Re-HOLD: Video Hand Object Interaction Reenactment via adaptive Layout-instructed Diffusion Model

Yingying Fan[1†] Quanwei Yang[2] Kaisiyuan Wang[3*] Hang Zhou[3] Yingying Li[3]
Haocheng Feng[3] Errui Ding[3] Yu Wu[1*] Jingdong Wang[3]
[1] School of Computer Science, Wuhan University
[2] University of Science and Technology of China [3] Baidu Inc.

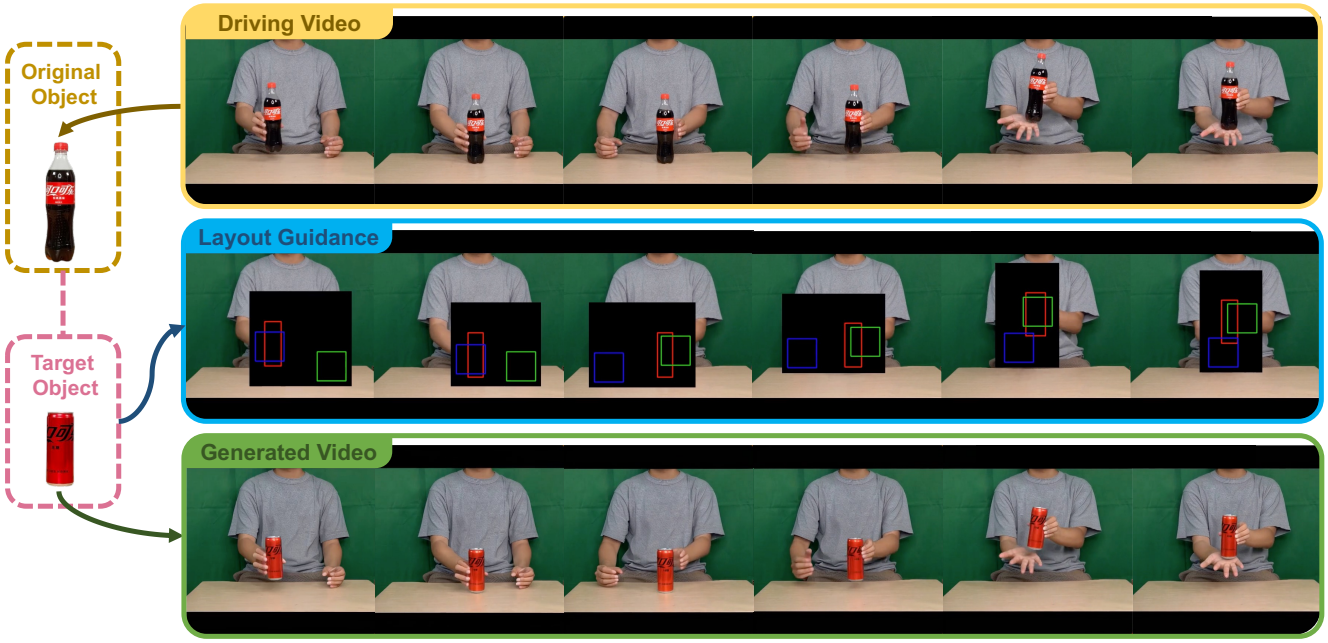

Figure 1. **Cross-Reenactment Results by our Re-HOLD framework.** Given a driving video and a target object, Re-HOLD can synthesize high-fidelity Human-Object Interaction (HOI) videos, even when the sizes of the target and original object differ significantly.

Abstract

Current digital human studies focusing on lip-syncing and body movement are no longer sufficient to meet the growing industrial demand, while human video generation techniques that support interacting with real-world environments (e.g., objects) have not been well investigated. Despite human hand synthesis already being an intricate problem, generating objects in contact with hands and their interactions presents an even more challenging task, especially when the objects exhibit obvious variations in size and shape. To tackle these issues, we present a novel video reenactment framework focusing on Human-Object Interaction (HOI) via an adaptive Layout-instructed Diffusion model (Re-HOLD). Our key insight is to employ specialized layout representation for hands and objects, respectively. Such representations enable effective disentanglement of hand modeling and object adaptation to diverse motion sequences. To further improve the quality of the HOI generation, we design an interactive textural enhancement module for both hands and objects by introducing two independent memory banks. We also propose a layout adjustment strategy for the cross-object reenactment scenario to adaptively adjust unreasonable layouts caused by diverse object sizes during inference. Comprehensive qualitative and quantitative evaluations demonstrate that our proposed framework significantly outperforms existing methods. Project page:

[†]Work done during an internship at Baidu Inc.
[*]Corresponding author

https://fyycs.github.io/Re-HOLD.

1. Introduction

With the rapid advancements in human video generation technology, digital human services have extended their reach into our daily routines (e.g., education, e-commerce, and multi-modal entertainment), leading to a notable rise in the demands of digital human videos. In response to these demands, numerous studies [14, 37, 40, 47, 50, 51, 54, 56, 58, 74] committed to 2D speaking animation have been proposed to create an interactive conversational experience.

On the other hand, human videos solely limited to lip movements or head movements synthesis fail to deliver a satisfactory user experience in real-world scenarios, which prompts the exploration of video generation for human body movements [4, 21, 23, 49, 53, 59, 73]. However, these approaches typically concentrate on holistic body motion modeling based on either 2D body poses [63] or implicit motion representation and take inadequate consideration of human hands, which are essential components in interactive scenarios. Although latest studies [13, 22, 75] continue to optimize hand synthesis by involving 3D hand mesh, they still cannot produce compelling results in more interactive scenarios with complex hand-object interaction (HOI). Nevertheless, HOI synthesis is a highly challenging research area, where even generating HOI images poses considerable difficulties, leading to video-level investigation particularly sparse. Thus how to achieve realistic HOI video synthesis remains an open problem. Previous image-level works [25, 66] built upon ControlNet [69] attempt to generate either an articulated hand or a full-body hand-grasping pose image for a given object. More recently, HOI-Swap [60] extends image-level HOI inpainting into a video-level framework by leveraging an additional stage for sequential frame warping. However, it can only produce object-centric videos with single-hand grasping operations and limited hand involvement.

Our primary focus is to devise a human-centric video generation system that enables reasonable HOI synthesis according to a source motion sequence and a target object. However, building such a system is non-trivial, since it entails three challenging problems: **1)** The physical interaction between hands and objects usually creates diverse occlusions, which leads to their intricate entanglement and easily causes artifacts at the hand-object interface. **2)** Both hands and objects exhibit high degrees of freedom and occupy only limited pixels in each video frame. Solely recovering either of them with detailed textures confronts a significant challenge. **3)** The non-negligible differences between distinct objects in shape and size inevitably affect the interacting position and degrade the realism of HOI, if the source motion sequence remains unchanged.

To tackle these problems, we propose a Video **Re**enactment framework for **H**and-**O**bject Interaction via **L**ayout-instructed **D**iffusion Model, namely **Re-HOLD**. Our key insight is to pursue hand-object disentanglement by involving specialized layout representations for hands and objects, respectively. Particularly, the layout representation for a video frame is composed of three detected bounding boxes, where two of them from hands (i.e., blue and green ones in Fig. 1) are shaped in a fixed size, while the rest (i.e., the red one in Fig. 1) has a varying size according to the object and depth. Notably, the hand layout representation exhibits *pose-invariance* and *size-invariance* providing merely positional information, which benefits the disentanglement between hands and objects. While such sparse layout representations may not achieve correct and fine-grained HOI synthesis, the disentanglement they provided allows us to introduce more representative instructions for better generation. Therefore, we further present a Hand-Object Interaction Restoration module to perform structure reshaping and texture refinement, where the former incorporates 3D hand meshes for better structural guidance and the latter relies on two independent memory banks and corresponding masks. Considering the gap between diverse objects under the cross-reenactment setting, we novelly design an adaptive strategy for layout adjustment at the inference stage, aiming to avoid producing unreasonable physical contact or interactive position. Extensive experiments demonstrate that our framework reenacts HOI videos with better fidelity than previous methods.

Our contributions are summarized as follows: **1)** We propose the first HOI reenactment framework for human-centric video generation which achieves realistic and reasonable HOI synthesis. **2)** Our proposed specialized layout representations for hands and objects along with our HOI Restoration Module, enable effective disentanglement and improved HOI modeling. **3)** Our proposed layout adjusting strategy is compatible with diverse objects, even ones with obvious gaps in shape and size, to generate reasonable interactions.

2. Related Work

2.1. Human Body Animation

Recent studies have leveraged powerful diffusion-based models to address the challenge of human body animation. One prominent approach involves using a UNet-based network [44] enhanced with cross-attention mechanisms [48] to inject additional information. PIDM [1] was the first to introduce classifier-free diffusion guidance for pose-guided human image generation, as further developed by [33]. DreamPose [23] utilizes UV maps as motion signals and performs conditional embeddings to achieve motion transfer. Similar concepts have been explored in other works

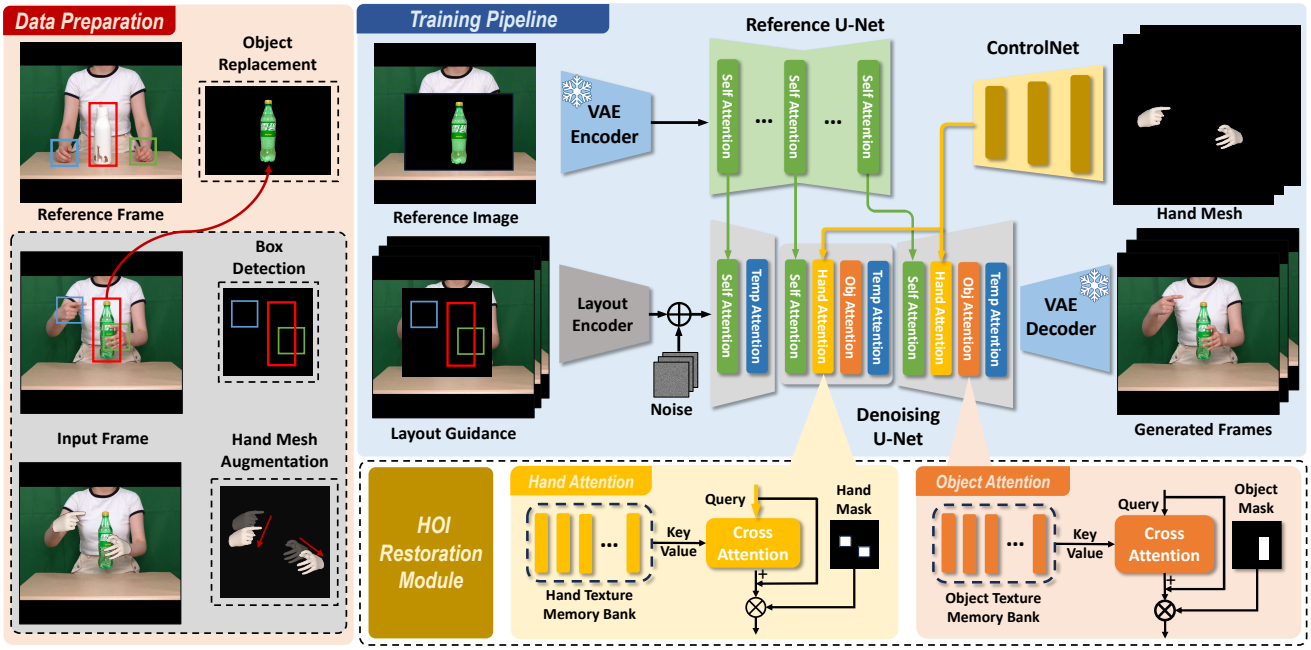

Figure 2. **Overview of our proposed Re-HOLD framework**. We propose a two-branch framework that consists of a Reference U-Net and a Denoising U-Net. The Reference U-Net takes a reference object image for object texture encoding while the denoising one takes noise latent and layout guidance as input for diffusion processing. To enhance the quality of HOI generation, we adopt the HOI Restoration Module for hand information and fine-grained object information restoration.

[4, 8, 13, 21, 53, 59, 61, 76]. RealisDance [75] takes DWPose [63], SMPL [2], and HaMeR [38] as input to generate realistic hand poses. While these approaches have yielded promising results, there are challenges when dealing with accurate HOI generations. Recently, InteractDiffusion [19] introduces the layout input to capture intricate interaction relationships. Inspired by this, we employ the sparse layout guidance as the driving signal and adaptively modify the hand-object interaction information during inference.

2.2. Video Generation and Editing

Numerous studies have focused on adapting pre-trained diffusion models for video generation and editing, including zero-shot [3, 10, 30, 41, 52, 62, 64] and one-shot-tuned [12, 28, 57, 78] learning frameworks. However, these approaches typically demand extensive per-video fine-tuning, which significantly hampers their practical application. Alternatively, other research [5, 7, 29, 32, 35, 39, 55, 67, 77] takes a training-based approach, where models are trained on large datasets to ensure they can serve as efficient editing tools during inference. However, none of these methods specifically address the adaptive generation of human-object interaction, and accurate HOI synthesis has always been a significant challenge in video generation tasks. In addition, a significant proportion of video generation and editing methodologies [3, 9–11, 28, 31, 34, 41, 42, 45, 46, 57, 62, 64] primarily depend on textual input for guidance during the editing process. However, textual prompts can sometimes fall short in accurately conveying the user's intentions. For our specific application, utilizing an object image as a form of guidance proves to be more precise and effective.

2.3. Human-Object Interaction

Given that human motion rarely occurs in isolation but rather within the context of objects or the surrounding environment, numerous methods [15–17, 71, 72] have been developed to explore the realistic integration of humans into scenes. A variety of methods have also focused on more fine-grained hand-object interactions [6, 20, 26, 65, 66, 68, 70]. DiffHOI [65] proposes a diffusion network to model the conditional distribution of geometric renderings of objects and leverage it to guide the novel-view rendering. GraspXL [68] unifies the generation of hand-object grasping motions across multiple motion objectives, diverse object shapes, and dexterous hand morphologies. Cg-hoi [6] focuses on generating realistic 3D human-object interactions from textual descriptions. Recently, HOI-Swap [60] presents a diffusion-based video editing framework for video object swapping with HOI awareness, ignoring the size and position change of hands and objects during object swapping. For practical digital human applications, we aim to generate two-hand human-object interactions using an adaptive layout-instructed diffusion model.

3. Method

In this section, we first describe our task formulation in Sec.3.1. Then we introduce the pipeline of our framework and its important components in Sec.3.2. The training strategy is described in Sec.3.3. The overview of our proposed method is shown in Fig.2.

3.1. Task Formulation

Task Formulation. HOI reenactment aims to generate reasonable interaction between hands and objects given a sequence of human motion signals and a target object $I_o \in \mathbb{R}^{H\times W\times 3}$. Here, sequential human motion signals in our Re-HOLD framework include layout guidances $V^l = \left\{I_1^l, I_2^l, \ldots, I_F^l\right\} \in \mathbb{R}^{F\times H\times W\times 3}$ and reconstructed hand meshes $V^h = \left\{I_1^h, I_2^h, \ldots, I_F^h\right\} \in \mathbb{R}^{F\times H\times W\times 3}$. The training for our framework is performed via a self-reconstruction manner, where the human motion signals and target objects are both from the source video $V = \{I_1, I_2, \ldots, I_F\} \in \mathbb{R}^{F\times H\times W\times 3}$. Our goal is to reconstruct V from random noise and these two conditional inputs.

At the inference stage, a reference image from another object $I_{o'}$ is provided to reenact the target video $V^{'}$. To guarantee the realism of the reenactment results, modification of the hand box and object box within the layout is conducted according to the difference between the two objects. Hand poses are kept unchanged to ensure that hands can effectively and adaptively interact with the new object.

3.2. Re-HOLD Framework Designs

Baseline Architecture. Correctly synthesizing hands, particularly fingers, is acknowledged as a quite challenging problem, which becomes even more difficult when it comes to generating in-contact objects and their interactions. Inspired by recent studies on human animation [4, 13, 21, 75], we initially formulated a baseline built upon a parallel-branch architecture to represent the human motion and the target object separately.

The upstream branch processes the reference image I_o of the target object to extract texture information by using a VAE encoder and a Reference U-Net. In the downstream branch, a Motion Encoder takes the human motion signals as input to integrate information about motion and structures. The subsequent Denoising U-Net effectively combines the extracted object texture and motion to predict noise intensity. Additional temporal attention layers are employed to improve temporal coherence. These two branches interact with each other through cross-attention mechanisms. Particularly, we follow [13, 75] to involve 3D hand meshes as the motion signals for better hand structure recovery due to the sufficient HOI information they provided (e.g., position, hand pose, and hand size).

However, we achieved two interesting observations under the cross-object reenactment setting: **1)** Generated objects fail to preserve their original structures and textures and hand synthesis suffers from severe deformation and distortion. **2)** Objects with an obvious gap in shape or size may result in physically unreasonable interactions or grasping positions. Therefore, we have conducted extensive explorations of motion instructions, hand-object restoration, and effective strategies for inference.

Layout Instruction. A possible explanation for the first observation is that object synthesis lacks effective guidance, causing variations in both shape and texture. In terms of the hand synthesis distortion, we speculate that the hand synthesis is strongly bonded with the positional instructions provided by hand meshes, thus the model tends to synthesize HOI scenes with similar hand positions, leading to the degradation in the final results.

To address these issues, we involve a set of layout representations composed of one bounding box for the object and two for the hands. The object box can compensate for the position and size instructions that the reference image cannot provide. The hand boxes are detected by a 2D key-point estimation approach [63], while the object box is derived from the object mask produced by the segmentation model [27]. Particularly, the hand boxes are limited to squares with a fixed size, which constructs a pose-invariance and size-invariance instruction for hands. Such hand instructions enable positional information disentanglement from the motion signals and provide basic interactive information for HOI synthesis.

To reduce computational complexity, we utilize a lightweight network as the layout encoder, comprising 4 convolution layers initialized with Gaussian weights, with the final projection layer using zero convolution. The layout feature $\mathbf{F}_l$ extracted by the layout encoder is then combined with the Gaussian noise ϵ_t and fed into the Denoising U-Net as the noisy latent.

Hand-Object Interaction Restoration Module. Although the layout instructions already provide HOI information, it is not sufficient for satisfactory hand-object recovery. To generate accurate and high-quality hand gestures, we reuse 3D hand meshes V^h reconstructed by HaMeR [38]. As mentioned above, in order to eliminate the over-reliance on hand positions, we apply simple augmentation on the positions of the hand meshes during training. As illustrated in Fig.2, the augmentation involves randomly shifting both hands in any direction from their original positions.

To capture robust and accurate hand pose information, we first utilize a ControlNet-like network to encode the hand mesh since it incorporates spatial and contextual cues into the generation process. The encoding process can be described as follows:

$$\mathbf{F}^h = \boldsymbol{C}\left(\mathbf{z}_t \mid I^h, t, \theta^{\mathrm{h}}\right), \tag{1}$$

where $\boldsymbol{C}\left(\cdot, \theta^{\mathrm{h}}\right)$ denotes the ControlNet [69], $\mathbf{z}_t$ is the noise

latent diffused at timestep t, $\mathbf{F}^h$ is a set of features output by the down blocks and middle blocks of ControlNet.

It is widely recognized that generating human hands and object textures is a challenging task [42]. We contend that relying solely on aligned hand and object features is insufficient for accurately recovering details. Thus, we propose two global memory banks to restore diverse hand poses and object information, respectively. Specifically, we develop a Hand-Object Interaction Restoration module for generating human hands and fine-grained object textures. This is achieved by constructing separate learnable memory banks for hands and objects: $\mathbf{B}_h \in \mathbb{R}^{N_h \times C_h}$ and $\mathbf{B}_o \in \mathbb{R}^{N_o \times C_o}$.

Alongside these two memory banks, we also design corresponding Hand-Attention and Object-Attention layers integrated into the U-Net architecture. The Hand Memory Bank effectively restores human hand textures, while the Object Memory Bank is designed to store object textures during training. The attention mechanism for Hand Attention and Object Attention can be defined as follows:

$$\mathbf{F}^a = \mathrm{Att}(\mathbf{F}, \mathbf{B}, \mathbf{B}) * M + \mathbf{F}, \tag{2}$$

$$\mathrm{Att}(\mathbf{F}, \mathbf{B}, \mathbf{B}) = \mathrm{Softmax}\left(\frac{(\mathbf{W}_Q \cdot \mathbf{F})(\mathbf{W}_K \cdot \mathbf{B})^\top}{\sqrt{d}}\right) \cdot (\mathbf{W}_V \cdot \mathbf{B}), \tag{3}$$

among them, $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ represent learnable weights for the cross-attention modules, $\mathbf{B}$ is the hand or object memory bank and M denotes the mask for the hand or object. For object attention, $\mathbf{F}$ is the Denoising U-Net feature. As for hand attention, $\mathbf{F} = \mathbf{F}^h$, $\mathbf{F}^a$ is then added to the Denoising U-Net feature after hand attention.

Adaptive Layout-Adjustment Strategy. In terms of the second observation, we develop an adaptive strategy for layout adjustment during cross-object reenactment to produce plausible HOI physical contact relationships. As illustrated in Fig.3, the process is divided into four steps: **1)** We initialize the centers on the four sides of each object box as potential contact points between the hand and the object. We identify the contact relationship between the hands and the box by calculating the distance from the center point of the hand box to the nearest contact point (named H2O distance). If H2O distance is less than a pre-defined threshold $\mathcal{T}$, the hand and the object are regarded as "in contact", otherwise they are not in contact. **2)** For each frame, we fix the center point of the object box and then adjust its height and width to match the size of the target object by calculating the adaptive ratio by frame. **3)** Then we horizontally adjust the position of each hand box to maintain the original H2O distance. **4)** Finally, we keep moving the object box until its bottom is consistent with the original box bottom. Following these four steps, we can effectively avoid floating objects and generate physically reasonable hand-object interactions, especially when handling objects with obvious gaps in size and shape.

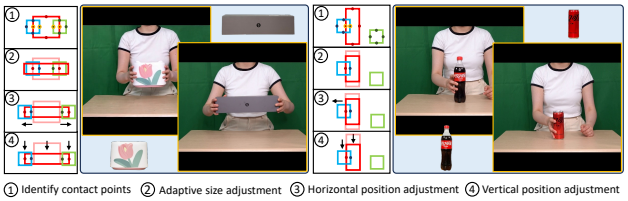

Figure 3. Schematic diagram of adaptive layout adjustment.

3.3. Training Objectives

Our framework is developed from Stable Diffusion (SD) [42], which employs an autoencoder to diffuse and denoise within the latent space. In the training phase, the diffusion process involves encoding the image to latent space $\mathbf{z}_0 = \mathcal{E}(\mathrm{x})$, and the random Gaussian noise is gradually added to $\mathbf{z}_0$ with T diffusion steps. Typically, the training object is to predict the added noise at various time steps through a learnable denoising U-Net, formulated as follows:

$$\mathbf{L} = \mathbb{E}_{\mathbf{z}_t, c, \epsilon, t}\left(\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, c, t\right)\|_2^2\right), \tag{4}$$

where ϵ_θ denotes the denoising U-Net, c is the conditional embeddings. During inference, $\mathbf{z}_T$ is sampled from random Gaussian distribution with the initial timestep T and is denoised back into $\mathbf{z}_0$. Finally, the decoder $\mathcal{D}$ reconstructs $\mathbf{z}_0$ to yield the generated image.

In our work, we adopt a two-stage training strategy to perform image-level HOI modeling and temporal HOI consistency modeling separately. The first stage is image-level HOI modeling, which focuses on accurately establishing HOI based on given image-level conditions. In this stage, we keep the VAE encoder fixed while training the remaining network components, excluding the temporal attention mechanism. Notably, towards the end of the first training phase, we place a special emphasis on hands and objects by only calculating the L1 loss of the corresponding region as the final loss every 10 iterations. In the second stage, we incorporate the temporal layer into the previously trained network to model the temporal consistency of the generated video frames. During this stage, the input is consecutive video frames and we only train the temporal layer while fixing the weights of the rest of the network.

4. Experiments and Results

4.1. Experimental Settings

Dataset. For our training, we collect a dataset consisting of 9 subjects with 14 objects. All videos are segmented into clips of 5 seconds each. For each subject data, we randomly select two objects that the subject has not seen as the test set. To enrich the diversity of the dataset, we follow HOI-Swap [60] to leverage a large-scale egocentric dataset HOI4D [36]

Dataset	Method	Cross-Reenactment			Self-Reenactment					
		hand fid. ↑	subj. cons. ↑	mot. smth. ↑	PSNR ↑	FID ↓	hand agr. ↑	hand fid. ↑	subj. cons. ↑	mot. smth. ↑
HOI4D	AnyV2V	0.183	0.591	0.952	28.821	185.961	0.206	0.291	0.877	0.982
	VideoSwap	0.924	0.907	**0.992**	31.964	158.964	0.611	0.987	0.915	0.990
	HOI-Swap	0.994	0.911	0.990	31.528	30.152	0.754	0.993	0.902	0.988
	Re-HOLD	**0.994**	**0.915**	0.991	**31.984**	**26.583**	**0.826**	**0.993**	**0.916**	**0.991**
Ours	AnyV2V	0.934	0.829	0.983	30.166	116.084	0.223	0.981	0.931	0.992
	VideoSwap	0.936	0.922	0.992	32.903	100.840	0.625	0.983	0.943	0.993
	AnimateAnyone	0.983	0.950	0.991	32.611	26.361	0.698	0.990	0.951	0.992
	RealisDance	0.989	0.948	0.991	32.784	26.337	0.749	0.992	0.951	0.993
	HOI-Swap	0.994	0.944	0.994	31.634	30.932	0.725	0.992	0.949	0.994
	Re-HOLD	**0.994**	**0.955**	**0.994**	**33.451**	**19.021**	**0.773**	**0.993**	**0.953**	**0.995**

Table 1. Quantitative results of our approach compared with SOTAs. 'Cross-Reenactment' means the target object is different from the original object while 'Self-Reenactment' is the self-reconstruction result on the test set.

for training. To meet our task setting, we only select videos that include a single object in HOI4D.

Implementation Details. For data pre-processing, we first crop out the hand-object interaction region excluding the human face based on the key points of the DWPose [63]. We then input cropped frames into HaMeR [38] to predict the MANO [43] model parameters for rendering the 3D hand mesh. The hand box is also extracted by DWPose [63] for the layout construction. We utilize LISA [27], a language-guided segmentation model for the extraction of the object mask. All video clips are pre-processed at a frame rate of 25 FPS with 512×512 resolution. The hand and object feature bank size is empirically set to 512. Since the size of the hand pose varies, while the hand box size remains the same, so we set $\mathcal{T}$ to half the size of the hand box in addition to 20. Following AnimateAnyone [21], we initialize the Denoising Branch and Reference Branch using Stable Diffusion V1.5 parameters. During inference, we use a DDIM sampler for 30 denoising steps. All experiments were completed on 4 A800s with a learning rate of 1e-5. For the first training stage, the batch size is 48 and F is 1. The training step is 100k, which takes about three days. For the second stage, batch size and F are set to 1 and 24 respectively with 50k training steps, taking about 2 days.

Evaluation Metrics. To demonstrate the effectiveness of our method, we comprehensively measure the self-reenactment results as well as the cross-reenactment results. Following HOI-Swap [60], we employ the HOI hand agreement to measure spatial alignment in the hand region, hand fidelity, subject consistency, and motion smoothness to evaluate general video quality. We exclude the HOI contact agreement metric due to the incorrect detection of objects. We also measure the quality of generated images from pixel space and feature space using PSNR and FID [18]. Hand agreement score, PSNR, and FID are only used for evaluating self-reenactment results due to the lack of ground truth in the cross-reenactment setting.

4.2. Comparison with Other Methods

Quantitative Results. For a more thorough comparison, we conduct two experimental settings, including self-reenactment and cross-object reenactment. Self-reenactment is performed only on the test set, which consists of unseen objects for each person. For cross-object reenactment, we displace the objects of the training set to the ones in the test set. The quantitative results of our methods compared with SOTAs are shown in Table 1. AnyV2V [24] and VideoSwap [12] are the state-of-the-art video editing methods. AnimateAnyone [21] and RealisDance[75] focus on generating human motions without interacting with objects. HOI-Swap [60] aims to swap the objects with HOI awareness. As illustrated in Table 1, Our method achieves top performance in both self-reenactment and cross-object reenactment across most metrics, proving its effectiveness. For instance, our method significantly outperforms other methods in PSNR and FID metrics, demonstrating the superiority of our approach in image generation. In addition, we obtain the best hand fidelity and hand agreement metrics indicating that Re-HOLD can synthesize accurate hand poses. Additionally, our approach maintains the best subject consistency, ensuring high-fidelity object textures in generated videos. Benefiting from the adaptive layout strategy, we can generate appropriate HOI details during inference.

Qualitative Results. For qualitative comparison, we provide results under both self-reenactment and cross-object reenactment settings. As shown in the left part of Fig.4, our method can generate realistic object texture though it is unseen by the person while other approaches fail to do so. Note that cross-object reenactment is more challenging due

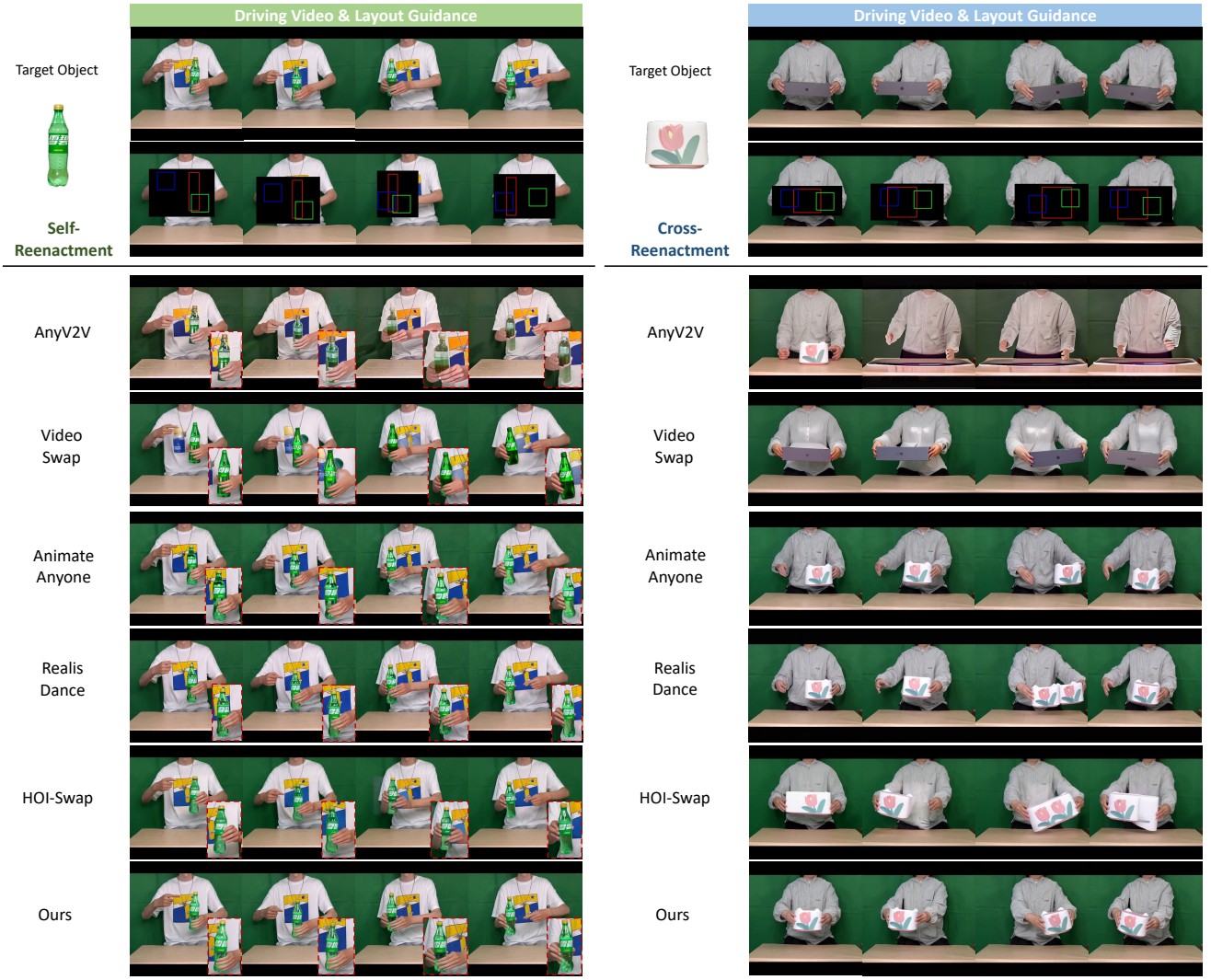

Figure 4. **Qualitative results compared with other methods.** Our approach achieves high-fidelity HOI details and satisfactory image quality in both self-reenactment and cross-object reenactment settings.

to the varied object shapes and sizes, but we still achieve proper HOI details and satisfactory image quality as indicated in the right of Fig.4. The results of the other method either appear as two objects in the image or generate another one instead of the target. In conclusion, our method fully exploits human-object interaction (HOI) information through adaptive layout guidance and the HOI Restoration Module, thereby enhancing video quality to meet the demands of various cross-reenactment scenarios.

Human Evaluation. We conduct a user preference study on our collected dataset to evaluate the performance of human-centric HOI video generation. There are 20 samples and 15 human voters in total. For each sample, we randomly present six video results generated with Re-HOLD and other SOTA methods to the human voter. The human voters are required to estimate the video results in three aspects: a) HOI Consistency: Does the video accurately reproduce the Human-Object Interaction in the driving video? b) Object Appearance Consistency: Does the object in the video have a consistent appearance with the target object? c) Temporal Consistency: How is the temporal coherence of this video? The rating score ranges from 1 to 5 and higher scores indicate better preference. The displayed result represents the ratio of the obtained score to the overall score. As shown in Table 2, our method achieves the highest scores compared with its counterparts.

4.3. Ablation Study

To better demonstrate the effectiveness of different components of our framework, we conduct experiments for all the

Method	HOI Consistency	Object Consistency	Temporal Consistency
AnyV2V	0.38	0.32	0.38
VideoSwap	0.52	0.22	0.44
AnimateAnyone	0.72	0.58	0.42
RealisDance	0.68	0.74	0.28
HOI-Swap	0.76	0.40	0.44
Re-HOLD	**0.92**	**0.92**	**0.88**

Table 2. User study of Re-HOLD and other SOTA methods.

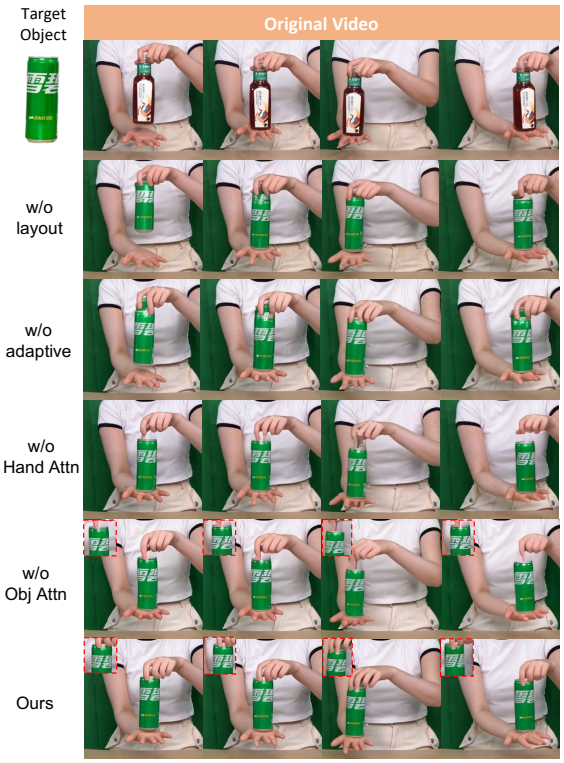

Figure 5. Qualitative ablation results of cross-reenactment when removing different components in our framework.

variants in our ablation studies, which are defined below. 1) "w/o layout": We directly use the 3D hand mesh renderings as the driving signal instead of the proposed layout guidance and the ControlNet is removed accordingly. 2) "w/o adaptive strategy": The adaptive layout adjustment strategy is not applied under the cross-reenactment setting. 3) "w/o hand attention": We remove the hand attention layers of the HOI Restoration Module from the Denoising U-Net. 4) "w/o obj attention": We remove the object attention layers of the HOI Restoration Module from the Denoising U-Net.

For quantitative results, we report the hand fidelity, and subject consistency metrics in Table 3. It is observed that hand attention can enhance the hand gesture quality, and

Variations	Self-Reenactment		Cross-Reenactment	
	hand agr. ↑	subj. cons. ↑	hand fid. ↑	subj. cons. ↑
w/o layout	0.753	0.950	0.993	0.950
w/o adaptive strategy	-	-	0.992	0.952
w/o hand attention	0.756	0.952	0.992	0.953
w/o obj attention	0.767	0.951	0.994	0.952
Ours	**0.773**	**0.953**	**0.994**	**0.955**

Table 3. Ablation study.

object attention brings more detailed information about objects. Our result in the subject consistency metric is higher than "w/o adaptive strategy" indicating the significance of this strategy. Note that the adaptive layout adjustment strategy is only performed in the cross-reenactment setting. Additionally, we present qualitative comparisons of cross-object reenactment to verify the effectiveness of the proposed module. As shown in Fig.5, the object is deformed without the layout guidance, demonstrating the validity of our proposed method.

5. Discussion and Conclusion

Conclusion. In this paper, we propose the video reenactment framework Re-HOLD, which achieves realistic and reasonable Human-Object Interaction (HOI) via an adaptive Layout-instructed Diffusion model. We first specialize in layout representations of hands and objects for effective hand-object disentanglement. Accordingly, we introduce a Hand-Object Interaction Restoration module to perform structure reshaping and texture refinement via two memory banks. To further bridge the gap between diverse objects under the cross-reenactment setting, we implement an adaptive layout adjustment strategy that enables the generation of plausible hand-object physical contacts. Both quantitative and qualitative assessments have demonstrated our framework's superiority over existing methods.

Limitations. Despite the success of our framework, we also recognize some limitations during the exploration. Our dataset is specifically designed to capture fundamental hand movements used for object display in live-streaming scenarios. As a result, our framework produces less satisfactory results when handling 3D object manipulation cases, such as generating a multi-view video of an object. This will be a focus of our future work.

Broader Impact. However, the potential for misuse of this technology is a significant concern. We will take measures to strictly monitor the content our model generates, restricting access to research-oriented applications only. We believe the responsible use of our model can foster positive societal development in academic research and everyday life.

Acknowledgment

This work was partially supported by the Yunnan provincial major science and technology special plan projects under Grant 202403AA080002 and the National Natural Science Foundation of China under grant 62372341.

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 2

[2] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *NeurIPS*, 2024. 3

[3] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *CVPR*, 2023. 3

[4] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *ICML*, 2024. 2, 3, 4

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

[6] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 3

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *CVPR*, 2023. 3

[8] Yingying Fan, Kaisiyuan Wang, Hang Zhou, Shengyi He, and Yu Wu. Rqtalker: Speech-driven 3d facial animation via region-aware vector quantization. In *ICASSP*, 2025. 3

[9] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *CVPR*, 2024. 3

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv*, 2023. 3

[11] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv*, 2023. 3

[12] Yuchao Gu, Yuchao Zhou, Gu, Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *CVPR*, 2024. 3, 6

[13] Jiazhi Guan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Shengyi He, Zhiliang Xu, Haocheng Feng, Errui Ding, Jingdong Wang, Hongtao Xie, et al. Talk-act: Enhance textural-awareness for 2d speaking avatar reenactment with diffusion model. *SIGGRAPH Asia*, 2024. 2, 3, 4

[14] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv*, 2024. 2

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *CVPR*, 2019. 3

[16] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, 2021.

[17] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH*, 2023. 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[19] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *CVPR*, 2024. 3

[20] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *NeurIPS*, 2022. 3

[21] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 2, 3, 4, 6

[22] Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *CVPR*, 2024. 2

[23] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, 2023. 2

[24] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv*, 2024. 6

[25] Patrick Kwon and Hanbyul Joo. Graspdiffusion: Synthesizing realistic whole-body hand-object interaction. *arXiv*, 2024. 2

[26] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. Lego: Learning egocentric action frame generation via visual instruction tuning. In *ECCV*, 2025. 3

[27] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 4, 6

[28] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. In *CVPR*, 2023. 3

[29] Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, and Forrester Cole. Generative omnimatte: Learning to decompose video into layers. *arXiv*, 2024. 3

[30] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *CVPR*, 2024. 3

[31] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018. 3

[32] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *CVPR*, 2024. 3

[33] Fangjian Liao, Xingxing Zou, and Waikeung Wong. Appearance and pose-guided human generation: A survey. *ACM Computing Surveys*, 2024. 2

[34] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *CVPR*, 2024. 3

[35] Xiao Liu, Xiaoliu Guan, Yu Wu, and Jiaxu Miao. Iterative ensemble training with anti-gradient control for mitigating memorization in diffusion models. In *ECCV*, 2024. 3

[36] 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 *CVPR*, 2022. 5

[37] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *TOG*, 2021. 2

[38] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024. 3, 4, 6

[39] Elia Peruzzo, Vidit Goel, Dejia Xu, Xingqian Xu, Yifan Jiang, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Vase: Object-centric appearance and shape manipulation of real videos. *arXiv*, 2024. 3

[40] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 2

[41] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *CVPR*, 2023. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 5

[43] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv*, 2022. 6

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[45] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *ACML*, 2024. 3

[46] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv*, 2022. 3

[47] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv*, 2024. 2

[48] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 2

[49] Kaisiyuan Wang, Lu Sheng, Shuhang Gu, and Dong Xu. Vpu: A video-based point cloud upsampling framework. *TIP*, 2022. 2

[50] Kaisiyuan Wang, Changcheng Liang, Hang Zhou, Jiaxiang Tang, Qianyi Wu, Dongliang He, Zhibin Hong, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Robust video portrait reenactment via personalized representation quantization. In *AAAI*, 2023. 2

[51] Kaisiyuan Wang, Hang Zhou, Qianyi Wu, Jiaxiang Tang, Zhiliang Xu, Borong Liang, Tianshu Hu, Errui Ding, Jingtuo Liu, Ziwei Liu, et al. Efficient video portrait reenactment via grid-based codebook. In *SIGGRAPH*, 2023. 2

[52] Ruoyu Wang, Yongqi Yang, Zhihao Qian, Ye Zhu, and Yu Wu. Diffusion in diffusion: Cyclic one-way diffusion for text-vision-conditioned generation. *ICLR*, 2023. 3

[53] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *CVPR*, 2024. 2, 3

[54] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *ICCV*, 2021. 2

[55] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2024. 3

[56] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv*, 2022. 2

[57] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *CVPR*, 2023. 3

[58] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv*, 2024. 2

[59] 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 *CVPR*, 2024. 2, 3

[60] Zihui Xue, Mi Luo, Changan Chen, and Kristen Grauman. Hoi-swap: Swapping objects in videos with hand-object interaction awareness. *NeurIPS*, 2024. 2, 3, 5, 6

[61] Quanwei Yang, Jiazhi Guan, Kaisiyuan Wang, Lingyun Yu, Wenqing Chu, Hang Zhou, ZhiQiang Feng, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Showmaker: Creating high-fidelity 2d human video via fine-grained diffusion modeling. *NeurIPS*, 2024. 3

[62] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023. 3

[63] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *CVPR*, 2023. 2, 3, 4, 6

[64] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, 2024. 3

[65] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *CVPR*, 2023. 3

[66] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 2, 3

[67] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv*, 2024. 3

[68] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *ECCV*, 2025. 3

[69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. 2, 4

[70] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *CVPR*, 2024. 3

[71] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *3DV*, 2020. 3

[72] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *CVPR*, 2020. 3

[73] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 2

[74] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 2

[75] Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and Fan Wang. Realisdance: Equip controllable character animation with realistic hands. *arXiv*, 2024. 2, 3, 4, 6

[76] 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 *ECCV*, 2024. 3

[77] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. *ICLR*, 2022. 3

[78] Zhichao Zuo, Zhao Zhang, Yan Luo, Yang Zhao, Haijun Zhang, Yi Yang, and Meng Wang. Cut-and-paste: Subject-driven video editing with attention control. *Neural Networks*, 2024. 3