbf{z}_t^V, \mathbf{z}_t^M), t)$ , then projected it onto the 2D image plane to obtain rendered motion video  $\tilde{\mathbf{M}}$ , which are subsequently embedded into the VAE latent space yielding  $\tilde{\mathbf{z}}_0^M$ . Finally, it is concatenated with the noised motion latent code  $\mathbf{z}_t^M$  mentioned in Sec. 3.3, forming an additional interaction guidance for the SViMo. On the other hand, the input of VID could come from the output of the SViMo. Therefore, the gradient of VID will be backpropagated into the SViMo during the training process, forming a gradient constraint path and thereby promoting its optimization (Eq. 7, row 2nd).

$$\begin{cases} (\mathbf{z}_t^V, \mathbf{z}_t^M) \xrightarrow{\text{VID}_{\text{no-grad}}} (\tilde{\mathbf{h}}_0, \tilde{\mathbf{o}}_0) \xrightarrow{\text{Proj.}} \tilde{\mathbf{M}} \xrightarrow{\text{VAE } \mathcal{E}} \tilde{\mathbf{z}}_0^M \xrightarrow{\text{Inter. Guid. to SViMo}} (\mathbf{z}_t^V, \mathbf{z}_t^M \oplus \tilde{\mathbf{z}}_0^M), \\ (\mathbf{z}_t^V, \mathbf{z}_t^M \oplus \tilde{\mathbf{z}}_0^M) \xleftarrow[\text{Gradient}]{\text{SViMo}} (\hat{\mathbf{z}}_0^V, \hat{\mathbf{z}}_0^M) \xleftarrow[\text{Gradient}]{\text{VID}} (\hat{\mathbf{h}}_0, \hat{\mathbf{o}}_0) \xleftarrow[\text{Gradient}]{\text{Loss}} \mathcal{L}_{\text{VID}}. \end{cases} \quad (7)$$

**Training Objectives.** The training process of our method involves two phases: initially warming up the VID based on Eq. 6, followed by closed-loop training where the SViMo and the VID are jointly optimized according to Eq. 6 and Eq. 5:  $\mathcal{L} = \omega_1 \mathcal{L}_{\text{SViMo}} + \omega_2 \mathcal{L}_{\text{VID}}$ .

## 4 Experiments

We conducted extensive experiments to validate the effectiveness of our proposed method. More information, such as additional results and limitation discussions, is provided in the **Appendix**.

### 4.1 Experimental Setup

**TACO dataset** [38] is a large-scale bimanual hand-object interaction dataset capturing diverse tool-use behaviors via multi-view video recordings and high-fidelity 3D motion annotations. Each task is defined as a triplet <tool category, action type, target object category>, describing tool-mediated interactions with objects. The dataset includes 2.5k interaction sequences, covering 20 object categories, 196 3D models, 14 participants, and 15 daily interaction types. It provides allocentric (4096×3000) and egocentric (1920×1080) video streams, totaling 5.2M frames at 30 Hz. To reduce computational load, we crop hand-object interaction regions, adjust their aspect ratio to 3:2, and downsample to 49 frames at 8 FPS. This results in videos with spatiotemporal resolution 416×624×49, aligning with CogVideoX’s [69] default settings while lowering spatial resolution. To mitigate overlap between test and training sets that might compromise evaluation of generalization, we implemented a two-stage data partitioning strategy. First, we reserved all instances involving specific actions (*e.g.*, hit), tools (*e.g.*, glue gun), and objects (*e.g.*, toy) as the initial test set, ensuring these elements are absent from the training set. Second, from the remaining instances grouped

Table 1: Comparison of video generation results. The best and second-best results are highlighted with **bold** and underline formatting.

| Method                                   | Type           | Training | Content Quality                |                                | Dynamic Quality         |                         | Overall $\uparrow$             |
|------------------------------------------|----------------|----------|--------------------------------|--------------------------------|-------------------------|-------------------------|--------------------------------|
|                                          |                |          | Subj. $\uparrow$               | Bkg. $\uparrow$                | TSmo. $\uparrow$        | Dyn. $\uparrow$         |                                |
| Hunyuan-13B [26]<br>Wan-14B [25]         | 3D V           | w/o      | <b>0.9632</b><br><u>0.9576</u> | <b>0.9627</b><br><u>0.9620</u> | <u>0.9889</u><br>0.9829 | 0.4900<br>0.8476        | 0.4493<br>0.7675               |
| Animate Anyone [21]<br>Easy Animate [61] | 2.5D V         | w/       | 0.9206<br>0.9243               | 0.9302<br>0.9358               | 0.9671<br>0.9657        | 0.9867<br><b>0.9933</b> | 0.8172<br>0.8297               |
| CogVideoX-5B [69]<br>Ours                | 3D V<br>3D V&M |          | 0.9404<br>0.9500               | 0.9477<br>0.9533               | 0.9858<br><b>0.9898</b> | <b>0.9933</b><br>0.9801 | <u>0.8727</u><br><b>0.8785</b> |

by <action, tool, object> triplets, we applied weighted sampling according to group sizes to obtain additional test data. Finally, the ratio of the training set to the test set is 9:1.

**Evaluation Metric.** For video evaluation, we use VBench [22] to assess two key dimensions: Content Quality (including **Subject Consistency** and **Background Consistency**) and Dynamic Quality (**Temporal Smoothness** and **Dynamic Degree**). More details about these metrics can be found in the Appendix. To address the partiality of individual metrics, we multiply them to derive a **Overall** score for holistic evaluation. For 3D interaction evaluation, we separately assessed hand poses and object point cloud sequences. For the former, we calculated **MPJPE** (Mean Per Joint Position Error) and **Motion Smoothness** metrics. For object evaluation, we measured the **Chamfer Distance** between generated and ground-truth point clouds. Additionally, we compute a comprehensive **FID** score via a pretrained interaction autoencoder.

## 4.2 Implementation Details

**Network Architecture.** Our proposed model generates hand-object interaction videos with resolution  $[H, W, N] = [416, 624, 49]$  and 3D motion sequences containing  $J = 42$  hand keypoints and  $K = 298$  object nodes. During training, timesteps are uniformly sampled from  $[0, 1000]$ , and the embedding dimension is  $d_{\text{time}} = 512$ , while only 50 steps are sampled during inference for acceleration. Text prompts (max length  $L = 226$ ) are embedded into features of dimension  $d_p = 4096$ . The video VAE uses spatial-temporal compression ratios  $[rw, rh, rn] = [8, 8, 4]$ , producing latent codes of size  $[52 \times 78 \times 13]$  with  $d_{\text{VAE}} = 16$  channels. After patchification via two-step stride convolution, video and motion tokens are compressed to 13182 tokens each. This results in a total multimodal token sequence length of  $226 + 13182 \times 2 = 26590$  (text + video + motion). The DiT backbone comprises 42 DiT Blocks, each with 48 attention heads, totaling 6.02B parameters.

**Training Details.** All models are trained on 4 NVIDIA A800-80G GPUs. With memory optimization techniques including DeepSpeed ZeRO-3 [49], gradient checkpointing, and BF16 mixed-precision trick, we achieve a per-GPU batch size of 4. We first warm up the VID for 5k steps, then conduct joint training with the SViMo. To enhance computational efficiency, we initially train at reduced resolution  $[H', W'] = [240, 368]$  for 30k steps before fine-tuning at full resolution for 5k steps. The weights of SViMo and VID terms in the training objectives are  $\omega_1 = 1$  and  $\omega_2 = 0.05$ .

## 4.3 Comparison with Previous Approaches

**Baselines.** For video generation performance, we compare with video models that follow the image-animation paradigm (2.5D Video Models), including Animate Anyone [21] and Easy Animate [61], as well as native 3D large video models, including Hunyuan-13B [26], Wan-14B [25], and CogVideoX-5B [69]. Particularly, due to the high training costs of the first two 3D models, we directly utilized them for zero-shot inference. For 3D motion generation quality, we compare with the classic MDM [57] and its latest improved version EMDM [74]. Our method generates motion from both images and text (image+text-to-motion), whereas MDM and EMDM are inherently text-to-motion systems. For fair comparison, we adapted both baselines into text-and-image-to-motion generators by supplying additional reference frames to their CLIP encoders.

**Quantitative and Qualitative Evaluation.** As shown in Table 1, our method achieves the highest overall score in video generation. Notably, individual metrics often conflict: high scores in one

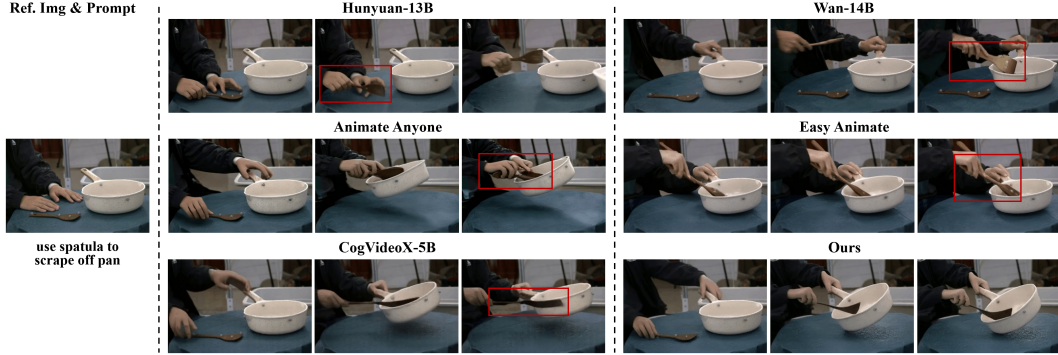


Figure 3: Visualization of videos. Red boxes highlight artifacts such as deformation, hallucinations, distortion, and implausible motions.

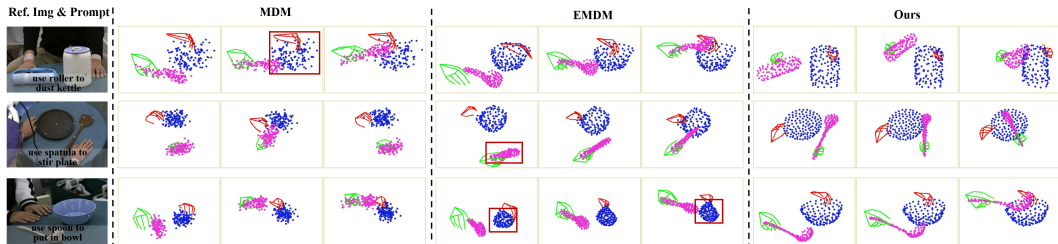


Figure 4: Quantitative comparison of motion generation results. Our method achieves smoother motions, sharper object point clouds, and higher instruction compliance.

aspect may compromise others. For instance, Hunyuan-13B [26] attains top subject/background consistency and second-highest temporal smoothness, but its near-static outputs (dynamic degree: 0.49) yield the lowest overall score. Wan-14B [25] also exhibits the same phenomenon, and the lack of TACO dataset fine-tuning results in poor instruction adherence and hallucinations (Fig. 3, Row 1). The 2.5D image animation methods [21, 61] achieve high dynamic degree yet exhibit inadequate content consistency and temporal smoothness, visually manifesting as distortions and temporal flickering (Fig. 3, Row 2). CogVideoX-5B [69] achieves the second-highest overall score, but the generated videos still exhibit inconsistencies, as shown on the left of the last row of Fig. 3. In contrast, our method benefits from the synchronized modeling of visual and dynamic, resulting in better comprehensive performance.

For motion generation, our method achieves superior performance across all metrics, as shown in Tab. 2. Qualitative results in Fig. 4 reveal that MDM [57] and EMDM [74] produce motions with poor instruction compliance and frame consistency. This stems from two limitations: 1) They compress both reference images and text prompts into 512 dimensions through the CLIP encoder, then simply concatenate them as denoising conditions, which dilutes the instruction signal. 2) Their motion models lack vision awareness, causing large discrepancies between generated point clouds and references. Contrastingly, our approach not only preserves input condition effectiveness through a triple-modality adaptive modulation mechanism, but also enhances object point cloud consistency with low-level visual priors.

**User Study.** To validate our method’s effectiveness, we conducted user studies for video and motion generation (Fig. 5). For video generation, 26 image-prompt pairs were used to generate videos with six models each, yielding 1,066 valid responses from 41 participants. Our method achieved a 78.42% preference rate, significantly outperforming all baselines. In motion generation, 10 image-prompt pairs produced 410 valid responses, with our results surpassing the baseline in 97.56% of cases. These results demonstrate the clear advantages of our video-motion synchronous diffusion model.

Table 2: Quantitative comparison against motion generation methods. The best results are in **bold**.

| Method    | Hand          |               | Object        | HOI           |
|-----------|---------------|---------------|---------------|---------------|
|           | MPJPE ↓       | MSmo. ↓       | Cham. ↓       | FID ↓         |
| MDM [57]  | 0.3382        | 0.0365        | 0.7915        | 0.4056        |
| EMDM [74] | 0.3255        | 0.0306        | 0.7788        | 0.3681        |
| Ours      | <b>0.1087</b> | <b>0.0255</b> | <b>0.1577</b> | <b>0.1050</b> |

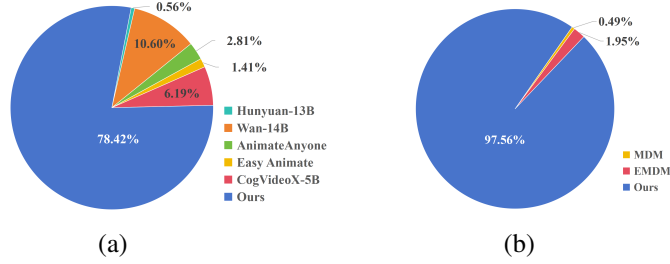


Figure 5: User studies for generated videos (a) and motions (b). We received 1,066 and 410 valid responses respectively, and our method significantly outperforms other baselines.

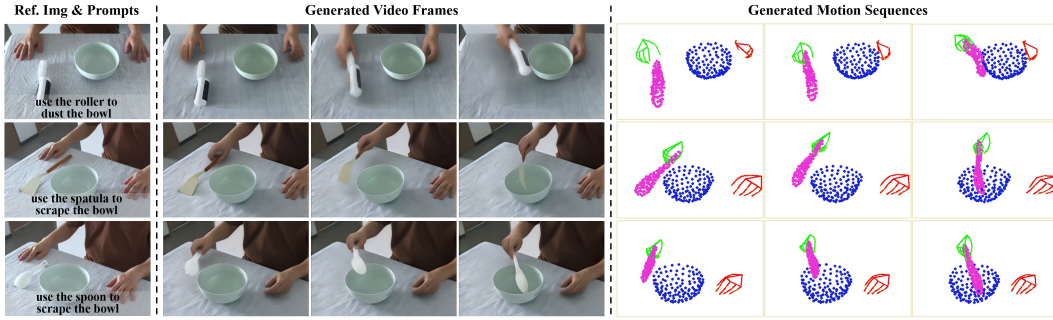


Figure 6: Zero-shot inference on real-world data. We collect data with everyday objects and generate high-fidelity, plausible HOI videos and 3D motions, demonstrating the generalizability of our method.

Table 3: Ablation studies on the synchronized diffusion and the vision-aware 3D interaction diffusion.

| Variants              | Content          |                 | Dynamic          |                 | Overall $\uparrow$ | Hand               |                    | Object             | HOI              |
|-----------------------|------------------|-----------------|------------------|-----------------|--------------------|--------------------|--------------------|--------------------|------------------|
|                       | Subj. $\uparrow$ | Bkg. $\uparrow$ | TSmo. $\uparrow$ | Dyn. $\uparrow$ |                    | MPJPE $\downarrow$ | MSmo. $\downarrow$ | Cham. $\downarrow$ | FID $\downarrow$ |
| SViMo w/ VID (Ours)   | 0.9534           | <b>0.9546</b>   | <b>0.9883</b>    | <b>0.9784</b>   | <b>0.8800</b>      | <b>0.0121</b>      | <b>0.0053</b>      | <b>0.0019</b>      | <b>0.0100</b>    |
| SViMo w/ Inter. Guid. | 0.9522           | <b>0.9546</b>   | 0.9877           | 0.9768          | 0.8770             | 0.0157             | 0.0060             | 0.0022             | <b>0.0100</b>    |
| SViMo w/ Grad. Cons.  | 0.9499           | 0.9525          | <u>0.9881</u>    | <u>0.9757</u>   | 0.8723             | <u>0.0141</u>      | <u>0.0058</u>      | <u>0.0021</u>      | 0.0124           |
| SViMo w/o VID         | <b>0.9543</b>    | 0.9545          | <b>0.9883</b>    | 0.9686          | 0.8719             | 0.0195             | 0.0070             | 0.0037             | 0.0546           |
| VModel w/ Pred. Mot.  | 0.9356           | 0.9392          | 0.9858           | 0.9675          | 0.8381             | 0.0202             | 0.0074             | 0.0040             | 0.0575           |

#### 4.4 Generalization and Extensibility

We evaluate our method on manipulation tasks involving common household objects such as rollers, spatulas, spoons, and bowls, for which we collect image-prompt pairs. These inputs are processed by our synchronized diffusion model to generate human-object interaction (HOI) videos and corresponding 3D interactions, as shown in Fig. 6. The successful generation of plausible interactions on real-world object categories demonstrates the generalization capability of our approach. This generalization stems from two key design principles. First, our plug-and-play synchronized diffusion architecture effectively leverages the visual-semantic priors encoded in large video foundation models. Second, the use of 3D point clouds as object representations provides detailed spatial geometry and enhanced physical awareness, which supports robust reasoning across different viewpoints.

Furthermore, the framework is inherently extensible to diverse HOI scenarios. By adjusting point cloud configurations, it readily handles multiple objects, complex geometries, or articulated structures. For example, adding objects only requires including additional point clouds, and articulated objects (e.g., drawers) can be decomposed into rigid sub-components modeled as separate point clouds. Increasing point density captures fine details in deformable or complex shapes. This flexibility highlights the method’s practical potential for modeling varied real-world interactions.

## 4.5 Ablation Study

**Effectiveness of Synchronized Diffusion.** We argue that integrating visual priors and physical dynamics into a synchronized diffusion process is essential for HOI video and motion generation. To validate our synchronized diffusion mechanism: (1) We first remove VID to avoid confounding factors (**SViMo w/o VID**). (2) Then we decompose it into two independent components: a motion generation model with only motion loss and a video generation model conditioned on groundtruth motion (**VModel w/ GT Mot. Guid.**). After training these models independently, we use the predicted motions from the former as conditions for the latter during inference (**VModel w/ Pred. Mot.**). The last two rows of Tab. 3 show that modeling video and motion independently not only leads to a 3.88% decrease in video overall score (0.8719 vs. 0.8381), but also results in a 5.31% degradation in motion FID (0.0546 vs. 0.0575). This highlights the importance of our synchronous diffusion model in enabling feature-level synergy between video and motion. This highlights the advantage of integrating visual priors and motion dynamics for our method.

**Impact of Vision-aware 3D Interaction Diffusion.** The vision-aware 3D interaction diffusion model forms a closed-loop feedback and co-evolution mechanism with the synchronized video-motion diffusion, by injecting interaction guidance and gradient constraints into the latter. To validate its effectiveness, we conduct four variants: removing VID entirely (**SViMo w/o VID**), applying only gradient constraints (**SViMo w/ Grad. Cons.**), providing only interaction guidance (**SViMo w/ Inter. Guid.**), and preserving the full VID (**SViMo w/ VID**). Evaluation results in Tab. 3 and training loss curves in Fig. 7 show that direct interaction guidance slightly outperforms gradient constraints, while the complete VID achieves the best performance.

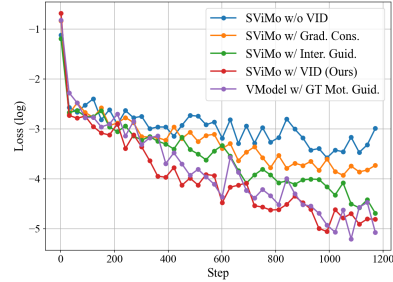


Figure 7: Video loss curves of different variants during the training process.

**Influence of 3D Data on Convergence Speed and Performance.** Our method simultaneously generates 2D videos and 3D motions. To evaluate the contribution of 3D motion data to training convergence and output quality, we conduct an ablation study in which 3D motion are replaced with 2D data (Ours-2D-VID).

The comparison is performed using three metrics: motion FID, video loss at 1K training steps, and overall video quality. As shown in Tab. 4, incorporating 3D motion diffusion leads to more plausible motions, faster convergence, and a higher overall video score. In contrast, the variant that completely removes the explicit motion generation module (SViMo w/o VID) exhibits further degradation in both convergence speed and final performance. This can be largely attributed to the fact that 3D motion representation enhances the model’s understanding of object scale and spatial occlusion relationships.

Table 4: Ablation analysis of the impact of 3d data on convergence speed and performance.

| Method        | Mo. FID ↓     | Vid. Loss @1K ↓ | Over. Qual. ↑ |
|---------------|---------------|-----------------|---------------|
| SViMo w/o VID | -3.4883       | -4.1912         | 0.8719        |
| Ours-2D-VID   | 0.0522        | -3.8265         | 0.8721        |
| Ours          | <b>0.0100</b> | <b>-4.9119</b>  | <b>0.8800</b> |

## 5 Conclusion

In conclusion, we propose a synchronized diffusion model that unifies hand-object interaction video generation and motion synthesis in a single diffusion process. By jointly modeling the co-evolution of appearance and motion, our method produces visually realistic and dynamically plausible results. A key component is the vision-aware 3D interaction diffusion model, which guides the denoising process via gradient constraints in a closed-loop pipeline, greatly improving video-motion consistency. The framework requires no predefined conditions and shows strong zero-shot generalization in real-world settings. It is also inherently extensible, easily adapting to diverse HOI scenarios, such as those with multiple, complex, or articulated objects, through simple adjustments to the 3D point cloud representation. This paradigm offers a promising direction for multimodal alignment and building world models that understand complex physical interactions.

## Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) No.62125107 and No.62272172.

## References

- [1] Tenglong Ao. Body of her: A preliminary study on end-to-end humanoid agent. *arXiv preprint arXiv:2408.02879*, 2024.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2024.
- [5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>, 3:1, 2024.
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [7] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024.
- [8] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.
- [9] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.
- [10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.
- [11] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thimo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15896–15908, 2025.
- [12] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5162–5171, 2022.
- [13] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024.
- [14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023.
- [15] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Giga-hands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17461–17474, 2025.



- [16] Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. Coohoi: Learning cooperative human-object interaction with manipulated object dynamics. *Advances in Neural Information Processing Systems*, 37:79741–79763, 2024.
- [17] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [18] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021.
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023.
- [21] 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.
- [22] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.
- [23] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7822–7832, 2023.
- [24] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024.
- [25] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.
- [26] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [27] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024.
- [28] Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, and Tae-Kyun Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 527–537, 2024.
- [29] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024.
- [30] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.
- [31] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024.
- [32] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*, 2024.
- [33] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [35] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025.
- [36] Siqi Liu, Yong-Lu Li, Zhou Fang, Xinpeng Liu, Yang You, and Cewu Lu. Primitive-based 3d human-object interaction modelling and programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3711–3719, 2024.
- [37] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.
- [38] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024.
- [39] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.
- [40] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [41] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omni-grasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems*, 37:2161–2184, 2024.
- [42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [43] Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. *arXiv preprint arXiv:2412.16212*, 2024.
- [44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [45] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023.
- [46] Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15061–15073, 2023.
- [47] 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.
- [48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [49] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- [50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.
- [51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

- [52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [53] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. *arXiv preprint arXiv:2409.08273*, 2024.
- [54] 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.
- [55] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.
- [56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.
- [57] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.
- [58] Rong Wang, Wei Mao, and Hongdong Li. Deepsimho: Stable pose estimation for hand-object interaction via physics simulation. *Advances in Neural Information Processing Systems*, 36:79685–79697, 2023.
- [59] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.
- [60] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physshoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.
- [61] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.
- [62] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023.
- [63] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. 2025.
- [64] Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37:52858–52890, 2024.
- [65] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024.
- [66] Ziyi Xu, Ziyao Huang, Juan Cao, Yong Zhang, Xiaodong Cun, Qing Shuai, Yuchen Wang, Linchao Bao, Jintao Li, and Fan Tang. Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation. *arXiv preprint arXiv:2411.17383*, 2024.
- [67] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022.
- [68] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.
- [69] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

- [70] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6639–6647, 2024.
- [71] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.
- [72] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *arXiv preprint arXiv:2409.09300*, 2024.
- [73] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022.
- [74] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2024.
- [75] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024.
- [76] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract clearly summarizes the core idea and highlights key innovations. The introduction explicitly states the main contributions, supported by theoretical analysis (Sec. 3) and experimental results (Sec. 4). The claims match the paper's findings, and the scope is well-defined.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper does not have a dedicated limitations section in the main body. We addressed potential limitations in the **appendix** due to space constraints, allowing the main text to focus on methodological details and results.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The paper presents the theoretical components including synchronized video-motion diffusion (SViMo), vision-aware 3D interaction diffusion (VID), and their closed-loop co-evolution mechanism in Sec. 3. This section describes the network architecture and loss functions. Additional training and inference pseudocode are included in the **Appendix**.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper provides a detailed experimental setup, including datasets, hyperparameters, and evaluation metrics. Reproduction steps are clearly outlined in the main text (Sec. 4). Both qualitative and quantitative results are presented for comparison. We commit to releasing the source code upon paper acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The TACO dataset [38] and the CogVideoX [69] foundation model we use are already open-sourced, and we commit to releasing the source code upon paper acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details in the experimental section (Sec. 4) and the **appendix**.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We follow the convention in prior works and report the performance number on the standard benchmarks.

Guidelines:



- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide sufficient information on the computer resources and other training details in Sec 4.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Due to space limitations, we discuss both potential positive societal impacts and negative societal impacts in the **appendix**.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly credit the creators or original owners of the assets (e.g., code, data, models) used in the paper, and explicitly mention and respect the corresponding licenses and terms of use in the references section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects requiring IRB Approvals.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## SViMo: Synchronized Diffusion for Video and Motion Generation in Hand-object Interaction Scenarios (Appendix)

In this supplementary material, we provide additional details on our methodology and experiments, and discuss the limitations, thereby enabling a more comprehensive and in-depth understanding of our proposed Synchronized diffusion for video and motion generation (SViMo). Below is the outline of all contents.

### A Pseudo-code for the Training and Inference Phases

In the main text, we introduced the Synchronized Video-Motion Diffusion  $\mathcal{G}_\theta$  (SViMo, Sec. 3.3), which focuses on video generation. We also presented the Vision-aware 3D Interaction Diffusion  $\mathcal{M}_\phi$  (VID, Sec. 3.4), designed for motion generation. Additionally, we described a closed-loop feedback mechanism between these two components (Sec. 3.5). Next, we will provide a more detailed description of the training (Alg. 1) and inference process (Alg. 2) using pseudocode. Note that our method does not require the rendered motion video as input during inference. We only require a reference image and a prompt as inputs, and we initialize the target outputs (2D video, 2D motion video, and 3D motion) with pure noise, and iteratively denoise them to obtain the final results. As detailed in Alg. 2, Line 4, denote the triple noisy latents at timestep  $t$  as  $\{z_t^V, z_t^M, (h_t, o_t)\}$ , we first apply single-step denoising using VID  $\mathcal{M}_\phi$  to produce the refined (but still noisy) 3D motion  $(\tilde{h}_0, \tilde{o}_0)$ . The refined 3D motions are then rendered back into a noisy 2D motion video latent  $\tilde{z}_0^M$ . Since  $\tilde{z}_0^M$  represents a more refined estimate than the original noisy latent  $z_t^M$ , we concatenated it with the latter as auxiliary information and input them into SViMo  $\mathcal{G}_\theta$ .

---

#### Algorithm 1 Joint training process of SViMo and VID.

---

**Input:** Reference image  $I$ , text prompt  $P$ , target video  $V$ , target 3D motion  $(h, o)$ , frozen video VAE encoder  $\mathcal{E}$ , SViMo network  $\mathcal{G}_\theta$ , VID network  $\mathcal{M}_\phi$ .  
**Output:** Optimized parameters of SViMo ( $\theta^*$ ) and VID ( $\phi^*$ ).  
1:  $\tilde{M} = \text{Proj}(h, o)$  ▷ rendered motion video projection  
2:  $z_0^V = \mathcal{E}(V)$ ,  $z_0^M = \mathcal{E}(\tilde{M})$ ,  $z_I = \mathcal{E}(I)$  ▷ calculate latent codes  
3: **while** not converged **do**  
4:    $t \sim \mathcal{U}\{1, \dots, T\}$  ▷ sample time step  $t$   
5:   Calculate diffused data  $z_t^V, z_t^M, h_t, o_t$  ▷ following Eq. 1  
6:    $(\tilde{h}_0, \tilde{o}_0) = \mathcal{M}_{\text{no-grad}}((h_t, o_t), (z_t^V, z_t^M), t)$  ▷ following Eq. 6  
7:    $\tilde{M} = \text{Proj}(\tilde{h}_0, \tilde{o}_0)$ ,  $\tilde{z}_0^M = \mathcal{E}(\tilde{M})$  ▷ direct interaction guidance following Eq. 7  
8:    $(\hat{z}_{0,\theta}^V, \hat{z}_{0,\theta}^M) = \mathcal{G}_\theta(z_t^V \oplus z_I, z_t^M \oplus \tilde{z}_0^M, P, t)$  ▷ inverse denoising  
9:    $(\hat{h}_{0,\phi}, \hat{o}_{0,\phi}) = \mathcal{M}_\phi((h_t, o_t), (\hat{z}_{0,\theta}^V, \hat{z}_{0,\theta}^M), t)$  ▷ indirect gradient constraint (Eq. 6, 7)  
10:    $\mathcal{L} = \omega_1 \mathcal{L}_{\text{SViMo}} + \omega_2 \mathcal{L}_{\text{VID}}$  ▷ calculate loss following Eq. 5, 6  
11:   update parameters  $\theta$  and  $\phi$  by gradient descent  
12: **end while**  
13: **return**  $\theta^* = \theta$ ,  $\phi^* = \phi$

---

### B More Implementation Details

#### B.1 VBench Metrics

We use some of the evaluation metrics from VBench [22] for quantitative analysis of the generated videos. The details of its implementation could not be fully presented in Sec. 4.1 of the main text due to page constraints, so we provide a supplement here.

**Subject Consistency** score is used to evaluate whether a subject (*e.g.*, a person or an object) remains consistent throughout a video. Specifically, we first extract image features from each frame of the video using DINO [6]. Then we calculate the cosine similarity between features of consecutive frames and between each frame and the first frame to characterize subject consistency. The

---

**Algorithm 2** Generation process of videos and motions.

---

**Input:** Reference image  $I$ , text prompt  $P$ , parameters of the noise scheduler  $\{\alpha_t, \beta_t\}_{t=1}^T$ , frozen video VAE encoder  $\mathcal{E}$  and corresponding decoder  $\mathcal{D}$ , trained SViMo network  $\mathcal{G}_\theta$  and VID network  $\mathcal{M}_\phi$ .

**Output:** Generated HOI video  $V$  and 3D motion  $(h_0, o_0)$ .

```
1:  $z_I = \mathcal{E}(I)$  ▷ calculate latent codes
2:  $z_T^V \sim \mathcal{N}(0, I), z_T^M \sim \mathcal{N}(0, I), (h_T, o_T) \sim \mathcal{N}(0, I)$  ▷ initialization
  for  $t = T, \dots, 1$  do
4:    $(\tilde{h}_0, \tilde{o}_0) = \mathcal{M}_{\text{no-grad}}((h_t, o_t), (z_t^V, z_t^M), t)$  ▷ following Eq. 6
      $\tilde{M} = \text{Proj}(\tilde{h}_0, \tilde{o}_0), \tilde{z}_0^M = \mathcal{E}(\tilde{M})$  ▷ direct interaction guidance following Eq. 7
6:    $(\hat{z}_0^V, \hat{z}_0^M) = \mathcal{G}_\theta(z_t^V \oplus z_I, z_t^M \oplus \tilde{z}_0^M, P, t)$  ▷ inverse denoising
      $(\hat{h}_0, \hat{o}_0) = \mathcal{M}_\phi((h_t, o_t), (\hat{z}_0^V, \hat{z}_0^M), t)$  ▷ indirect gradient constraint (Eq. 6, 7)
8:    $\mu_V = \left( \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} z_t^V + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \hat{z}_0^V \right)$ , same as  $\mu_M$  and  $\mu_{h,o}$ 
      $\sigma^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} (1-\alpha_t)$  ▷ constant value
10:   $z_{t-1}^V \sim \mathcal{N}(\mu_V, \sigma^2 I), z_{t-1}^M \sim \mathcal{N}(\mu_M, \sigma^2 I), (h_{t-1}, o_{t-1}) \sim \mathcal{N}(\mu_{h,o}, \sigma^2 I)$ 
  end for
12:  $V = \mathcal{D}(z_0^V)$  ▷ decode into raw video
  return  $V, (h_0, o_0)$ 
```

---

calculation formula is as follows:

$$S_{\text{Subj.}} = \frac{1}{T-1} \sum_{t=2}^T \frac{1}{2} (\langle d_1 \cdot d_t \rangle + \langle d_{t-1} \cdot d_t \rangle), \quad (8)$$

where  $d_t$  is the DINO feature of the  $t^{\text{th}}$  video frame.

**Background Consistency.** High-quality videos should not only ensure the subject’s appearance remains consistent throughout the video, but also should not overlook the consistency of the background, such as scenes, rooms, and tabletops. VBench [22] computes the background consistency score by utilizing the cosine similarity of CLIP [47] image features, with the calculation process being similar to Eq. 8:

$$S_{\text{Bkg.}} = \frac{1}{T-1} \sum_{t=2}^T \frac{1}{2} (\langle c_1 \cdot c_t \rangle + \langle c_{t-1} \cdot c_t \rangle), \quad (9)$$

where  $c_t$  is the CLIP feature of the  $t^{\text{th}}$  video frame. Notably, we crop the HOI region, following the common practice in HOI generation [24], because the global background context has a limited influence on fine-grained local interactions. Our experiments demonstrate that baseline models still exhibit hallucinated artifacts even with cropped regions. These artifacts, which include the spontaneous generation of objects, violate temporal coherence (see Fig. 9, rows 1-2). Therefore, this metric remains valuable for assessing video consistency.

**Temporal Smoothness** score is based on prior knowledge that video motion should be smooth, *i.e.*, linear or quadratic, over very short time intervals (a few consecutive video frames). To quantify the smoothness of a generated video  $V = \{v_1, v_2, \dots, v_{2T}\}$ , VBench [22] first removes odd-numbered frames  $(v_1, v_3, \dots, v_{2T-1})$ . Then it uses a video frame interpolation model [33] to generate the removed odd-numbered frame sequence  $(\hat{v}_1, \hat{v}_3, \dots, \hat{v}_{2T-1})$ . Next, it calculates the mean absolute error between the reconstructed frames and the original removed frames to measure whether the generated video satisfies the prior knowledge embedded in the video frame interpolation model. Finally, this value is normalized to the range [0, 1]:

$$S_{\text{TSmo.}} = \frac{1}{T} \sum_{t=1}^T \frac{255 - \text{MAE}(v_{2t-1}, \hat{v}_{2t-1})}{255}. \quad (10)$$

**Dynamic Degree.** The three metrics mentioned above, *i.e.*, Subject Consistency, Background Consistency, and Temporal Smoothness, can individually quantify the performance of generated videos

in certain specific aspects. However, a completely static video may achieve very high scores in them, which is not the desired outcome. To evaluate the dynamic degree of generated videos, VBench [22] employs RAFT [56] to calculate optical flow strengths between adjacent frames. The average of the top 5% highest optical flow intensities is taken as the video’s dynamic score. Once this score exceeds a predefined threshold  $\tau_{op}$ , the video is considered dynamic, otherwise, it is deemed static. The overall dynamic score is calculated as follows:

$$S_{\text{Dyn.}} = \frac{1}{N} \sum_{n=1}^N \mathbb{I} \left( \text{Avg} \left( \text{Top}_{t=2:T}^{5\%} (\text{RAFT}(v_{t-1}, v_t)) \right) > \tau_{op} \right), \quad (11)$$

where  $N$  is the number of generated videos.

## B.2 Training Loss for 3D Motions

We generate 3D motions using a vision-aware 3D interaction diffusion model. During training, we apply separate supervision for hand and object motions. The hand loss is computed as:

$$\mathcal{L}_{\text{hand}} = \left\| \mathbf{h} - \hat{\mathbf{h}} \right\|_2^2 + 0.2 \cdot \left\| \mathbf{h}' - \hat{\mathbf{h}}' \right\|_2^2 + 0.05 \cdot \left\| \mathbf{h}'' - \hat{\mathbf{h}}'' \right\|_2^2, \quad (12)$$

where the three terms represent the hand pose and its first-order and second-order differences between frames. For the 3D object point cloud sequence  $\mathbf{o} \in \mathbb{R}^{[B, N, K, 3]}$ , we first split the object into tool  $\mathbf{o}_{\text{tool}} \in \mathbb{R}^{[B, N, K/2, 3]}$  and target  $\mathbf{o}_{\text{target}} \in \mathbb{R}^{[B, N, K/2, 3]}$ , then calculate losses separately for these two components and average them to obtain the final object loss. Firstly, we directly compute the Chamfer distance between each frame’s generated point cloud and the ground truth to ensure consistency in shape contours. Furthermore, to enforce smoothness of the inter-frame motion, we first reshape the data into  $[B \times N, K/2, 3]$ , then calculate pairwise Chamfer distances between consecutive frames for both ground truth and generated motion, which captures the movement dynamic in the point cloud sequence. Finally, we compute the mean absolute error (MAE) between predicted values and ground truth motion dynamics:

$$\begin{aligned} \mathcal{L}_1 &= \text{Avg} \{ D_{\text{chamf}}(\mathbf{o}, \hat{\mathbf{o}}) \}, \\ \mathcal{L}_2 &= \text{Avg} \left\{ \frac{1}{N-1} \sum_{n=2}^N \text{MAE} (D_{\text{chamf}}(\mathbf{o}_n, \mathbf{o}_{n-1}), D_{\text{chamf}}(\hat{\mathbf{o}}_n, \hat{\mathbf{o}}_{n-1})) \right\}, \\ \mathcal{L}_{\text{obj}} &= \mathcal{L}_1 + 0.1 \cdot \mathcal{L}_2, \end{aligned} \quad (13)$$

where  $\mathbf{o}$  consists of  $\mathbf{o}_{\text{tool}}$  and  $\mathbf{o}_{\text{target}}$ .

## B.3 Pretrained 3D Interaction Reconstruction Model

To assess the difference between the distribution of the generated 3D interactions and that of the ground truth motions, we trained a 3D action reconstruction model on the TACO [38] dataset using the variational autoencoder architecture from Diverse Sampling [12]. During evaluation, we utilized features from the residual graph convolutional layers of this model to compute the FID score.

# C Additional Experimental Results

## C.1 More Ablation Studies

Our model is built upon the CogVideoX-5B [69] model and extended to a video-motion joint generation model. Additional parameters include the input and output projection layers, and triple modality modulation modules in all 42 DiT Blocks, resulting in a total of 6.02 billion parameters. To reduce training computational consumption, we tested two model variants. First, we added motion modality modulation modules only to even-numbered DiT Blocks while retaining only the original text and video modulation modules in odd-numbered DiT Blocks, *i.e.*, **Only Even Layers**. This design slightly reduced the total parameter count. Second, we applied **LoRA Training**, which uses low-rank decomposition of the original model parameters to significantly decrease the number of trainable parameters. Experimental results in Tab. 5 show that both variants produce lower-quality videos and motion compared to our default settings. This indicates that effective alignment and fusion of video and motion modality features are essential for generating high-quality outputs.

Table 5: Ablation studies on the synchronized diffusion and the vision-aware 3D interaction diffusion.

| Variants         | Content          |                 | Dynamic           |                 | Overall $\uparrow$ | Hand               |                     | Object             | HOI              |
|------------------|------------------|-----------------|-------------------|-----------------|--------------------|--------------------|---------------------|--------------------|------------------|
|                  | Subj. $\uparrow$ | Bkg. $\uparrow$ | Tsmoo. $\uparrow$ | Dyn. $\uparrow$ |                    | MPJPE $\downarrow$ | MSmoo. $\downarrow$ | Cham. $\downarrow$ | FID $\downarrow$ |
| Ours             | 0.9534           | 0.9546          | 0.9883            | <b>0.9784</b>   | <b>0.8800</b>      | <b>0.0121</b>      | <b>0.0053</b>       | <b>0.0019</b>      | <b>0.0100</b>    |
| Only Even Layers | <b>0.9535</b>    | <b>0.9552</b>   | 0.9884            | 0.9757          | 0.8783             | 0.0297             | 0.0103              | 0.0127             | 0.1368           |
| LoRA Training    | 0.9458           | 0.9523          | <b>0.9892</b>     | 0.9688          | 0.8631             | 0.0671             | 0.0211              | 0.0555             | 0.4671           |

Table 6: Generalization performance on more distant HOI scenarios.

| HOI Data for Evaluation                 | Overall Video Quality $\uparrow$ | 3D Motion FID $\downarrow$ |
|-----------------------------------------|----------------------------------|----------------------------|
| Ours (on diverse camera viewpoint data) | 0.8703                           | 0.0214                     |
| Ours (on completely unseen verbs)       | 0.8712                           | 0.0197                     |
| Ours (on TACO test data)                | <b>0.8800</b>                    | <b>0.0100</b>              |

## C.2 More Results of Our SViMo

**Qualitative Results of Our SViMo.** We present two additional videos and motion generation results as shown in Figure 8. These results demonstrate videos and motions that are both reasonable and consistent. Considering that image frames cannot effectively showcase video effects, we provide videos in the supplementary materials for a more vivid demonstration.

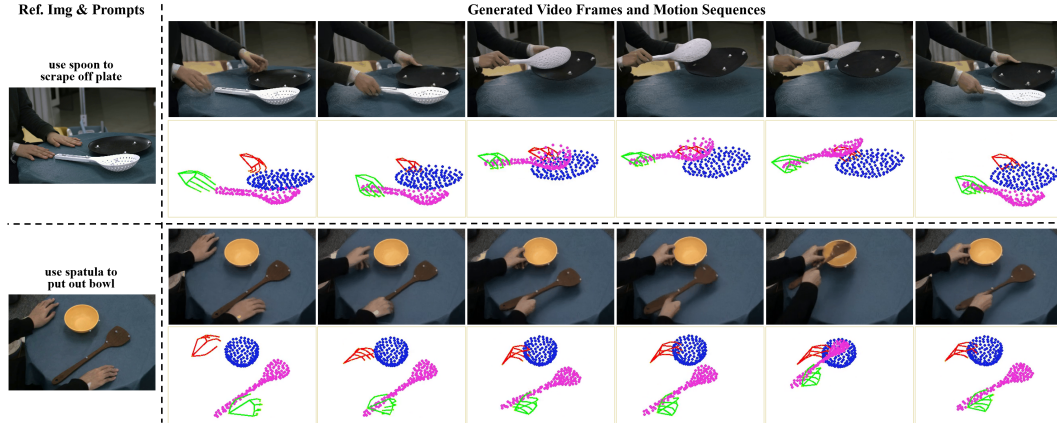


Figure 8: Visualization of the generated videos and motions. We provide vivid demonstrations in the video of the supplementary materials.

**Generalization to More Distant HOI Scenarios.** To enable a more rigorous evaluation of generalization capability, we curated an additional test set of 100 instances, consisting entirely of unseen human-object interaction (HOI) categories with novel verbs. These were sourced from TACO [38], OakInk2 [71], and GigaHands [15]. While TACO provides 12 camera views, we further enhanced viewpoint diversity by selecting approximately 70 instances from OakInk2 (4 views) and GigaHands (51 views). As shown in Tab. 6, both video and motion metrics on this unseen HOI set remain comparable to those on the main TACO test set. For the generation involving novel verbs, although a slight performance drop is observed, the results stay reasonably close to the main set.

## C.3 Comparisons of Our SViMo and Other Methods

**Diversity and Physical Metrics of 3D Motions.** We add three new evaluation metrics: diversity, penetration rate, and contact rate. The results in Tab. 7 show that our method achieves lower diversity scores than baselines. This occurs because baseline methods produce 3D motions with visible artifacts, including chaotic object point clouds and significant shape/scaling distortions relative to reference images (Fig. 4). Consequently, higher diversity scores don't necessarily indicate better



Table 7: Additional metrics of the generated 3D motion.

| Method    | Diversity $\uparrow$ | Penetration Rate $\downarrow$ | Contact Rate $\uparrow$ |
|-----------|----------------------|-------------------------------|-------------------------|
| MDM [57]  | <b>6.5308</b>        | 0.0904                        | 0.8107                  |
| EMDM [74] | 5.7847               | 0.0844                        | 0.8251                  |
| Ours      | 2.4135               | <b>0.0473</b>                 | <b>0.8328</b>           |

Table 8: The influence of motion information on baseline video models.

| Method                                         | Motion Availability | Motion Type            | Overall Video Quality $\uparrow$  |
|------------------------------------------------|---------------------|------------------------|-----------------------------------|
| Easy Animate [61]<br>CogvideoX-5B [69]         | w/o                 | w/o                    | 0.8330<br>0.8684                  |
| Animate Anyone [21]                            |                     | Input GT Motion        | 0.8209                            |
| Easy Animate [61]<br>CogvideoX-5B [69]<br>Ours | w/                  | Train with motion data | 0.8366<br>0.8698<br><b>0.8800</b> |

Table 9: Comparison on memory and time consumption.

| Method              | Training   |               |                   | Inference     |                          |
|---------------------|------------|---------------|-------------------|---------------|--------------------------|
|                     | Params (B) | Peak VRAM (G) | Time Per Step (s) | Peak VRAM (G) | Latency for 50 steps (s) |
| Animate Anyone [21] | 2.17       | 39.98         | 2.45              | 6.84          | 20.81                    |
| CogVideoX-5B [69]   | 5.57       | 56.32         | 11.61             | 20.59         | 48.35                    |
| Ours                | 5.82       | 61.52         | 18.06             | 23.35         | 71.31                    |

motion quality. Additionally, our method achieves superior performance in both penetration rate and contact rate metrics, indicating superior physical plausibility.

**Influence of Motion Information on Baseline Video Models.** Compared to baseline methods performing video generation alone, our approach jointly generates video and motion, and learns a visual-motion joint representation through adaptive multimodal modulation. To enable fair comparisons with baselines and assess the contribution of the modulation mechanism to our performance gains, We augmented baseline video models with the motion modality. Specifically, Animate Anyone [21] inherently uses Motion GT as input. For EasyAnimate [61] and CogVideoX [69], we trained them using both the original video and the rendered motion video. Results in Tab. 8 show that Animate Anyone produces videos of notably lower overall quality due to severe inter-frame flickering (see Fig. 3 in the main text and supplementary videos). For EasyAnimate and CogVideoX, while motion-aware training yielded marginal gains, improvements remained modest. This suggests that multimodal modulation fusion plays a more critical role in our approach than the motion modality alone.

**VRAM Consumption and Training/Inference Time.** We compared VRAM consumption and training/inference time with Animate Anyone [21] and CogVideoX-5B [69]. Results in Tab. 9 show that Animate Anyone requires significantly lower VRAM consumption and shorter time for both training and inference. This efficiency stems from its incremental design: it extends a 2D image generation model with temporal attention modules to enable video synthesis. In contrast, both CogVideoX and our method employ unified spatial-temporal full attention. Compared with CogVideoX, our approach simultaneously generates both 2D videos and 3D motion sequences, introducing only minimal additional parameters and maintaining computational cost within a reasonable range.

**Additional Comparisons of Video Generation Results.** We present additional video generation results from three more cases, as shown in Figures 9, 10, and 11. Other baseline methods’ generated results exhibit certain artifacts, such as flickering, distortion, and implausible movements, which are highlighted in red. Our method demonstrates superior performance compared to theirs.

#### C.4 Failure Cases

While our approach can generate high-realistic videos and plausible motions in most cases, it occasionally exhibits certain artifacts such as penetration, low dynamics, and object inconsistency, as



Figure 9: Comparison of video generation results. The artifacts in the videos generated by the baseline methods are highlighted in red. Refer to the video in the supplementary material for vivid demonstrations.

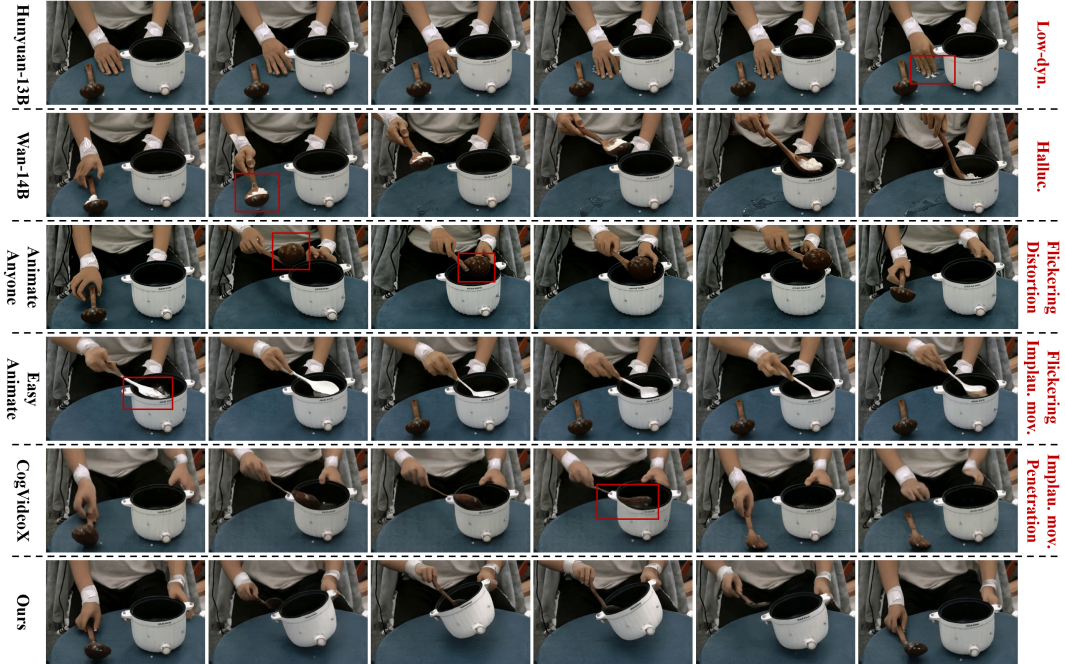


Figure 10: Comparison of video generation results. The artifacts in the videos generated by the baseline methods are highlighted in red. Refer to the video in the supplementary material for vivid demonstrations.

shown in Figure 12. However, the overall performance is still better than the baseline method. Addressing these issues, we need to explore stronger foundational models on one hand and leverage larger-scale video-motion datasets for training on the other hand.



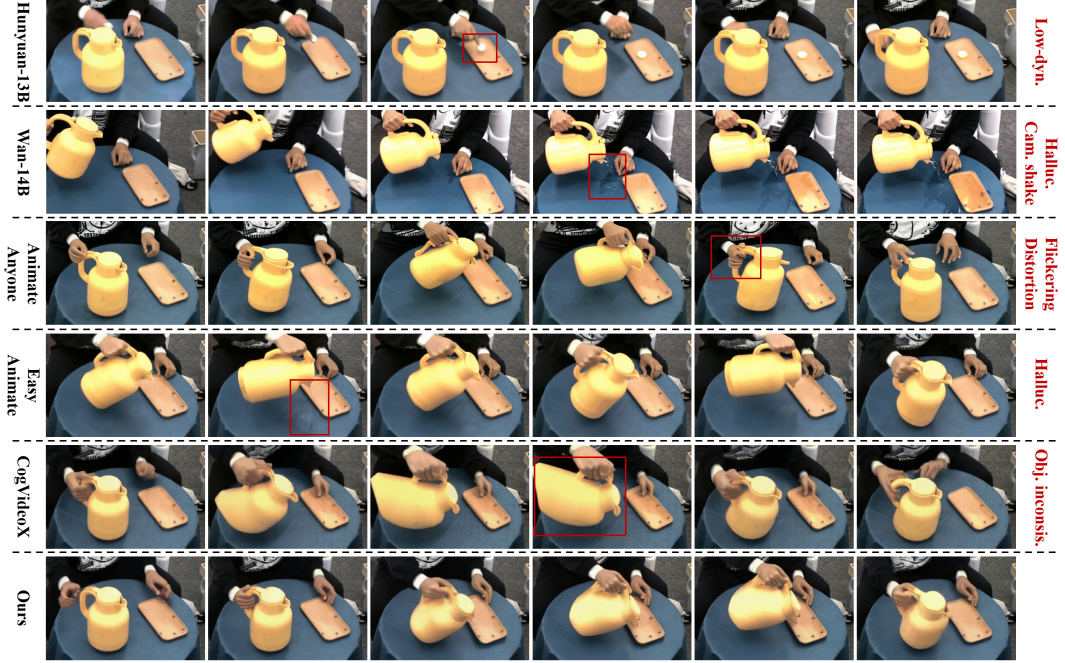


Figure 11: Comparison of video generation results. The artifacts in the videos generated by the baseline methods are highlighted in red. Refer to the video in the supplementary material for vivid demonstrations.



Figure 12: Visualization of failure cases. Red boxes highlight artifacts.

## D Limitations and Future Work

In fine-grained hand-object interaction scenarios, simultaneously generating visually high-fidelity videos and physically plausible motions remains a challenging task. While our synchronized diffusion model has made significant progress in addressing this challenge, three key limitations still exist. First, our method relies on a foundation model trained on large-scale video data, which is then fine-tuned on a smaller dataset of video-3D motion pairs. Although the latter has a relatively smaller

data scale, it remains essential for the expansion. Second, while our approach can generate diverse interaction motions, the produced 3D object point clouds are currently restricted to rigid, simple objects and struggle with structurally complex geometries. Third, the capabilities of the pre-trained foundation model directly impact both training efficiency and final performance. For instance, using lightweight LoRA (low-rank adaptation) strategies with the CogVideoX [69] foundation model results in suboptimal outcomes, even a full-parameter finetuned model has potential blurring artifacts when sampling at reduced resolution.

To address these challenges, future work should focus on three directions. First, replacing the non-differentiable 3D trajectory representation with differentiable neural representations (*e.g.*, NeRF-style formulations [42]) could enable video-only supervision without requiring explicit 3D annotations. This would transform our vision-aware 3D interaction diffusion model (VID) into a large reconstruction model, potentially resolving the second limitation. Second, continuous following of advanced open-source foundation models is necessary, as their evolving capabilities directly affect training stability and output quality. Third, integrating visual reinforcement learning strategies [40] could further enhance generation fidelity. These improvements would collectively advance the field toward more robust and scalable hand-object interaction synthesis.

## E Broader Impacts

Our proposed synchronized video-motion diffusion model generates hand-object interaction videos alongside corresponding motion sequences, achieving both visual realism and physically plausible dynamics. This overcomes the limitations of prior approaches that prioritized only visual quality or physical accuracy in isolation. The model demonstrates generalization capabilities to real-world scenarios, making significant progress in this research domain. Moreover, the joint diffusion paradigm offers valuable insights for cross-modal information integration in multimodal generative models. Practical applications span multiple fields, including game development, digital human animation, and embodied robotics, with potential societal benefits across these domains.

While acknowledging the positive impacts of this technology, we also recognize its dual-use nature. The system could be misused to generate synthetic media that violates privacy or perpetuates biases. To address these concerns, we commit to ethical development practices and transparent implementation protocols. Our research emphasizes proactive risk mitigation to ensure responsible innovation that maximizes social benefits while minimizing potential harms.