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

GenHOI: Towards Object-Consistent Hand–Object Interaction with Temporally Balanced and Spatially Selective Object Injection

Xuan Huang[1,2]*, Mochu Xiang [1,3]*, Zhelun Shen[1]†, Jinbo Wu[1], Chenming Wu[1]
Chen Zhao[1], Kaisiyuan Wang[1], Hang Zhou[1], Shanshan Liu[1], Haocheng Feng[1], Wei He[1], Jingdong Wang[1]
[1]Department of Computer Vision Technology(VIS), Baidu Inc., China
[2]Shenzhen Campus of Sun Yat-Sen University, China [3]Northwestern Polytechnical University, China
shenzhelun@pku.edu.cn

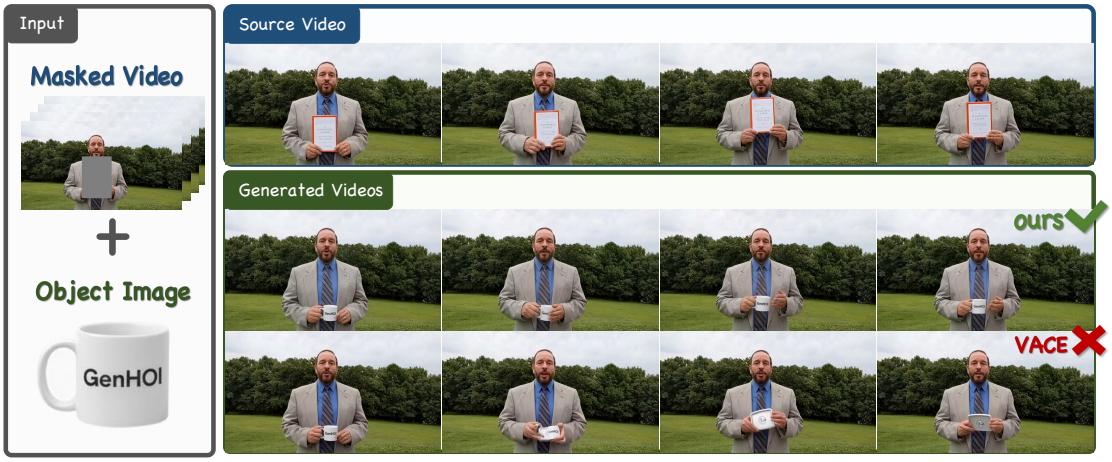

Figure 1. **Comparison with representative method.** All-in-one video editing models (*e.g.*, VACE) benefit from large-scale Internet training data, they still struggle to maintain object consistency across frames. In contrast, our method achieves both strong generalization and natural, visually consistent interactions between the human and the object. **We strongly encourage readers to refer to our website for additional video visualization results.**

Abstract

Hand–Object Interaction (HOI) remains a core challenge in digital human video synthesis, where models must generate physically plausible contact and preserve object identity across frames. Although recent HOI reenactment approaches have achieved progress, they are typically trained and evaluated in-domain and fail to generalize to complex, in-the-wild scenarios. In contrast, all-in-one video editing models exhibit broader robustness but still struggle with HOI-specific issues such as inconsistent object appearance. In this paper, we present GenHOI, a lightweight augmentation to pretrained video generation models that injects reference-object information in a temporally balanced and spatially selective manner. For temporal balancing, we propose Head-Sliding RoPE, which assigns head-specific temporal offsets to reference tokens, distributing their influence evenly across frames and mitigating the temporal decay of 3D RoPE to improve long-range object consistency. For spatial selectivity, we design a two-level spatial attention gate that concentrates object-conditioned attention on HOI regions and adaptively scales its strength, preserving background realism while enhancing interaction fidelity. Extensive qualitative and quantitative evaluations on unseen, in-the-wild scenes demonstrate that GenHOI significantly outperforms state-of-the-art HOI reenactment and all-in-one video editing competitors. Project Page: https://xuanhuang0.github.io/GenHOI/.

1. Introduction

Hand–Object Interaction (HOI) is a fundamental aspect of digital human content creation, particularly in domains such as online education and e-commerce. Although significant progress has been made in digital human generation, achieving realistic HOI remains a major challenge due to

*Equal contribution. Work was done when intern at Baidu.
†Corresponding authors. Zhelun Shen is the project lead.

the high variability in object shapes, human poses, and interaction orientations. Consider a scenario in which a designer aims to combine a product image into a real-world broadcast video, as illustrated in Fig. 1. Two objectives must be met to achieve a convincing result. First, the interaction between the hand and the object should appear natural and physically plausible. Second, the visual consistency of the reference object, such as color, texture, and logo, must be preserved across frames. Realizing these objectives is difficult because the shapes, poses, and orientations of objects vary widely, requiring the model to reason about physical contact and contextual semantics.

Recent HOI reenactment works have made encouraging progress. For example, HOI-Swap [53] extends image-level inpainting to the video domain by sequential frame warping, while Re-HOLD [7] introduces a layout-guided diffusion model for controllable HOI synthesis. However, these methods are typically trained and evaluated on in-domain data and fail to generalize to real-world in-the-wild scenes. All-in-one video editing models, such as VACE [18] exhibit stronger generalization due to their large-scale internet data pretraining. Nevertheless, they still struggle with HOI-specific challenges, e.g., maintaining object consistency. These limitations highlight the need for a model that generalizes robustly to unseen environments while ensuring both realism and identity preservation.

In this paper, we propose GenHOI to address the aforementioned challenges. GenHOI aims to improve HOI reenactment beyond source-domain datasets and extend it to challenging in-the-wild scenarios. To this end, we augment existing video generation models with a lightweight module designed for temporally balanced and spatially selective object information injection. This approach enables the network to learn the HOI task more effectively while minimizing degradation to its original video generation capabilities. For temporal balancing, we introduce *Head-Sliding RoPE*, which assigns head-wise temporal offsets to reference-object tokens. This method distributes their influence evenly across frames, thereby mitigating temporal degradation of the reference response when injected into video tokens during self-attention. For spatial selectivity, we introduce a spatial attention gate that controls where information flows and how strongly it is applied across the spatial dimension.

We validate our approach through comprehensive qualitative and quantitative evaluations, all performed on unseen scenes to ensure a fair assessment of generalization. The results show that our framework significantly outperforms state-of-the-art HOI reenactment methods and all-in-one video editing models, achieving higher realism and smoother hand motions and interactions. Figure 1 presents a visual comparison with our strongest competitor, VACE[18], confirming our clear advantage in object consistency. This improvement primarily stems from our design, which injects reference-object information in a temporally balanced and spatially selective manner. We summarize the contributions of this paper as follows.

- We introduce Head-Sliding RoPE, which assigns head-specific temporal offsets to reference tokens to balance their influence across frames, mitigating the temporal decay of 3D RoPE and improving object consistency in long video generation.
- We design a two-level spatial attention gate that concentrates object-conditioned attention on HOI regions and adaptively scales its strength, improving interaction fidelity while preserving background realism.
- Through the synergy of Head-Sliding RoPE and the spatial attention gate, we augment existing video generation models with a lightweight module that provides temporally balanced, spatially selective object information injection. This boosts hand–object realism and object-identity consistency while minimizing degradation to the base model's video generation capabilities.

2. Related Work

2.1. Human Body Animation

The rapid advancement of diffusion models has greatly accelerated progress in human body animation. A major line of research builds upon U-Net architectures [31] enhanced with cross-attention mechanisms [40] to integrate contextual or textural information into the generation process. Among early works, DreamPose [19] employs UV maps as motion signals and conditional embeddings for motion transfer, inspiring subsequent methods such as AnimateAnyone [16], DISCO [44], MimicMotion [62], MagicPose [5], ShowMaker [54], and TalkAct [12]. Recently, diffusion transformers (DiT) have demonstrated superior scalability, leading to DiT-based architectures for human animation, including HumanDiT [9], Human4DiT [33], UniAnimate [46], and DreamActor-M1 [25]. While human–object interaction (HOI) reenactment can be viewed as a specialized subtask of human animation, this general line of work still faces challenges in maintaining consistent human identity and object appearance across long video sequences. This highlights the need for models explicitly designed for the unique constraints of HOI reenactment.

2.2. Diffusion-based Video Editing and Inpainting

A substantial body of research has focused on adapting pretrained diffusion models for video editing and inpainting. Many frameworks [4, 8, 10, 11, 21–23, 27, 30, 34, 35, 49, 55, 56] rely on textual prompts for control. While convenient, this approach often lacks the precision needed to describe fine-grained object details or spatial relationships especially in complex Human–Object Interaction (HOI) sce-

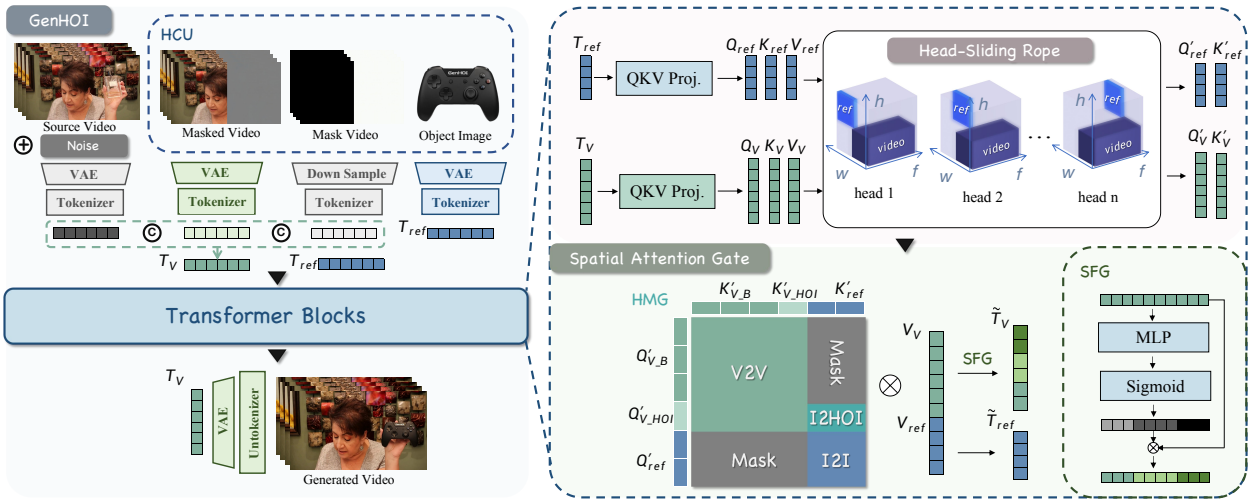

Figure 2. **Overview of the proposed framework.** The model integrates the HOI Condition Unit, Head-Sliding RoPE, and Spatial Attention Gate for temporally balanced and spatially selective HOI reenactment. HMG denotes hard mask gate and SFG is the soft flow gate.

narios. Recently, VideoPainter [3] further advances the field with a DiT-based dual-branch architecture that separates foreground generation from background preservation. Despite these advances, most methods still depend on text inputs, which can be ambiguous. Concurrently, VACE [18] was the first to introduce an all-in-one model for video creation and editing that supports diverse condition combinations (including image reference and inpainting), a paradigm similar to ours. Despite their powerful, general-purpose capabilities, these 'all-in-one' models still struggle with HOI-specific challenges, particularly maintaining object consistency and fidelity.

2.3. Human-Object Interaction

Alongside recent advances in broad human-centric tasks like head-body animation and lip-syncing [13, 24, 26, 38, 42, 43, 45, 47, 50, 63], research is also beginning to address the more fine-grained challenge of modeling HOI [6, 15, 20, 57–60]. These interactions are critical for achieving realistic human behavior in both static and dynamic settings. Among these, DiffHOI [57] employs a diffusion-based framework to model the conditional distribution of object renderings, thereby facilitating novel-view synthesis of HOIs. GraspXL [59] offers a unified framework for generating hand-object grasping motions across a range of object geometries and hand morphologies. Cg-HOI [6] focuses on generating plausible 3D human-object interactions from textual descriptions, enhancing controllability in semantic-to-motion generation. HOI-Swap [53] extends static image-level HOI inpainting to a video-based framework by incorporating sequential frame warping. Although these models achieve impressive performance on in-domain datasets, they often struggle to generalize to in-the-wild scenarios. This limitation largely stems from the heavy reliance on task-specific priors and extensive modifications to the base foundation models.

3. Proposed Method

3.1. Overview

Given a source video $\mathbf{V} = \{I_1, \ldots, I_F\} \in \mathbb{R}^{F\times H\times W\times 3}$ and an object reference image $I_{\text{ref}} \in \mathbb{R}^{H\times W\times 3}$, our goal is to reenact realistic hand–object interactions (HOI) between hands and the target object. Our framework is trained in a self-supervised reconstruction manner, where the reference video $\mathbf{V}_{\text{r}} \in \mathbb{R}^{F\times H\times W\times 3}$, the binary mask video $\mathbf{V}_{\text{mask}}$ and the reference image I_{ref} are derived from the original source video $\mathbf{V}$. The training objective is to reconstruct $\mathbf{V}$ from random noise $\mathbf{X}_{\text{rand}}$ conditioned on V_r, I_{ref} and $\mathbf{V}_{\text{mask}}$, formulated as:

$$\mathbf{V} = \mathcal{D}(\mathcal{M}(\mathbf{X}_{\text{rand}}, \mathcal{E}(I_{\text{ref}}), \mathcal{E}(V_r), \psi(\mathbf{V}_{\text{mask}}))), \quad (1)$$

We employ a pretrained VAE encoder–decoder $(\mathcal{E}, \mathcal{D})$ to compress videos into latent representations and reconstruct outputs. A DiT-based denoiser $\mathcal{M}$ predicts iterative denoising. The operator ψ denotes an average pooling operation.

To adapt a general video generation model to the HOI reenactment task, we introduce a HOI Condition Unit that injects HOI-related cues in a manner consistent with the original input organization. Next, built upon the DiT framework, we propose Head-Sliding RoPE and a spatial attention gate to inject reference object information in a temporally balanced and spatially selective manner. During inference, a novel object reference image is supplied to reenact the target video. An overview of the pipeline is shown in Fig. 2. We describe the HOI Condition Unit in Sec. 3.2 and the temporally balanced, spatially selective attention mechanism in Sec. 3.3.

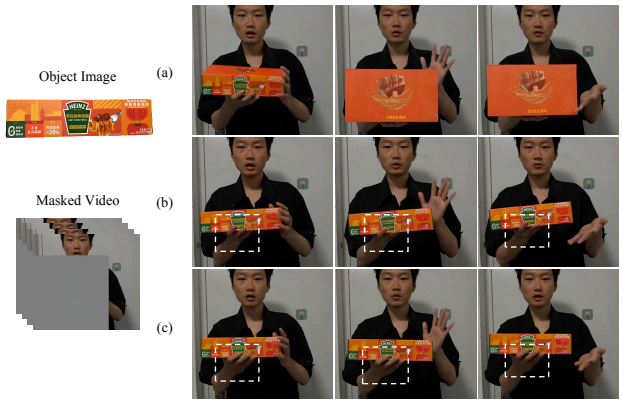

Figure 3. Visualization comparison between different reference object injection methods: (a) only HOI condition unit (HCU); (b) HCU + ref-in-bbox conditioning; (c) HCU + the proposed Attention.

3.2. HOI Condition Unit

To adapt a pretrained video generation model to the HOI reenactment task without introducing additional network branches or parameters, we propose a HOI Condition Unit (HCU). This module injects HOI-relevant cues directly into the latent input stream, preserving the model's original generative capacity while enabling effective HOI-specific learning.

Given an input video $\mathbf{V}$ and its corresponding binary mask video $\mathbf{V}_{\text{mask}}$, we first construct the reference video $\mathbf{V}_{\text{r}}$ that defines the inpainting regions:

$$\mathbf{V}_{\text{r}} = \begin{cases} \mathbf{V}, & F = 0, \\ (1 - \mathbf{V}_{\text{mask}}) \cdot \mathbf{V} + \mathbf{V}_{\text{mask}} \cdot \lambda, & F > 0, \end{cases} \tag{2}$$

where $\mathbf{V}_{\text{mask}}$ identifies the hand–object interaction (HOI) regions, λ is a constant, set to 127 in our network, which becomes 0 after normalization, and F denotes the frame index in the input video. We then project all inputs into the latent space and concatenate them along the channel dimension to form the input of the DiT model:

$$\mathbf{L}_{\mathbf{v}} = \text{Concat}(\mathbf{X}_{\mathbf{t}}, \mathcal{E}(\mathbf{V}_{\text{r}}), \psi(\mathbf{V}_{\text{mask}})), \tag{3}$$

where $\mathbf{X}_{\mathbf{t}}$ represents the noisy target video latent. In this way, we reformulate the video generation process as a *first-frame-conditioned video inpainting task*, where reference video $\mathcal{E}(\mathbf{V}_{\text{r}})$ provides background context and mask video $\psi(\mathbf{V}_{\text{mask}})$ tells the model where edits should occur. This formulation enables the model to better preserve background consistency while focusing generation on the HOI regions.

3.3. Temporally Balanced, Spatially Selective Attention

A key challenge of HOI is how to effectively inject reference object information into the model to ensure visual consistency across diverse object shapes and orientations. A straightforward approach is to incorporate this information from the first frame. However, as shown in Fig. 3(a), relying solely on the first frame often leads to noticeable object degradation over time.

Another commonly adopted strategy is the "ref-in-bbox" method, which directly pastes the reference object image into the inpainting region of $\mathbf{V}_{\text{r}}$:

$$\mathbf{V}_{\text{r}} = \begin{cases} \mathbf{V}, & F = 0, \\ (1 - \mathbf{V}_{\text{mask}}) \cdot \mathbf{V} + \mathbf{V}_{\text{mask}} \cdot \mathbf{I}_{\text{ref}}, & F > 0, \end{cases} \tag{4}$$

where $\mathbf{I}_{\text{ref}}$ denotes the object reference image. As illustrated in Fig. 3 (b), the "ref-in-bbox" approach can improve object appearance consistency in long video sequences. However, this encourages the model to treat the task as simple inpainting rather than interaction. Consequently, the model fails to generate natural interaction, instead just "pasting" the reference into the inpainting region.

In contrast, we propose Temporally Balanced, Spatially Selective Attention to address these issues. The overall pipeline is shown in the right part of Fig. 2. Specifically, *Head-Sliding RoPE* assigns head-wise temporal offsets to the reference object tokens, distributing their influence evenly across frames and mitigating temporal degradation of the reference response. Meanwhile, a *Spatial Attention Gate* focuses attention on interaction regions (e.g., HOI regions) while suppressing background interference.

3.3.1. Head-Sliding RoPE

Rotary Positional Embedding (RoPE) [36] is widely adopted in DiT architectures to encode spatiotemporal positional information. We extend the standard video diffusion 3D RoPE to construct our Head-Sliding RoPE, which is applied to both query and key embeddings within self-attention layers. For the query embeddings, our RoPE is defined as

$$\begin{bmatrix} Q'_{ref} \\ Q'_{V} \end{bmatrix} = \begin{bmatrix} Q_{ref} \\ Q_{V} \end{bmatrix} \odot [R_f, R_h, R_w] \tag{5}$$

where $Q_{\text{ref}} \in \mathbb{R}^{N \times D}$ and $Q_V \in \mathbb{R}^{M \times D}$ denote the query embeddings for reference-image tokens and video tokens, respectively. The rotary embedding components $R_f \in \mathbb{R}^{(N+M) \times D_1}$, $R_h \in \mathbb{R}^{(N+M) \times D_2}$, and $R_w \in \mathbb{R}^{(N+M) \times D_3}$ encode the frame, height, and width dimensions.

In settings involving both video tokens and conditional (reference object) tokens, prior works [37, 52] typically assigns conditional tokens to a separate coordinate space, following the formulation of [36], where $Q' = Qe^{jm\theta}$. Under this design, the formulation is as follows:

$$[R_f, R_h, R_w] = \begin{bmatrix} \begin{pmatrix} e^{j(-1)\theta} \\ \cdots \\ e^{j(-1)\theta} \end{pmatrix}_{N \times D_1}, & \begin{pmatrix} e^{j(h_I^0 + H_V)\theta} \\ \cdots \\ e^{j(h_I^N + H_V)\theta} \end{pmatrix}_{N \times D_2}, & \begin{pmatrix} e^{j(w_I^0 + W_V)\theta} \\ \cdots \\ e^{j(w_I^N + W_V)\theta} \end{pmatrix}_{N \times D_3} \\ \begin{pmatrix} e^{jf^0\theta} \\ \cdots \\ e^{jf^M\theta} \end{pmatrix}_{M \times D_1}, & \begin{pmatrix} e^{jh_V^0\theta} \\ \cdots \\ e^{jh_V^M\theta} \end{pmatrix}_{M \times D_2}, & \begin{pmatrix} e^{jw_V^0\theta} \\ \cdots \\ e^{jw_V^M\theta} \end{pmatrix}_{M \times D_3} \end{bmatrix} \tag{6}$$

while h_I^i and w_I^i denote the height and width indices of the i-th reference image token, and h_V^i and w_V^i denote the indice of the video tokens. Here, H_V and W_V represent the spatial lengths of video tokens in their respective dimensions. Although this design enables the network to distinguish between video tokens and object reference tokens, it also has an inherent limitation. Due to the nature of RoPE, larger positional distances result in weaker attention responses. Consequently, when conditional tokens are assigned a fixed frame index (e.g., -1), their influence on video tokens becomes uneven across the temporal dimension. The response is strongest at the earliest frames and weakest at the latest ones. This imbalance leads to degradation of object fidelity in HOI reenactment.

To overcome this issue, we revise the formulation into the proposed Head-Sliding RoPE:

$$[R_f] = \begin{bmatrix} \left(e^{j\left(\left\lceil \frac{N_f}{N_{head}} n_{head} \right\rceil\right)\theta}\right)_{N \times D_1} \\ \begin{pmatrix} e^{jf^0\theta} \\ \vdots \\ e^{jf^m\theta} \end{pmatrix}_{M \times D_1} \end{bmatrix} \tag{7}$$

where n_{head} denotes the n-th head in the multi-head attention mechanism, N_f is the total number of frames in the latent video representation, and N_{head} is the total number of attention heads, f^i is frame index. In this formulation, the frame index assigned to conditional tokens *slides* (Visualization in Fig.2 Head-sliding RoPE) across different attention heads, effectively averaging their attention responses over the entire temporal span of the video. This design balances the interaction between conditional and video tokens across frames while maintaining distinct spatial coordinates, thereby enabling the model to differentiate token types and preserve high-fidelity object appearance throughout the sequence.

3.3.2. Spatial Attention Gate

We formulate HOI reenactment as a video inpainting problem, where the main challenge is to concentrate modeling capacity on the hand–object interaction (HOI) regions rather than the comparatively easy background. To this end, we introduce a two-level spatial attention gate.

Hard Mask Gate Reference-object tokens are helpful for synthesizing hand–object interaction (HOI) regions but can be detrimental to background areas: within HOI regions, the model needs guidance from the reference tokens to determine which object to inject, whereas in background regions the reference video $\mathbf{V}_\mathrm{r}$ already provides sufficient context; excessive influence from the reference tokens may introduce artifacts.

To address this, we apply a hard mask gate that enforces information flow only passes from reference-object tokens

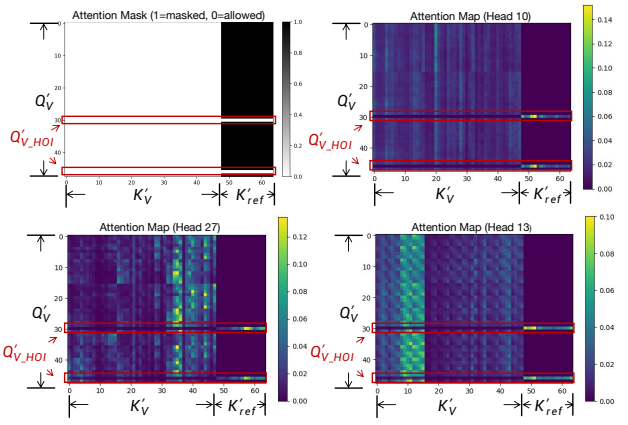

Figure 4. Visualization of the applied hard mask gate and the resulting attention maps. Multiple attention maps across different heads are shown, highlighting the pronounced effect of the proposed mechanism. The red box indicates the interaction between queries in HOI regions and keys from both video and reference object tokens.

to HOI tokens while preventing their influence on background tokens. Specifically, we denote the HOI-region and background-region video tokens as $T_{V_{HOI}}$ and T_{V_B}, respectively. For the mapped query, key, and value embeddings $Q = [Q_{V_{HOI}}, Q_{V_B}, Q_{ref}]$, $K = [K_{V_{HOI}}, K_{V_B}, K_{ref}]$, $V = [V_{V_{HOI}}, V_{V_B}, V_{ref}]$, where the subscripts V_{HOI}, V_B, and ref denote HOI, background, and object tokens, respectively, we introduce a binary mask M to control the information flow during attention computation:

$$M_{m,n} = \begin{cases} 0 & \text{if } m \in \{Q_{V_B}\} \text{ and } n \in \{K_{ref}\}, \\ 0 & \text{if } m \in \{Q_{ref}\} \text{ and } n \in \{K_{V_{HOI}}, K_{V_B}\}, \\ 1 & \text{elsewhere.} \end{cases} \tag{8}$$

Here, m and n index the query and key tokens, respectively. The final attention output is given by

$$T_{out} = \text{softmax}\left(\frac{M \odot QK^\intercal}{\sqrt{d_k}}\right) V. \tag{9}$$

This hard mask gate permits HOI queries to attend to object keys while blocking background queries from doing so, and it prevents object queries from attending back to video keys. Consequently, information from reference tokens flows into HOI regions (or remains self-regularized) without contaminating the background. See visualization in HMG (Hard Mask Gate) part of Fig.2. Fig. 4 visualizes the attention map $\text{softmax}\left(\frac{M \odot (QK^\top)}{\sqrt{d_k}}\right)$ after applying the hard mask gate. As shown, object keys yield zero responses to background queries, while producing strong activations for HOI-region queries, which are often even stronger than those induced by video keys. This observation confirms that the proposed hard mask gate effectively blocks information flow from the

reference object to background regions and encourages HOI queries to rely on reference object keys.

Soft Flow Gate. Beyond the hard truncation above, we add a soft, token-wise gate that dynamically modulates the updated video tokens. This design is inspired by the common practice in LLMs [28]. Specifically, we decompose the output as $T_{\text{out}} = [T'_v, T'_{ref}]$, where T'_v collects the updated video tokens. We then compute per-token gating coefficients via LayerNorm layer $\mathcal{LN}$ and FCN $\mathcal{F}$, followed by a sigmoid:

$$\begin{aligned} G_v &= \sigma\big(\mathcal{F}(\mathcal{LN}(T'_v))\big), \\ \tilde{T}_v &= G_v \odot T'_v, \end{aligned} \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid and $\odot$ denotes element-wise multiplication. The gated representation $\tilde{T}_v$ adaptively amplifies informative regions and suppresses redundant responses. Corresponding visualization is shown in SFG (Soft Flow Gate) part in Fig.2.

In conclusion, the *Hard Mask Gate* controls *where* information can flow, and the *Soft Flow Gate* scales *how strong* it should be based on content. Their complementarity defines the *Spatial Attention Gate*, which concentrates attention on HOI regions, protects the background, and injects reference cues robustly for high-fidelity reenactment.

4. Experimental Results

4.1. Implementation Details

Implementation. We implement the proposed method on top of the pretrained Wan-14B-I2V model [41]. Training is conducted on a dataset containing approximately 19,000 videos using 16× NVIDIA H100 (80 GB) GPUs for three days. The learning rate is set to 1×10^{-5}. During training, the reference object image and the first frame are extracted from the original video, while during inference, the reference image is provided by the user and the first frame is generated using image editing methods [1, 2].

Evaluation Dataset. We evaluate the performance of our approach on in-the-wild scenes using the AnchorCrafter dataset [51], which provides diverse human reenactment videos captured across various scenes and identities. To ensure high visual quality, we select videos with a resolution of at least 720p and reorganize the dataset to include 50 videos for self-reenactment and 50 videos for cross-reenactment. The resulting subset is referred to as AnchorCrafter_HOI.

Baselines. We conduct a comprehensive comparison between our method and several state-of-the-art (SOTA) approaches. The compared methods cover different lines of work, including the human body animation models UniAnimate-DiT [46] and MimicMotion [62], the all-in-one video editing framework VACE [18], and the HOI reenactment model HOI-Swap [53]. In the cross-reenactment setting, we ensure a fair comparison by supplying the same edited first frame to all methods that require an initial frame.

4.2. Evaluation Setting

We evaluate the proposed method in three settings:

Self-reenactment. In this setting, the model reconstructs the original video from a masked source video and the corresponding object image of the video. We assess visual fidelity using PSNR [32], SSIM [48], and LPIPS [61]; perceptual realism with FID [14] and FVD [39]; and object consistency via Object CLIP (OC) [17], which computes CLIP similarity [29] within the object bounding box.

Cross-reenactment. In this setting, the method replaces the object in the source video with a novel one absent from the original sequence. Visual quality is evaluated using FID [14] and FVD [39], and a user study is conducted due to the lack of ground truth.

User Study. We conducted a user study with 30 participants who evaluated 40 short videos and 20 long videos from AnchorCrafter_HOI. The participants rated the generated videos based on two criteria: (1) Reference fidelity with the reference image, (2) Video quality. Each criterion was assessed on a scale from 1 to 5, where 1 represents the lowest quality and 5 the highest. The user study was approved by the IRB.

4.3. Comparison with other methods

We compare GenHOI against state-of-the-art methods on both short and long video generation in Tab. 1.

Short video generation evaluation We first evaluate the short video generation capability of each method. As shown, HOI-Swap, the representative HOI reenactment method, fails to generalize to in-the-wild datasets, achieving only 24.29 PSNR in self-reenactment. In contrast, our method generalizes well and achieves state-of-the-art results across all metrics in both self- and cross-reenactment. Compared with the strongest competitor VACE, our method outperforms it on all metrics by a clear margin. This improvement mainly stems from our superior ability to maintain object consistency. User study results in Tab. 1 further confirm this claim, showing a clear advantage over VACE in reference fidelity (4.6 vs. 2.79).

We visualize cross-reenactment results in Fig. 5 (down). The proposed method successfully handles objects of various shapes and sizes, enabling transformations between large and small objects as well as different object types. For example, transforming a bag into a lipstick, a milk tea cup, or even a non-existent magical wand. Such flexibility is highly useful in practical applications, significantly reducing production costs in e-commerce scenarios. Fig. 5 (up) presents a visual comparison with existing methods. As illustrated, previous methods often fail to maintain temporal

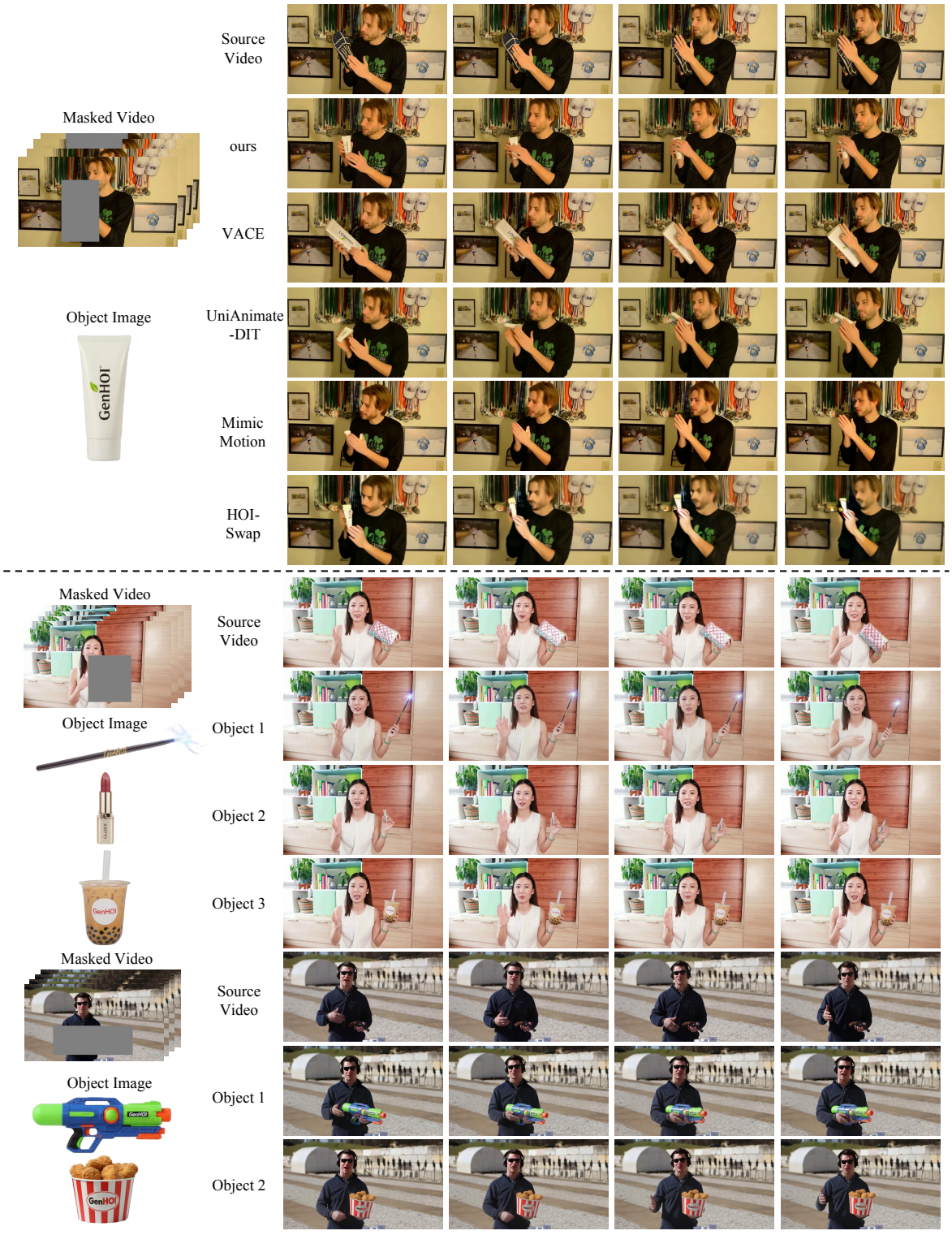

Figure 5. Up: Qualitative comparison with state-of-the-art methods. Down: Cross-reenactment results of the proposed method on in-the-wild videos, demonstrating robust and flexible object reenactment across various shapes, sizes, and categories.

Table 1. Quantitative comparison on the AnchorCrafter_HOI dataset. We evaluate performance under two settings: short video generation (81 frames) and long video generation (401 frames). VQ denotes video quality and RF denotes Reference Fidelity.

Setting	Method	Self-Reenactment						Cross-Reenactment		User Study	
		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	OC ↑	FID ↓	FVD ↓	VQ ↑	RF ↑
Short video generation	VACE	28.60	0.937	0.056	34.83	211.2	0.880	68.67	524.7	3.942	2.796
	UniAnimate-DiT	22.20	0.754	0.179	44.01	325.1	0.846	76.40	640.5	3.496	2.971
	MimicMotion	20.13	0.685	0.206	48.89	395.1	0.777	68.50	608.5	2.821	2.094
	HOI-Swap	24.29	0.843	0.173	50.67	352.1	0.787	76.72	570.5	1.475	1.201
	Ours	**31.71**	**0.952**	**0.036**	**11.53**	**67.95**	**0.937**	**58.83**	**429.5**	**4.487**	**4.636**
Long video generation	VACE	26.32	0.937	0.054	40.83	195.9	0.882	81.77	782.0	3.143	2.287
	UniAnimate-DiT	18.94	0.695	0.295	64.08	559.7	0.823	87.46	881.7	2.061	1.506
	MimicMotion	17.37	0.633	0.305	59.19	488.0	0.759	85.63	776.2	2.002	1.428
	Ours	**30.69**	**0.951**	**0.037**	**9.78**	**42.17**	**0.932**	**57.62**	**515.0**	**4.458**	**4.534**

Table 2. Ablation study on the components of proposed method. HS RoPE denotes head-sliding RoPE, SAG is spatial attention gate and FlF denotes first-last-frame condition.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	OC ↑
HCU	28.25	0.942	0.058	22.89	248.6	0.907
HCU + separate RoPE	29.73	0.945	0.050	22.66	223.8	0.908
HCU + ref-in-bbox	30.34	0.945	0.044	18.23	101.9	0.919
HCU + HS RoPE	30.88	0.951	0.039	17.92	103.9	0.915
HCU + HS RoPE + SAG	31.21	0.952	0.038	16.79	98.09	0.920
HCU + HS RoPE + SAG + FLF(full)	**31.71**	**0.952**	**0.036**	**11.53**	**67.95**	**0.937**

or object consistency, whereas our approach generates natural and visually coherent human–object interactions across videos.

Long video generation evaluation The proposed method achieves an even larger performance lead in long video generation. Compared with the second-best method, VACE, the PSNR gap in self-reenactment increases from 3.11 to 4.37, the FVD gap in cross-reenactment expands from 95.2 to 267, and the user study video quality gap rises from 0.5 to 1.3. These results further confirm the superiority of our method in producing temporally coherent and high-quality long videos.

4.4. Ablation Study

We perform various ablation studies to show the effectiveness of each component in our network. All experiments are reported on the AnchorCrafter_HOI dataset in self-reenactment setting. HCU denotes finetuning the video generation model with the proposed HOI condition unit, which serves as our baseline. Results are shown in Table 2. We also encourage readers to refer to the **supplementary material** for more comprehensive experimental results.

Head-Sliding RoPE vs. ref-in-bbox vs. 3D RoPE As the strategy for injecting reference-object information is central to our approach, we provide a detailed comparison of three alternatives. "Separate RoPE" denotes the separate rotary positional embedding used in Stand In [52] and OminiControl [37], while ref-in-bbox [7] follows the formulation in Eq. 4. As shown by our results, the proposed Head-Sliding RoPE improves PSNR by 1.15 dB over separate RoPE and by 0.54 dB over ref-in-bbox. Similar gains are observed across the other evaluation metrics. We attribute these improvements to Head-Sliding RoPE's ability to effectively average the attention responses of conditional tokens over the full temporal span of the video.

Impact of the Spatial Attention Gate. Adding the spatial attention gate increases PSNR from 30.88 to 31.12 dB and reduces FVD from 103.9 to 98.09 (lower is better), underscoring the necessity of imposing a spatial constraint.

Impact of the First–Last-Frame Condition. The first–last-frame (FLF) conditioning strategy is commonly used in video generation models [41]. In our framework, enabling FLF raises PSNR from 31.21 to 31.71 dB, indicating that anchoring the sequence with boundary frames benefits generation quality.

Model Parameter Analysis. Overall, the proposed module augments the pretrained video generation model with only 157M additional learnable parameters, accounting for approximately 0.95% of the original 16.5B-parameter model. This lightweight design offers two main advantages: (1) it largely preserves the original model's generalization ability, and (2) it introduces minimal computational overhead. Consequently, our model can achieve state-of-the-art HOI reenactment performance using only limited training data.

5. Conclusion

In this work, we present GenHOI, a novel framework designed to advance realistic hand–object interaction reenactment. Specifically, we introduce a HOI condition unit that adapts existing video generation models to the HOI reenactment task. Furthermore, we propose head-sliding RoPE and a spatial attention gate to achieve temporally balanced and spatially selective object information injection. Together, these mechanisms enable the model to maintain natural hand–object interactions and preserve object identity, even in unseen, in-the-wild scenarios. Comprehensive experiments demonstrate that GenHOI consistently outperforms existing HOI reenactment and all-in-one video editing methods in both qualitative and quantitative evaluations.

References

[1] Nano Banana – AI Image Editor. https://nanobanana.ai/, 2025. Accessed: 2025-11-13. 6

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 6

[3] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *SIGGRAPH*, 2025. 3

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *CVPR*, 2023. 2

[5] 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

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

[7] Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Errui Ding, Yu Wu, and Jingdong Wang. Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17550–17560, 2025. 2, 8

[8] 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. 2

[9] Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025. 2

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv*, 2023. 2

[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. 2

[12] 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

[13] 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. 3

[14] 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. *Advances in neural information processing systems*, 30, 2017. 6

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

[16] 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

[17] Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, et al. Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. *arXiv preprint arXiv:2506.08797*, 2025. 6

[18] 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. 2, 3, 6

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

[20] 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

[21] 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. 2

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

[23] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *CVPR*, 2024. 2

[24] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *TOG*, 2021. 3

[25] Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, Tianshu Hu, and Yongming Zhu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance, 2025. 2

[26] 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. 3

[27] 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. 2

[28] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025. 6

[29] 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. 6

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

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

[32] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 6

[33] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024. 2

[34] 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. 2

[35] 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. 2

[36] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[37] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025. 4, 8

[38] 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. 3

[39] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

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

[41] Team Wan. Wan: Open and advanced large-scale video generative models, 2025. 6, 8

[42] 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. 3

[43] 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. 3

[44] 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

[45] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *ICCV*, 2021. 3

[46] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025. 2, 6

[47] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv*, 2022. 3

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[49] 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. 2

[50] 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. 3

[51] 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. 6

[52] Bowen Xue, Qixin Yan, Wenjing Wang, Hao Liu, and Chen Li. Stand-in: A lightweight and plug-and-play identity control for video generation. *arXiv preprint arXiv:2508.07901*, 2025. 4, 8

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

[54] 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. 2

[55] 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. 2

[56] 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. 2

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

[58] 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.

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

[60] 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

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[62] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 2, 6

[63] 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. 3