

AnchorCrafter: Animate Cyber-Anchors Selling Your Products via Human-Object Interacting Video Generation

Ziyi Xu, Ziyao Huang¹, Juan Cao¹, Yong Zhang¹, Xiaodong Cun¹, *Member, IEEE*, Qing Shuai¹, Yuchen Wang, Linchao Bao¹, and Fan Tang¹, *Member, IEEE*

Abstract—The generation of anchor-style product promotion videos presents promising opportunities in e-commerce, advertising, and consumer engagement. Despite advancements in pose-guided human video generation, creating product promotion videos remains challenging. In addressing this challenge, we identify the integration of human-object interactions (HOI) into pose-guided human video generation as a core issue. To this end, we introduce AnchorCrafter, a novel diffusion-based system designed to generate 2D videos featuring a target human and a customized object, achieving high visual fidelity and controllable interactions. Specifically, we propose two key innovations: the HOI-appearance perception, which enhances object appearance recognition from arbitrary multi-view perspectives and disentangles object and human appearance, and the HOI-motion injection, which enables complex human-object interactions by overcoming challenges in object trajectory conditioning and inter-occlusion management. Extensive experiments show that our system improves object appearance preservation by 7.5%, and achieves the best video quality compared to existing state-of-the-art approaches. It also outperforms existing approaches in maintaining human motion consistency and high-quality video generation.

Index Terms—Diffusion models, human video generation, digital human synthesis, interaction synthesis.

I. INTRODUCTION

VIEWING unboxing videos and live-streamed product promotions by content creators and live streamers—collectively referred to as anchors—on platforms such as YouTube, TikTok, Douyin, etc, has become an integral part of the online shopping experience. Recent advancements in computer graphics, including video generation [1], [2] and human animation [3], [4], have made it possible to automate the

Received 13 June 2025; revised 2 February 2026; accepted 5 February 2026. Date of publication 9 February 2026; date of current version 17 June 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62572458, in part by Beijing Major Science and Technology Project under Contract Z251100008125009, and in part by CCF-Tencent Rhino-Bird Open Research Fund Recommended for acceptance by S. Zhu. (*Corresponding author: Fan Tang.*)

Ziyi Xu, Ziyao Huang, Juan Cao, Yuchen Wang, and Fan Tang are with the Institute of Computing Technique, Chinese Academy of Sciences, Beijing 100045, China (e-mail: tfan.108@gmail.com).

Yong Zhang is with Meituan, Beijing 100102, China.

Xiaodong Cun is with Great Bay University, Guangdong 523000, China.

Qing Shuai and Linchao Bao are with Tencent AI Lab, Shenzhen 518054, China.

Project page including data, code, and Huggingface demo: <https://github.com/cangcz/AnchorCrafter>.

Digital Object Identifier 10.1109/TVCG.2026.3662720



Reference Input MimicMotion StableAnimator Ours

Fig. 1. Existing methods accurately follow human poses but struggle with realistic hand-object interactions, often misinterpreting the object as part of the human, leading to static animations. In contrast, our approach ensures natural and dynamic movement by precisely synthesizing human-object interactions while preserving object appearance.

creation of such unboxing and promotional content. However, achieving high-fidelity temporal consistency, object realism, and controllable motion generation remains a formidable challenge.

Pose-guided human video generation [5], [6], [7], [8], [9], [10] aligns closely with anchor-style product promotion videos. Existing diffusion-based methods generate temporally consistent, high-fidelity human videos using pose and appearance references. Make-Your-Anchor [8] pioneered personalized anchor-style video generation, while methods like MimicMotion [7] and StableAnimator [9] animate static human images with precise poses, adaptable for anchor-style video generation.

However, due to the lack of effective human-object interaction (HOI) capabilities [11], [12], [13], these methods are unable to generate human-driven product demonstrations. As illustrated in Fig. 1, objects are often treated as static textures, which restricts interactive realism and hinders the modeling of complex hand-object interactions.

Recent studies on HOI generation [11], [12], [13] primarily focus on hand-centric videos or image domains, which do not offer the degrees of freedom necessary for anchor-style product promotion videos. Therefore, seamlessly integrating HOI into pose-guided human video generation is essential for achieving natural and dynamic human-object interactions in synthesized videos.

To generate anchor-style product promotion videos, we introduce AnchorCrafter, a novel object-centric customization framework. Unlike person-centric approaches such as Make-Your-Anchor [8], our method tailors a video diffusion model

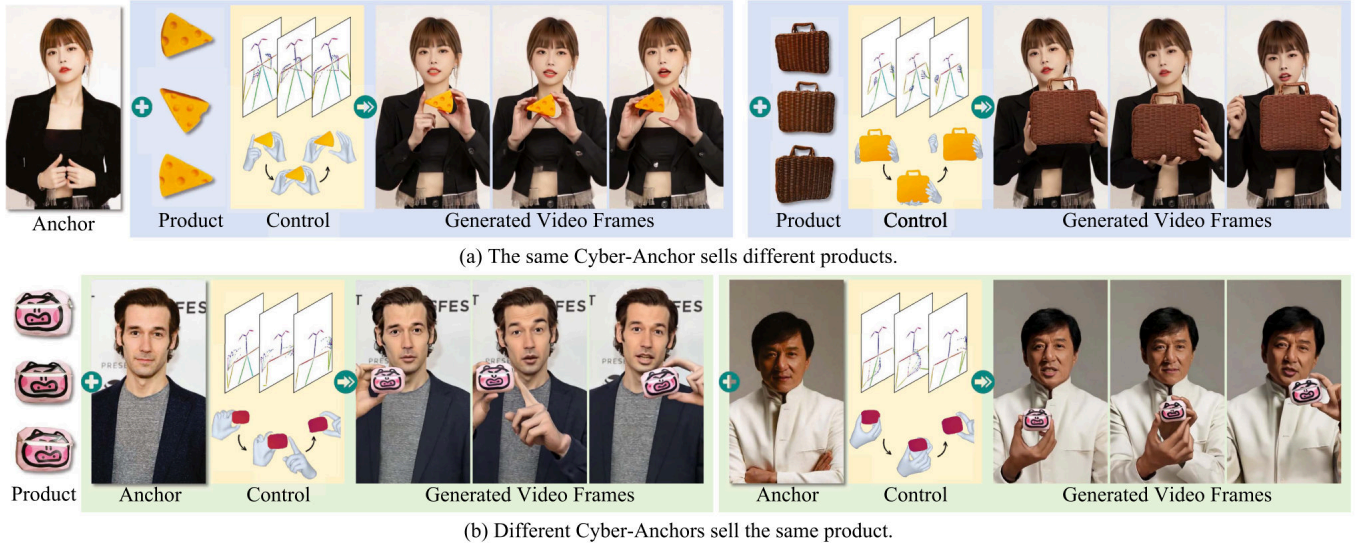


Fig. 2. We propose AnchorCrafter, a diffusion-based human video generation framework for creating high-fidelity anchor-style product promotion videos by animating reference human images with specific products and motion controls. By incorporating human-object interaction into the generation process, AnchorCrafter achieves high preservation of object appearance and enhanced interaction awareness.

by collecting a one-minute interaction video of a specific object. As shown in Fig. 2, the customized model enables arbitrary anchors to naturally interact with the target object in diverse poses. Specifically, we refine human and object representations via HOI-appearance perception, which integrates multi-view features and adopts a decoupled architecture to effectively disentangle human and object appearances. However, without precise motion control, objects remain disconnected from human interactions. To achieve natural and coordinated object behavior, we propose the HOI-motion injection module, which leverages depth maps and 3D hand meshes to provide fine-grained motion guidance to the model. In addition, interaction artifacts are alleviated through the adoption of occlusion-handling strategies. Furthermore, conventional training objectives often underrepresent the subtle dynamics of hand-object interactions, which undermines the realism of generated content. To enhance interaction fidelity, we introduce an HOI-region reweighting loss that strategically emphasizes interaction-critical regions during training, leading to reinforced fine-grained object representations and improved realism in modeled interactions.

We conduct extensive experiments on real-world objects and demonstrate that AnchorCrafter achieves object appearance preservation and interaction awareness results, which current approaches do not accomplish. Moreover, we collected and open-sourced a digital human interaction dataset to fill the gap in the field of interactive digital human generation.

In summary, our contributions are as follows:

- We propose AnchorCrafter, a novel system incorporating human-object interaction into pose-guided human video models and generating realistic anchor-style product promotion videos by animating reference anchors.
- We propose an HOI-appearance perception that utilizes a novel multi-view feature fusion and decoupled injection structure to achieve object appearance preservation, an HOI-motion injection to condition object motion and

handle inter-occlusion, and an HOI-region reweighting loss to enhance object details.

- We conducted quantitative and qualitative evaluations to demonstrate the effectiveness of AnchorCrafter, comparing it with state-of-the-art diffusion-based human video generation and editing methods.

II. RELATED WORK

A. Pose-Guided Human Video Generation

Pose-guided human video generation utilizes sequential poses to control human motion in generated videos, commonly applied in scenarios such as dance and speech video generation. With rapid advancements in diffusion models [14], [15] and pose-controlled human image generation [16], [17], pose-guided human video generation has achieved remarkable progress.

Common practices leverage ControlNet [18], a parallel control branch attached to diffusion UNet, to inject human skeletal motion sequences [3], [6], [8], [19], [20], thereby achieving advanced pose guidance. AnimateAnyone [5] proposed ReferenceNet to control human appearance, which injects information through the self-attention mechanism. Champ [21] enhances motion control by combining four different types of poses. MimicMotion [7] engineered a lightweight pose guidance network to incorporate human pose conditions. To further improve facial detail embedding, StableAnimator [9] introduces a Face Encoder, enhancing precision in facial synthesis.

Recent studies have explored replacing UNet with diffusion transformer (DiT) to achieve better temporal consistency in video generation. Based on the DiT framework, HumanDiT [22] employs prefix latent reference strategies, ensuring visual consistency across extended sequences. UniAnimate-DiT [23], built upon Wan2.1 [24], incorporates Low-Rank Adaptation [25] to facilitate efficient digital human synthesis while reducing computational overhead. Similarly to UNet-based approaches, both models leverage pose sequences to control human motion

within generated images. However, despite these advancements, DiT-based frameworks still require large-scale datasets for training to uphold video quality and coherence, posing challenges in practical deployment.

These studies demonstrate that by incorporating human poses such as OpenPose [26], DensePose [27], SMPL-X [28], control over human movement can be achieved. However, the important human-object interaction generation is often overlooked, despite its vital importance for real-world applications such as anchor-style product promotion video generation. In contrast, AnchorCrafter incorporates HOI into human video generation to address this issue.

B. Human-Object Interaction Generation

The goal of HOI is to generate visual content that accurately represents the relationships between humans and objects. In the domain of 3D reconstruction, EasyHOI [29] enables hand-object interaction reconstruction from a single image, whereas DICE [30] facilitates single-image reconstruction of hand-face interactions. In the field of motion generation [31], [32], some studies focus on interactive motion synthesis, aiming to generate human-object interactions with enhanced realism. InterDiff [33], TeSMo [34] and HOIAnimator [35] predict human motions during interactions with objects based on 3D object representations. PMP [36] learns to use part-wise motion priors to physically interact with environments. IMoS [37] generates full-body human and 3D object motions from textual input, but its focus is primarily on hand-grasping small objects. TextIM [38] synthesizes human interactive motions with part-level semantic accuracy by aligning textual descriptions to interactive body movements.

However, human-object interaction remains relatively underexplored in the field of content generation. Hand-object interaction image generation [39] focuses on static synthesis tasks. HOI-Swap [11] instead targets hand-centric videos, employing a two-stage editing framework to swap the object in hand while preserving hand-object interactions. ReHold [13] proposes the HOI Restoration Module, which injects hand and object details into the model to enhance hand-object interaction. MIMO [40] investigates controllable character video synthesis with spatially decomposed modeling, but it focuses on replacing human characters in interactive scenes and requires per-frame object and background inputs. Meanwhile, HOIGen-1M [41] introduces a large-scale text-to-video dataset for human-object interaction, providing a valuable resource for general HOI generation. However, its scope does not fully correspond to application scenarios such as e-commerce advertising and virtual anchors that demand high-fidelity object preservation and controllable motion. VirtualModel [12] stands as the most relevant work in literature, albeit within the domain of image generation. It excels in synthesizing interactive images of characters by leveraging input objects and human poses. Existing methods primarily focus on HOI image generation or object swapping in hand-centric videos, leading to the absence of the ability to generate videos with sufficient degrees of freedom, particularly videos that include the entire human body. In contrast, our approach integrates HOI capabilities into a human video generation model, enabling the generation of high-quality, controllable anchor-style product promotion videos.

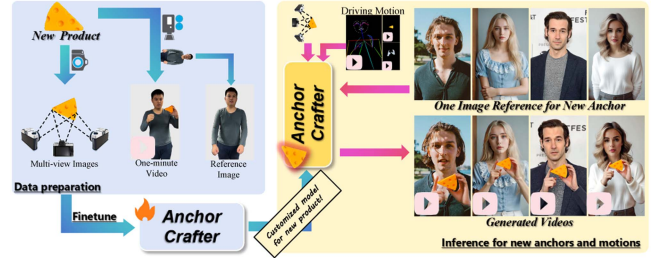


Fig. 3. Fine-tuning with a oneminute video enables customization for new products. After fine-tuning, our method could generate arbitrary anchor videos selling the product with various unseen motions.

III. SYSTEM SETTING

Overview: Driven by cutting-edge progress in pose estimation [42], [43] and enhanced by access to large-scale human datasets [44], [45], [46], [47], techniques [6], [7], [19], [45] for producing high-fidelity, controllable human videos have become increasingly mature. Current open-source human-object interaction datasets remain limited in both coverage and fidelity, posing challenges for anchor-style product video generation. To better address customized product interactions, we frame the task as a learning paradigm for customized products: learns fine-grained texture and appearance from a one-minute interaction video with the product. By leveraging this learned representation, the model generalizes effectively, producing interaction videos with diverse human appearances and precisely controlled motions. The learning framework of our system consists of two stages: training on diverse datasets to capture the distribution of HOI videos, and fine-tuning on specific objects, enabling more precise and adaptable interaction modeling.

Training: During training, each source video $X = \{x^{1:f}\}$ with f frames enables the model to learn the target object O and its interaction patterns with humans. To ensure accurate appearance modeling, we use a reference image I_H for the anchor and multi-view images $I_O = \{i_O^1, i_O^2, i_O^3\}$ for the object. We employ multiply comprehensive conditions to control the HOI motions, including the sequences of human skeletons $P = \{p^{1:f}\}$, 3D hand meshes $H = \{h^{1:f}\}$ and depth maps of the object $D = \{d^{1:f}\}$. The system then generates a sequence of frames $Y = \{y^{1:f}\}$ featuring a target human and customized object with controlled interactions. To improve interaction modeling, we collect a foundational dataset of human-object interactions. Given the diversity of objects and dataset constraints, our primary focus is refining interaction control rather than capturing intricate texture details.

Fine-tuning: During fine-tuning, we collect videos exceeding one minute in length for each object requiring demonstration, as illustrated in Fig. 3. Benefiting from the training phase, where the model learns human-object interaction concepts, the fine-tuning stage requires only a short video to grasp object textures. This enables the model to generate previously unseen poses and novel anchor interactions. We maintain that collecting one-minute videos is both practical and essential for achieving precise and realistic object rendering, while still allowing flexible replacement of presenters and actions in commercial scenarios.

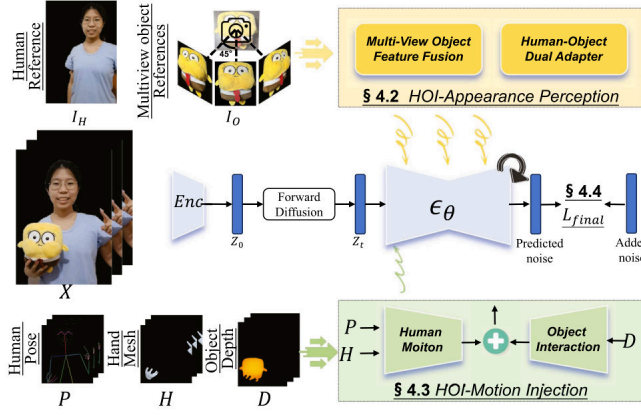


Fig. 4. Training pipeline for AnchorCrafter: Based on a video diffusion model, AnchorCrafter injects human and multi-view object references into the video via HOI-appearance perception. The motion is controlled through HOI-motion injection, with the training objective reweighted in the HOI region.

Inference: In the inference stage, we use HOI motions performed by actors to drive the generation of anchor videos. Actors demonstrate the required interaction actions by holding the products O learned during fine-tuning, and the extracted control conditions P , H , and D will serve as inputs to the system. Subsequently, our model can generate product-promotion videos for any unseen anchor image I_H .

IV. METHODOLOGY

As illustrated in Fig. 4, our system's pipeline consists of two core components: HOI-appearance perception and HOI-motion injection. We first introduce the video diffusion model architecture utilized in AnchorCrafter in Section IV-A. To enhance the reconstruction of both human and object appearances, we design HOI-appearance perception (Section IV-B), which employs an appearance-decoupling strategy to ensure accurate representation. To regulate the movement of humans and objects, we integrate HOI-motion injection (Section IV-C), which introduces precise control signals into the system. Additionally, to refine interaction fidelity, we introduce HOI-region reweighting loss (Section IV-D), which strategically enhances hand-object interaction details during inference.

A. Video Diffusion Model

AnchorCrafter is based on a video diffusion model [48] architecture, comprising a diffusion UNet [15], [49] with temporal layers, denoted as ϵ_θ , and a variational autoencoder (VAE) [50] composed of an encoder Enc and a decoder Dec for compressing and uncompressing the video frames. During training, the video sequence X is encoded into the latent space as $Z_0 = Enc(X) = \{z_0^{1:f}\}$. The core training objective for the video diffusion model is:

$$L_{diff} = \mathbb{E}_{\epsilon \sim N(0, I), Z, c, t} [\|\epsilon - \epsilon_\theta(Z_t, c, t)\|_2^2], \quad (1)$$

where c is conditional signal, $Z_t = \{z_t^{1:f}\}$ is the latent feature diffused from Z_0 through a deterministic Gaussian process over t timesteps by adding noise ϵ . During inference, the initial noises

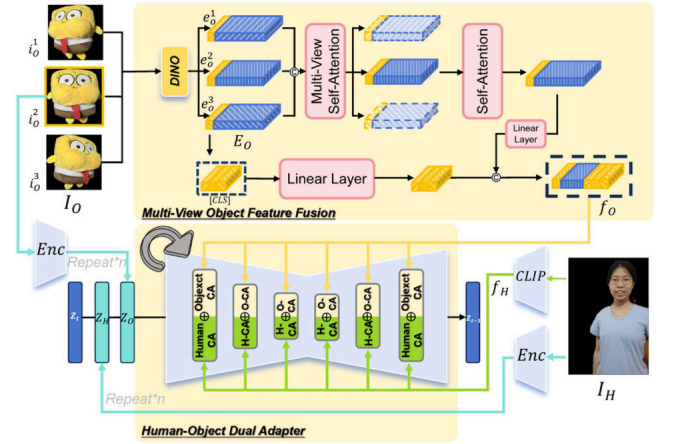


Fig. 5. HOI-appearance perception: The feature of the target object f_O is extracted through multi-view object feature fusion and combined with the human reference feature f_H within a human-object dual adapter to achieve improved disentanglement results.

$U = \{u^{1:f}\}$ sampled from Gaussian noise are denoised with T timesteps to get the estimated $\hat{Z}_0$, and the output video with $Y = Dec(\hat{Z}_0) = \{y^{1:f}\}$.

B. HOI-Appearance Perception

As illustrated in Fig. 5, the HOI-appearance perception design focuses on extracting appearance features for humans and objects, which are then integrated into the backbone network. Referring to previous approaches [5], [7], [19], [51], we integrate global context and local details from input images for human appearance.

Specifically, the VAE encoder Enc maps human images to a latent space compatible with the diffusion model $Z_H = Enc(I_H)$, which preserves the local detail. Subsequently, Z_H is repeated f times and concatenated with the UNet input Z_t . Meanwhile, the CLIP [52] features of the human image are extracted as $f_H = CLIP(I_H)$ for global representation.

To accurately depict objects, especially the promoting product, the model must perceive their shape and texture from multiple perspectives. Instead of a single reference image, we utilize multi-view object reference images $I_O = \{i_O^1, i_O^2, i_O^3\}$ captured from 45° left, front, and 45° right angles. Similar to human reference, the VAE feature of the front reference i_O^2 is extracted as $Z_O = Enc(i_O^2)$ and then repeated f times and concatenated with Z_t . Different from the process for human reference, we propose multi-view object feature fusion to extract richer information from multi-view images of the input object images.

1) *Multi-View Object Feature Fusion:* We introduce the multi-view object feature fusion to understand object appearance inherently. The process starts by feeding the multi-view object reference I_O into the pre-trained DINOv2-large model [53], resulting in the extraction of embedding $E_O = \{e_O^1, e_O^2, e_O^3\} \in R^{n \times m}$, $n = 1370$, $m = 1024$. Then, E_O is processed through two distinct branches, as illustrated in Fig. 5. In the first branch, we designed a multi-view self-attention layer to merge object features from three perspectives, where three object embeddings of shape $n \times m$ from DINO are concatenated along the n

channel and then a self-attention is performed. Subsequently, we extract the e_O^2 of the embedding, perform self-attention, and employ a linear network to project the features into a reduced dimensional space of $16 \times m$. In the other branch, we concatenate the first [CLS] feature from E_O as a vector of size $R^{3 \times m}$ for global representation, then pass through a linear layer. Finally, concatenate two branches' outputs as the final multi-view object representation, $f_O \in R^{19 \times m}$. This module extracts object features precisely from images captured from three perspectives, extracts 3D consistency features from multi-view images [54], and facilitates the final reproduction of objects.

Unlike earlier methods, HOI-Swap [111] extracts DINO features from a single image without multi-view fusion, while [55] leverages a 3D representation model to fine-tune 2D extractors. In contrast, we directly utilize DINO features from multiple views and introduce an feature fusion module, which enables faithful preservation of object appearance in HOI video generation.

2) *Human-Object Dual Adapter*: We observed that directly integrating object features f_O into UNet along with human features f_H leads to the appearance entanglement of the objects and humans. We propose a human-object dual adapter by replacing the cross-attention layers in each block of the diffusion UNet to achieve better disentanglement between humans and objects. The human feature f_H extracted from CLIP [52] and the object feature f_O obtained by multi-view object feature fusion are each fed into a cross-attention layer, and the formula can be written as:

$$HumanCA := Softmax\left(\frac{QK_H^T}{\sqrt{d}}\right) \cdot V_H, \quad (2)$$

and

$$ObjectCA := Softmax\left(\frac{QK_O^T}{\sqrt{d}}\right) \cdot V_O, \quad (3)$$

where $Q = W_Q \cdot Z_t$, $K_H = W_{KH} \cdot f_H$, $K_O = W_{KO} \cdot f_O$, $V_H = W_{VH} \cdot f_H$, $V_O = W_{VO} \cdot f_O$, $Z_t = \{z_t^{1:f}\}$ is the input latent frames on diffusion timestep t and W_Q, W_{KH}, W_{KO}, W_V are learnable weights for attention modules. Obtain the final feature set by summing the output features from the *HumanCA* and *ObjectCA* layers. Dual adapter effectively extracts features of objects and humans while achieving the disentanglement of the two entities.

C. HOI-Motion Injection

We propose an HOI-motion injection module that explicitly models hands and occlusion to address the challenge of generating specified interactive motions.

A critical challenge in conditioning the movement of objects is providing an unambiguous representation of their orientation and positional trajectory in 3D space. To address this, the depth maps P serve as the primary input for trajectory estimation. The depth maps of the target object are first estimated and cropped, and then processed through a lightweight ninelayer convolutional network. The resulting features are added elementwise to the output of the first convolutional layer of the UNet. Due to the suboptimal performance of existing pose extractors in handling

complex hand movements, we complement the commonly used human skeleton P , which controls the overall pose, with the extraction of 3D mesh sequence H of the hands. Since hand poses are prone to errors under occlusion, we explicitly handle occlusions between hands and objects during interactions by masking the regions of the 3D mesh where the object occludes the hands. The skeleton P and the masked hand meshes H are concatenated and processed through a convolutional network with the same architecture as the objectinteraction branch but with unshared parameters. The features extracted by this network are subsequently integrated into the UNet.

We observe that spatial differences between the input pose sequences P , H , and D and the reference human image I_H affect the generated results. For instance, when reference I_H depicts a person closer to the camera, while the input pose P is farther from the camera and includes more body parts, the input pose does not correspond to the reference image, leading to distortion in the consistency and accuracy of the generated video. To mitigate this problem, we explicitly handle spatial mismatches between condition sequences, thereby reducing scale and viewpoint artifacts. Let the 2D body keypoints in P be

$$P_{\text{coord}} = \{(x_k^j, y_k^j)\}_{k=1}^K, \quad j = 1, \dots, f, \quad (4)$$

and the reference keypoints in I_H be

$$I_{\text{coord}} = \{(x_k, y_k)\}_{k=1}^K. \quad (5)$$

The symbol f denotes the total number of frames, while k represents the index of a point within frame j . For each frame j , the vertical scale and translation are estimated via least squares:

$$(\alpha_y^j, \beta_y^j) = \arg \min_{a,b} \sum_{k=1}^K (y_k - (a y_k^j + b))^2. \quad (6)$$

The horizontal scale is derived as $\alpha_x^j = \alpha_y^j \cdot \frac{\text{width}(I_H)}{\text{height}(I_H)}$, and β_x^j is obtained by aligning $\alpha_x^j x_k^j$ with x_k . Finally, the similarity transform

$$\begin{aligned} x_k^{j,\text{aligned}} &= \alpha_x^j x_k^j + \beta_x^j, \\ y_k^{j,\text{aligned}} &= \alpha_y^j y_k^j + \beta_y^j \end{aligned} \quad (7)$$

maps P_{coord} into the spatial frame of I_H . We further apply this transform to all motion conditions H and D , uniformly rescaling and shifting their coordinates into the same reference frame as I_H . This ensures consistent spatial alignment and improves the overall quality of the generated videos. Compared with conventional skeleton- or mesh-based guidance [7], [21], [39], [56], our design incorporates 3D hand meshes, depth maps, and occlusion-masking strategies, substantially enhancing the precision of human-object interaction modeling.

D. HOI-Region Reweighting Loss

By constructing modules for HOI, our model enables control over the generated humans, objects, and their interactions. Accordingly, based on the diffusion model's loss function described in (1), our training objective can be formulated with $c = \{P, H, D, I_H, I_O\}$. However, we observed that standard

training loss caused the model to fail to learn object appearances adequately.

To mitigate this, we propose a HOI-region reweighting loss, which enhances the model's attention to interaction regions throughout the training process:

$$L_{object} = \eta \frac{S}{S_{obj} + S_{hand}} M_{inter} \odot L_{diff}, \quad (8)$$

where S , S_{obj} , S_{hand} is the area of the whole image, object, and hands, M_{inter} indicates the mask of the interaction region consisting of object and hands, and η is a hyperparameter. Using the inverse of the image area occupied by the interaction region as a weight, the model can assign higher training importance to the interaction area. The final loss is:

$$L_{final} = (1 - M_{inter}) \odot L_{diff} + L_{object}. \quad (9)$$

After applying the final loss, pose-driven capability across diverse human appearances is preserved, and objects are learned precisely. It is worth noting that our HOI-region reweighting loss differs from previous HOI loss designs. For example, [57] addresses HOI detection in static images and employs a multi-task detection loss combined with relation-embedding alignment. In contrast, our loss is tailored for HOI video generation, spatially reweighting small but critical interaction regions to enhance generative fidelity.

V. EXPERIMENTS

This chapter provides a comprehensive evaluation of our system. We first introduce the dataset and experimental setup (Section V-A), followed by benchmarking against SOTAs to evaluate appearance preservation, motion control, and interaction realism (Section V-B). To further validate our design choices, we conduct ablation studies on HOI appearance perception, HOI motion injection, loss reweighting, and fine-tuning strategies (Section V-C), complemented by a user study (Section V-D). Finally, we discuss the scalability of our method (Section V-E).

A. Experimental Settings

1) *Dataset*: For training, due to the lack of an open-source dataset specifically suited to our task, we constructed a custom dataset comprising two types of videos: human-only videos and human-object interaction videos.

- *Human-only dataset*: The human-only subsets comprising a total of 9,906 videos originate from two open-source datasets: HumanVid [45] and DreamDance [47]. HumanVid includes a variety of resolutions and encompasses both internet-sourced and generated data. We filtered the HumanVid to retain only vertical videos from the internet-sourced subset featuring a single person, resulting in a collection of 5,848 videos. Meanwhile, DreamDance is a human dance dataset providing URLs for each video. Due to access restrictions, we successfully downloaded 4,058 videos.
- *Human-object interaction dataset*: The human-object interaction subset includes 356 real-world videos recorded using mobile devices, as well as 44 livestream recordings

collected from online platforms. These videos feature 15 participants interacting with 286 distinct objects, primarily comprising common household items such as toys, cups, and boxes. Each video has a duration of approximately 20 seconds at 30 FPS, capturing natural and spontaneous human-object interactions. To ensure diverse and complete object representation, we photograph objects from three perspectives: frontal, 45° left rotation, and 45° right rotation. For a small number of objects lacking three-view coverage, duplicated images were used to approximate multi-view input.

For testing, we compiled a dataset of 80 videos, featuring interactions between eight individuals sourced from internet and five distinct objects. For each object, we provide approximately one minute of interaction footage for fine-tuning, together with two evaluation clips of about seven seconds each. The individuals engage with the objects in diverse poses, ensuring a broad range of interaction scenarios. The training and testing datasets have been made public.

Pre-processing: For each video, we use DWPose [42] to extract human skeletal pose sequences, following [7]. For the videos we recorded, we additionally employ HaMeR [58] to extract hand 3D mesh sequences, SAM2 [59] to extract object motion masks, and ViTA [60] to produce depth maps. For the human-only videos, since they do not involve objects and rapid hand movements result in blurring, we use black pixels to represent objects, hand sequences, and object depth maps.

2) *Implementation Details*: The pre-trained weight of MimicMotion [7] is used for basic pose-driven human animation. We freeze VAE, CLIP, and DINO, training all other parameters on seven gpus with 40 GB. In the first stage, we train for 12,000 iterations at a resolution of 512×768 with ten frames per sample. In the second stage, we train for 4,000 iterations at 576×1024 with six frames per sample. During these two processes, 65% of the training data is sampled from Human-object interaction dataset, while 35% is sampled from Human-only dataset.

During the fine-tuning phase, we train five objects simultaneously. To enhance the generalization capability for face generation and mitigate the impact of limited human diversity, 80% of the training data is sampled from the fine-tuning dataset, while the remaining 20% comes from DreamDance. The model undergoes 3,500 iterations at a resolution of 576×1024 , requiring approximately three to four hours. The learning rate is set to 10^{-5} with a 300-iteration warm-up. The HOI-region reweighting loss is applied with $\eta = 1$ to enhance object detail learning.

During inference, CFG is set to 4.0. And we apply the time-aware position shift fusion from Sonic [61] to generate long videos, processing ten frames per sample. Inference for a 100-frame video takes five minutes and requires 22.84 GB of VRAM.

3) *Comparison Methods*: We compare AnchorCrafter with seven state-of-the-art methods: MimicMotion [7], StableAnimator [62], UniAnimate-DiT [23], VACE [24], FlexiAct [63], MIMO [40], and Make-Your-Anchor [8]. MimicMotion and StableAnimator are pose-guided human video generation methods based on UNet. UniAnimate-DiT is a digital human synthesis model leveraging the DiT architecture, while VACE is an all-in-one DiT-based framework designed for both video creation

TABLE I

THE QUANTITATIVE RESULTS OF OUR METHOD WERE COMPARED WITH THOSE OF SOTAs AND ABLATION STUDIES. OUR METHOD SIGNIFICANTLY OUTPERFORMS EXISTING APPROACHES REGARDING NUMERICAL PERFORMANCE FOR APPEARANCE PRESERVATION OF OBJECTS WHILE ALSO MATCHING OR EXCEEDING CURRENT METHODS IN IMAGE AND VIDEO QUALITY AND HUMAN POSE CONTROL CAPABILITY. SUBJECT CONSISTENCY (SUBJ-CONS), BACKGROUND CONSISTENCY (BACK-CONS) AND MOTION SMOOTHNESS (MOT-SMTH) ARE PERCENTAGES.

Method	FID↓	FVD↓	Obj-CLIP↑	Face-Cos↑	LMD (Hand)↓	VLM Score↑	Subj-Cons↑	Back-Cons↑	Mot-Smth↑
MimicMotion	163.77	2470.75	0.843	0.60	20.43	0.56	93.24	93.23	98.62
StableAnimator	171.11	2596.31	0.846	0.55	19.14	0.66	92.95	91.96	98.38
UniAnimate-DiT	212.83	2964.71	0.772	0.73	-	0.23	93.73	91.73	98.88
VACE	160.84	2543.83	0.857	0.40	22.78	0.78	91.50	91.44	98.55
Ours	158.03	2385.39	0.921	0.70	12.19	0.96	95.43	94.96	98.91
w/o Human-Object Dual Adapter	156.56	2404.90	0.892	0.62	12.33	0.88	95.30	94.75	98.91
w/o Multi-View Self-Attention	159.84	2450.67	0.904	0.69	12.41	0.90	95.37	94.74	98.87
w/o Hand Mesh Sequence	162.00	2427.71	0.918	0.68	13.19	0.91	95.88	95.13	98.92
w/o Depth Map Sequence	162.31	2484.67	0.901	0.65	12.75	0.90	94.97	94.66	98.82
w/o HOI-Region Reweighting Loss	159.21	2396.31	0.904	0.65	12.46	0.93	95.40	94.91	98.94
w/o Fine-Tuning	178.99	2404.77	0.847	0.71	12.83	0.94	94.75	94.32	98.85

and editing. Since these models do not directly accept object inputs, we prepared human images with objects composited onto them for comparison. FlexiAct, MIMO, and Make-Your-Anchor target different problem settings and have additional usage constraints. FlexiAct is a motion transfer framework that typically requires fine-tuning for specific actions; MIMO focuses on replacing human characters in interactive scenes but is restricted to full-body scenarios; and Make-Your-Anchor performs subject-specific fine-tuning from a short target video and leverages explicit motion guidance; Consequently, we conduct a qualitative analysis to compare the performance of FlexiAct, MIMO and Make-Your-Anchor.

4) *Metrics*: We evaluate our system across multiple perspectives, including video quality, object appearance understanding, ensuring a comprehensive assessment of interaction realism and detail preservation. Specifically, for video quality, we report FID [64] and FVD [65], and adopt VBench [66] to measure subject consistency (Subj-Cons), background consistency (Back-Cons), and motion smoothness (Mot-Smth). For object generation, we introduce Object-CLIP, which evaluates appearance consistency by computing the CLIP cosine similarity between generated objects and reference objects. For human generation, we employ AdaFace [67] to extract facial features and calculate the cosine similarity between generated and reference faces (denoted as face-cos). Hand consistency is measured using Landmark Mean Distances (LMD) [8], leveraging OpenPose [26]. In addition, we incorporate the Qwen2.5-VL [68] to provide strict evaluation of video quality as VLM score. The evaluation prompt enforces rigorous criteria: (i) the human must interact continuously and naturally with the object (e.g., holding, displaying, operating); (ii) the object must remain consistently visible without artifacts such as penetration, misalignment, or flickering; and (iii) human actions must strictly follow logical and physical rules of object usage. The scoring is binary: 1 indicates perfect quality with flawless interaction and motion logic, while 0 indicates failure if any artifact or unnatural interaction is observed.

B. Main Results

1) *Quantitative Results*: Table I presents the quantitative evaluation of our approach. Due to the complexity of

UniAnimate-DiT’s alignment strategy, ground truth for pose is unavailable; consequently, we did not assess the accuracy of hand poses.

The results highlight the superior performance of our method in object appearance reproduction and human generation. Regarding object appearance preservation, our approach achieves the highest Object-CLIP score among all evaluated methods. Notably, some baselines also attain relatively high scores; however, this is largely because they treat objects as extensions of human clothing or background, causing them to remain static rather than distinct, independent entities.

For human body generation, our method excels in both facial and hand synthesis. The Face-Cos metric significantly surpasses that of the baseline MimicMotion and is comparable to UniAnimate-DiT, indicating that introducing object concepts does not compromise human generation quality. It is worth noting that UniAnimate-DiT is a 14B-parameter model, whereas our framework is substantially smaller (only 1.5B), further highlighting the efficiency of our design.

Furthermore, VBench evaluations affirm that our generated videos exhibit superior subject consistency, background stability, and motion smoothness, while FID and FVD further indicate improved overall visual quality and spatiotemporal fidelity. In addition, the VLM Score provides a strict assessment of interaction realism, confirming that our method produces natural and coherent human–object interactions. Collectively, these findings validate the high visual quality and fidelity of our approach.

2) *Qualitative Results*: Our qualitative results, illustrated in Fig. 6, demonstrate the efficacy of our approach. In these figures, red boxes denote insufficient object appearance fidelity, orange boxes highlight inaccuracies in hand synthesis, and green boxes indicate unrealistic anchor placements. The examples on the left pertain to stationary object interactions, whereas those on the right involve object displacement. Our method achieves superior preservation of both human and object appearances while faithfully capturing their dynamic behaviors.

MimicMotion and StableAnimator generate high-quality, pose-driven human images; however, they exhibit coupling artifacts, treating objects as static extensions of clothing rather than independent entities. When displacement occurs, objects either remain fixed in place or exhibit visual artifacts. UniAnimate-DiT maintains human appearance relatively well but suffers from

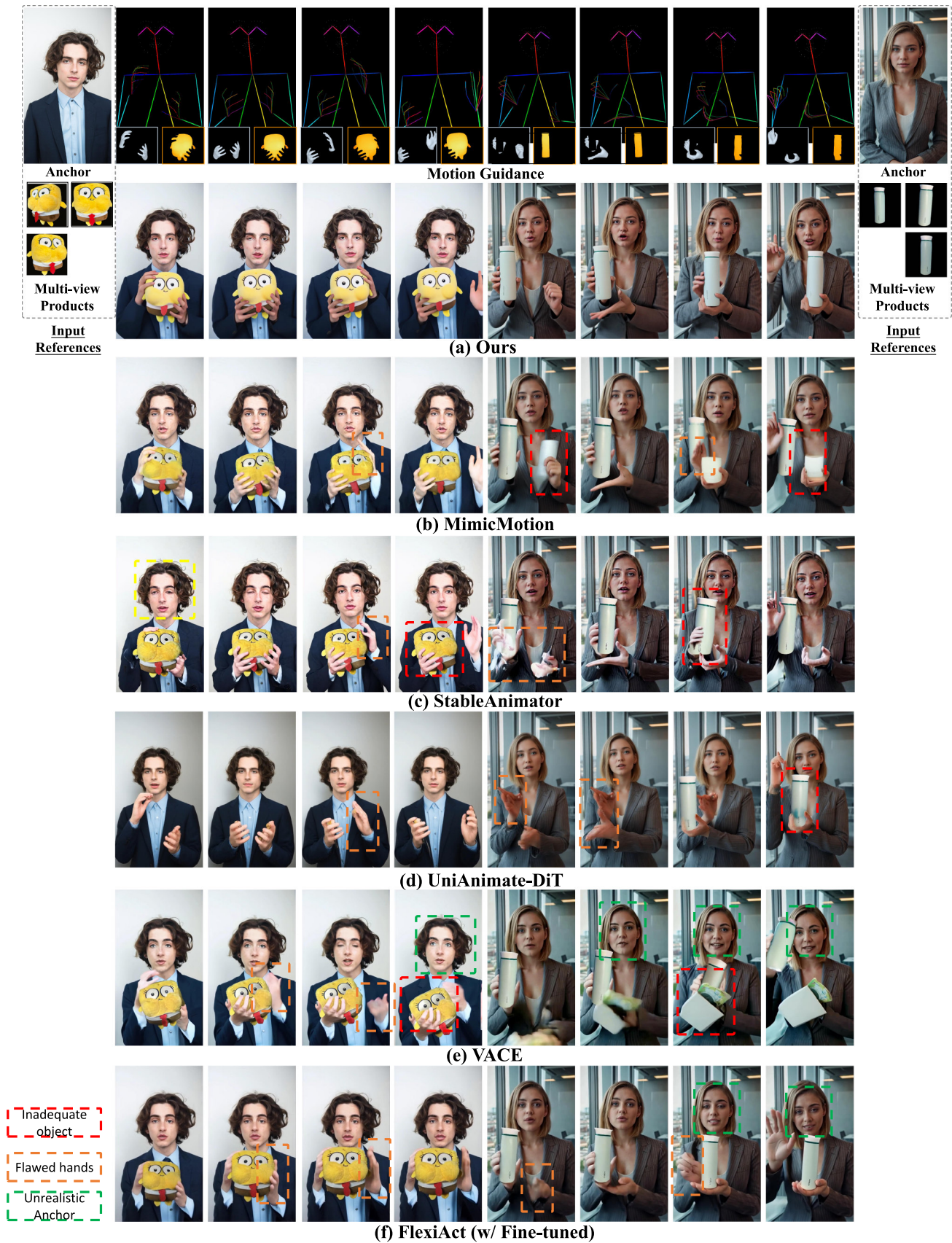


Fig. 6. Qualitative comparisons with other methods. Other approaches fail to preserve the visual integrity of individuals and objects or maintain reasonable interactive motion, making them incapable of completing character interaction tasks.

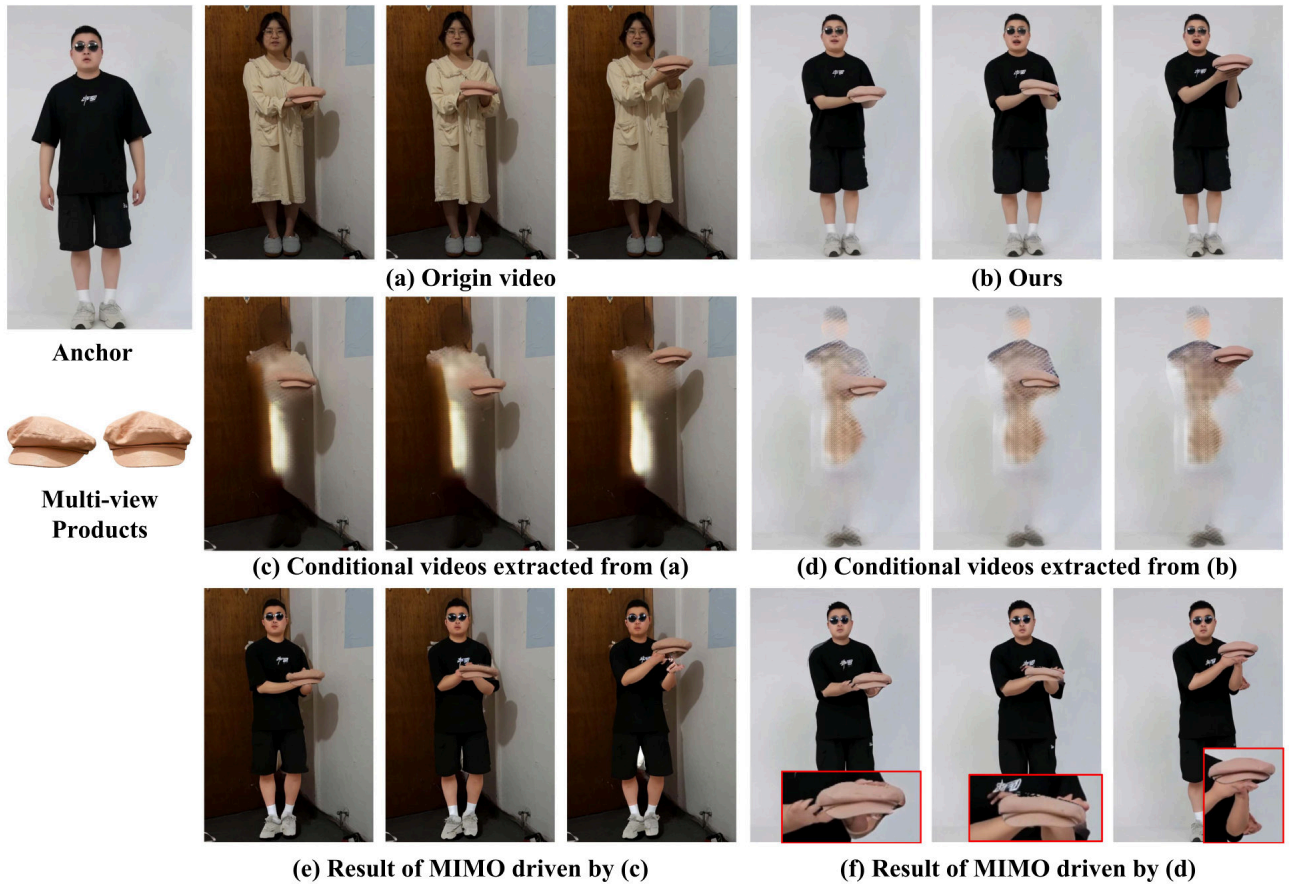


Fig. 7. Compared qualitatively with MIMO, the generated results of MIMO are more like pasted images and do not naturally blend with the background and objects.

object disappearance in certain frames and produces the least accurate hand synthesis. VACE struggles to preserve human appearance, leading to severe object deformation. Due to the complexity of hand-object interactions, all models except ours fail in generating realistic hand representations. For FlexiAct, we conducted 4,000 rounds of fine-tuning on the corresponding test videos, enabling it to fully learn object motion and appearance. While it successfully achieves object displacement, the overall visual quality remains suboptimal, and it fails to transfer interactive motions to novel subjects.

Fig. 7 presents results comparison between our method and MIMO. Following MIMO's official preprocessing pipeline, it requires per-frame inputs of human pose, background, and object, functioning primarily as a character replacement model. When driven by the original video, the generated human appears overlaid rather than naturally integrated with the background and objects. We further attempted to drive MIMO using conditions extracted from our results. This ensures perfectly matched human conditions and reduces its generation difficulty, but substantial errors remain in the hand-object interaction regions.

Fig. 8 presents a case comparison between our method and Make-Your-Anchors. We fine-tune Make-Your-Anchors on a short anchor video. Despite this adaptation, it still produces severe distortions in the hands, and once the object starts moving, it often suffers from disappearance and deformation artifacts. In

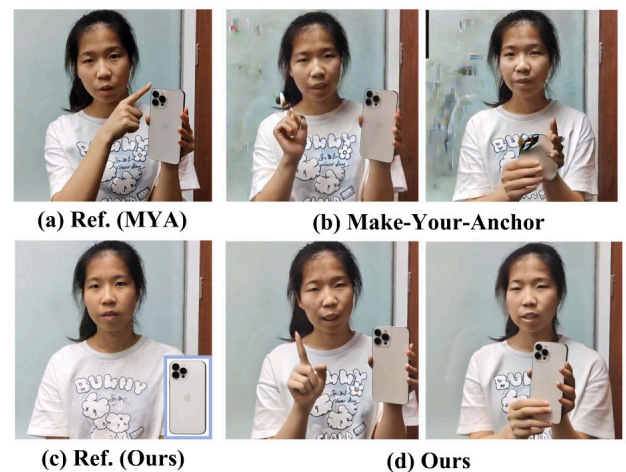


Fig. 8. Compared qualitatively with Make-Your-Anchors.

contrast, our method better preserves the manipulated object appearance and yields more coherent hand-object interaction across frames.

By leveraging HOI-appearance perception and HOI-motion injection, our method effectively preserves both the appearance and motion dynamics of humans and objects, producing high-fidelity human-object interaction videos.



Fig. 9. Ablation studies. Our modules improve the preservation of the object and its interactions with the hands. The error section has been enlarged in the lower right corner.

C. Ablation Study

Validation of human-object dual adapter: We conducted detailed ablation experiments to evaluate the effectiveness of the proposed human-object dual adapter. As shown in Table I, removing the adapter leads to a notable drop in the VLM Score, indicating that the generated interactions become less natural and coherent. Meanwhile, the Object-CLIP score

also decreases, suggesting that simply injecting latent object representations is insufficient for robust object perception, and coupling with human subjects further causes a decline in face-cos. The qualitative results in Fig. 9 further demonstrate that both humans and backgrounds exert significant influence on objects. In contrast, the human-object dual adapter effectively learns object representations while decoupling human and object features.

Validation of multi-view self-attention: As shown in Table I, removing the multi-view self-attention branch leads to a slight decrease in Obj-CLIP and VLM Score, as multi-view self-attention facilitates the extraction of more object information. Since these variations occur at a fine-grained level, they are difficult for CLIP to fully capture, resulting in only a marginal drop in Obj-CLIP. However, the decline in VLM Score indicates that interaction realism is affected when view-point adaptation is absent. As illustrated in Fig. 9, when the object's perspective shifts significantly, its logo undergoes deformation. Multi-view self-attention effectively adapts to different viewpoints and reconstructs object appearance more accurately, thereby improving both quantitative and qualitative performance.

Validation of Hand Mesh Sequence: We removed the hand 3D meshes information injection from the HOI-motion injection. As shown in Table I, the absence of detailed hand guidance resulted in a decrease in LMD (hand). Since errors are typically localized to specific fingers and are difficult for OpenPose to fully detect, the overall decline is not significant. The qualitative results illustrated in Fig. 9 show that the complexity of hand-object interactions, coupled with inaccurate hand pose guidance provided by DWPose, leads to the generation of artifacts. These artifacts become particularly noticeable when the hand is occluded by objects.

Validation of Depth Map Sequence: We further ablated the depth map sequence to evaluate its contribution to HOI-motion injection. As shown in Table I, removing depth information resulted in an increase in LMD (hand) and a decline in both subject and background consistency. This demonstrates that depth maps play a critical role in preserving spatial relationships between humans and objects, particularly under occlusion or complex interaction scenarios. Without explicit depth guidance, the model tends to misinterpret relative positions, leading to artifacts such as misaligned hand-object contact or object deformation. The qualitative results in Fig. 9 further confirm that depth maps provide essential geometric constraints, enabling more accurate reconstruction of object boundaries and enhancing the realism of human-object interactions.

Validation of HOI-region reweighting loss: In conjunction with the HOI-region reweighting loss, we assigned higher weights to the hand-object regions during training to enhance the learning of objects and hands. As shown in Table I, insufficient emphasis on the hand-object regions causes the model to overfit to the human figures in the training set while failing to adequately learn object features, leading to inaccuracies in the appearance of both the object and the human figure. Fig. 9 further demonstrates that, without adequate region enhancement, the model struggles to capture fine object details effectively.

Validation of Fine-Tuning: During the fine-tuning phase, we focus on learning the intricate texture details of objects. Table I demonstrates that fine-tuning significantly enhances the Object-CLIP score. As illustrated in Figs. 9 and 10, when presented with a new object without finetuning, the model successfully generates appropriate interactive actions together with its contour. However, it fails to capture fine-grained texture details. This observation indicates that our foundational model already possesses the capability for human-object interaction. Through



Fig. 10. Ablation study of fine-tuning. After fine-tuning, the model can learn the texture details of objects.

fine-tuning, the model further develops the ability to perceive and represent fine-grained texture details.

D. User Study

We conducted a user preference evaluation comparing AnchorCrafter with four state-of-the-art methods. Each participant reviewed 30 randomly selected videos for each method, scoring them based on five criteria:

- *Appearance Preservation (Human):* Evaluates the consistency between humans and the reference images.
- *Appearance Preservation (Object):* Focuses on maintaining object details such as texture and shape.
- *Motion Accuracy (Human):* Assesses the precision of human movements, particularly poses and gestures.
- *Motion Accuracy (Object):* Ensures that objects' motions align with intended physics and interactions.
- *Overall Quality:* Provides a comprehensive evaluation of visual appeal, coherence, and stability.

Each criterion was rated on a five-point scale, with five being the highest score. Before the evaluation, participants were given detailed instructions, reference videos, and corresponding conditions to ensure fairness and clarity throughout the process. A total of 50 participants were invited, including 25 males and 25 females, achieving balanced gender representation. The participants consisted of 30 computer vision researchers, 10 artists, and 10 individuals from diverse backgrounds. This diverse group provided comprehensive feedback from various perspectives. As shown in Table II, our approach consistently ranked the highest across all five criteria, excelling particularly in motion accuracy and appearance diversity. This demonstrates its strength in generating realistic and stable human-object interaction videos.

E. Discussions

1) *Generalization in Multi-View Tasks:* Human-object interaction tasks frequently require displaying objects from diverse

TABLE II
USER STUDY SCORES. THE RATING SCORE IS ON A SCALE FROM ONE TO FIVE,
WHERE FIVE IS THE HIGHEST SCORE AND ONE IS THE LOWEST

Method	Appearance		Motion		Overall
	Human	Object	Human	Object	
MimicMotion [7]	3.52	1.84	3.68	2.08	3.08
StableAnimator [9]	3.02	2.18	3.64	1.96	2.96
UniAnimate-DiT [23]	4.02	1.28	3.34	1.56	3.14
VACE [24]	2.28	1.92	2.42	2.16	2.62
Ours	4.12	4.68	4.58	4.82	4.64

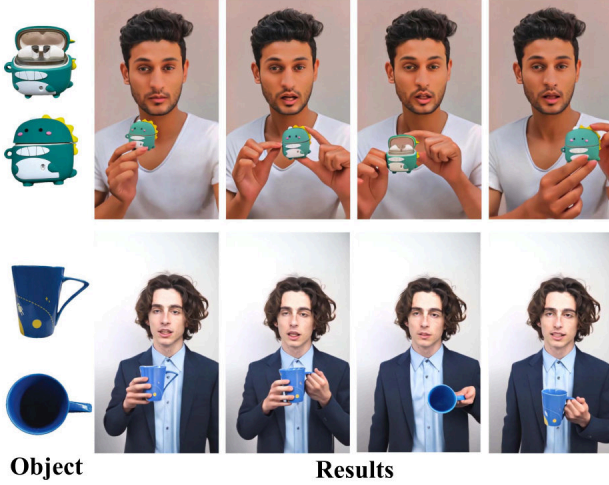


Fig. 11. AnchorCrafter is capable of performing more complex actions and adapting to arbitrary object viewpoints, demonstrating the adaptability and interactive ability of our model in intricate scenarios.

perspectives. To this end, the proposed method utilizes multiple images to capture intricate details from various viewpoints. For simplicity, the experiment emphasizes the front view and two side views of the object. In practice, these views can be substituted with other perspectives during fine-tuning, such as the object’s rear or even its interior. As shown in Fig. 11, leveraging depth perception and multi-view images, we successfully achieve the complex task of “opening the headphone case”. Furthermore, we demonstrate another example with a mug, where both the front view and the interior of the cup are captured. This experiment highlights that our framework can adapt to arbitrary viewpoints, effectively preserving object details and enabling realistic interactions. Overall, these results showcase our model’s adaptability and interactive capability in intricate scenarios.

2) *Generalization Across Shape-Similar Objects*: During inference, generate interaction videos with a specific object but don’t have access to it, can use a similarly shaped object instead. By capturing footage with a similar object, we can extract human poses, 3D hand mesh, and object depth as conditioning inputs. Our model exhibits a certain degree of generalization capability, enabling accurate inference even with similar-shaped objects. As shown in Fig. 12, we utilized videos recorded with different phone models to drive video generation for an iPhone. For objects with similar shapes, template videos can be captured to guide multiple objects.

3) *Generalization to Complex Backgrounds*: Real-world product promotion videos are often recorded in cluttered environments rather than clean studio settings. To evaluate the



Fig. 12. Generalization over similar object shapes. Driving iPhone interaction videos with another mobile phone.

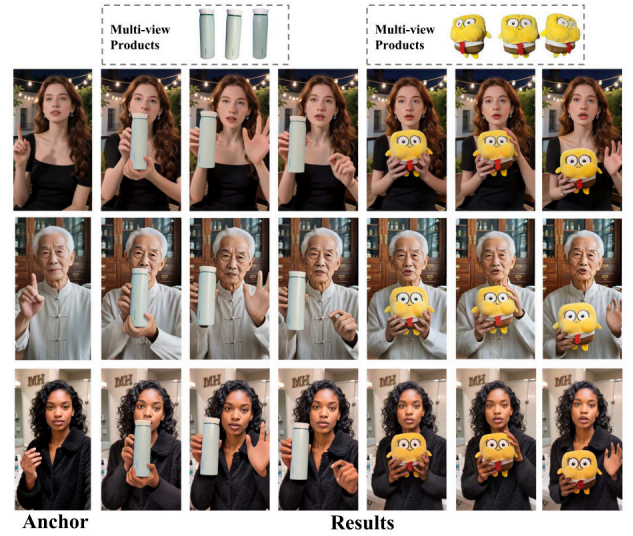


Fig. 13. AnchorCrafter results under cluttered backgrounds (human subjects from DreamActor [69]).

robustness of AnchorCrafter in such scenarios, we conducted experiments on diverse and complex backgrounds using human subjects from the DreamActor [69] dataset. As shown in Fig. 13, our method successfully preserves object appearance and generates natural human–object interactions.

4) *Pose-Confidence Strategy*: Errors introduced by pre-processing models can adversely affect the quality of generated videos. This issue can be mitigated through a confidence-based filtering strategy for pose estimation. Specifically, in the recognition results of DWPose, keypoints with confidence scores below a predefined threshold are discarded and not rendered. This design suppresses unreliable signals, ensuring that only high-confidence poses are propagated into subsequent stages. As illustrated in Fig. 14, this preliminary pose-confidence strategy reduces artifacts caused by inaccurate detections and enhances the robustness of AnchorCrafter in complex human–object interaction scenarios.

F. Limitations

AnchorCrafter is primarily evaluated on rigid objects and exhibits suboptimal performance when handling transparent objects (Fig. 15). Additionally, inaccuracies in preprocessing models such as DWPose introduce errors into the generated results.

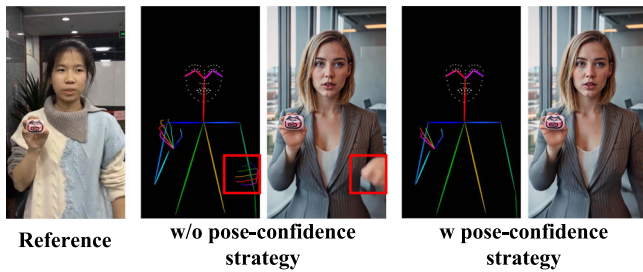


Fig. 14. Example of DWPose recognition results with and without the pose-confidence strategy. Lowconfidence keypoints are suppressed to improve robustness.

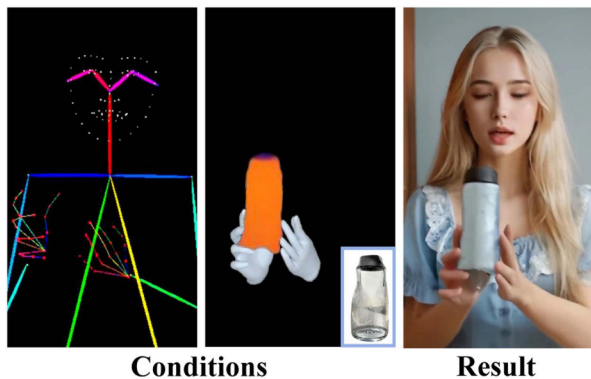


Fig. 15. Limitations. Handling transparent objects remains challenging.

In future work, we will conduct a more in-depth exploration of the challenges associated with transparent and non-rigid objects.

VI. CONCLUSION

We present AnchorCrafter, a novel diffusion-based system for anchor-style product promotion video generation that incorporates human-object interaction into pose-guided human video generation. Unlike approaches that typically rely on single-view object features, our system introduces a multi-view object feature fusion and disentangled appearance injection strategy to ensure consistent and high-fidelity reproduction of both anchors and products, a precise HOI-motion injection mechanism that integrates 3D hand meshes, depth maps, and occlusion-masking strategies to achieve controllable and realistic interactions, and an HOI-region reweighting loss that emphasizes interaction-critical regions during training to enhance fine-grained object details and realism. Extensive experiments demonstrate that AnchorCrafter significantly outperforms existing methods, achieving superior object appearance preservation, interaction fidelity, and motion consistency, and by addressing these unique challenges AnchorCrafter establishes a task-specific framework tailored for anchor-style product demonstrations, advancing the state of the art in human-object interactive video generation.

REFERENCES

[1] J. Xing et al., “Make-your-video: Customized video generation using textual and structural guidance,” *IEEE Trans. Visual. Comput. Graph.*, vol. 31, no. 2, pp. 1526–1541, Feb. 2025.

[2] Y. Zhang et al., “MotionCrafter: Plug-and-play motion guidance for diffusion models,” *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 10, pp. 8372–8384, Oct. 2025.

[3] L. Qu, J. Shang, M.-L. Lam, and H. Fu, “Controllable human video generation from sparse sketches,” *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 10, pp. 7243–7253, Oct. 2025.

[4] R. Shao, Y. Pang, Z. Zheng, J. Sun, and Y. Liu, “Human4Dit: 360-degree human video generation with 4D diffusion transformer,” *ACM Trans. Graph.*, vol. 43, no. 6, pp. 1–13, 2024.

[5] L. Hu, “Animate anyone: Consistent and controllable image-to-video synthesis for character animation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8153–8163.

[6] Z. Xu et al., “Magicanimate: Temporally consistent human image animation using diffusion model,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1481–1490.

[7] Y. Zhang et al., “MimicMotion: High-quality human motion video generation with confidence-aware pose guidance,” in *Proc. Int. Conf. Mach. Learn.*, 2025, pp. 1–15.

[8] Z. Huang et al., “Make-your-anchor: A diffusion-based 2D avatar generation framework,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6997–7006.

[9] S. Tu et al., “StableAnimator: High-quality identity-preserving human image animation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 21096–21106.

[10] D. Chang et al., “X-Dyna: Expressive dynamic human image animation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 5499–5509.

[11] Z. S. Xue, R. Luo, C. Chen, and K. Grauman, “HOI-swap: Swapping objects in videos with hand-object interaction awareness,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 77132–77164.

[12] B. Chen, C. Zhong, W. Xiang, Y. Geng, and X. Xie, “VirtualModel: Generating object-ID-retentive human-object interaction image by diffusion model for E-commerce marketing,” 2024, *arXiv:2405.09985*.

[13] Y. Fan et al., “Re-HOLD: Video hand object interaction reenactment via adaptive layout-instructed diffusion model,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 17550–17560.

[14] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. Adv. neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[16] K. Li, J. Zhang, Y. Liu, Y.-K. Lai, and Q. Dai, “PONA: Pose-guided non-local attention for human pose transfer,” *IEEE Trans. Image Process.*, vol. 29, pp. 9584–9599, 2020.

[17] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, “PISE: Person image synthesis and editing with decoupled GAN,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7982–7990.

[18] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3836–3847.

[19] T. Wang et al., “Disco: Disentangled control for realistic human dance generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9326–9336.

[20] S. Tu et al., “MotionEditor: Editing video motion via content-aware diffusion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 7882–7891.

[21] S. Zhu et al., “Champ: Controllable and consistent human image animation with 3D parametric guidance,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 145–162.

[22] Q. Gan et al., “HumanDiT: Pose-guided diffusion transformer for long-form human motion video generation,” 2025, *arXiv:2502.04847*.

[23] X. Wang et al., “UniAnimate: Taming unified video diffusion models for consistent human image animation,” *Sci. China Inf. Sci.*, vol. 68, 2025, Art. no. 200103.

[24] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu, “Vace: All-in-one video creation and editing,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 17191–17202.

[25] E. J. Hu et al., “LORA: Low-rank adaptation of large language models,” in *Proc. Int. Conf. Learn. Representations*, 2022, Art. no. 3.

[26] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[27] R. A. Güler, N. Neverova, and I. Kokkinos, “DensePose: Dense human pose estimation in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7297–7306.

- [28] G. Pavlakos et al., “Expressive body capture: 3D hands, face, and body from a single image,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.
- [29] Y. Liu et al., “Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild,” in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 7037–7047.
- [30] Q. Wu et al., “Dice: End-to-end deformation capture of hand-face interactions from a single image,” in *Proc. 13th Int. Conf. Learn. Representations*, 2025, pp. 1–24.
- [31] Z. Wu et al., “Doodle your motion: Sketch-guided human motion generation,” *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 9, pp. 6538–6549, Sep. 2025.
- [32] B. Ji, Y. Pan, Z. Liu, S. Tan, and X. Yang, “SPORT: From zero-shot prompts to real-time motion generation,” *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 10, pp. 7171–7183, Oct. 2025.
- [33] S. Xu, Z. Li, Y.-X. Wang, and L.-Y. Gui, “Interdiff: Generating 3D human-object interactions with physics-informed diffusion,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 14928–14 940.
- [34] H. Yi, J. Thies, M. J. Black, X. B. Peng, and D. Rempe, “Generating human interaction motions in scenes with text control,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 246–263.
- [35] W. Song et al., “Hoianimator: Generating text-prompt human-object animations using novel perceptible diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 811–820. [Online]. Available: <https://api.semanticscholar.org/CorpusID:272725056>
- [36] J. Bae, J. Won, D. Lim, C.-H. Min, and Y. M. Kim, “Pmp: Learning to physically interact with environments using part-wise motion priors,” in *Proc. ACM SIGGRAPH 2023 Conf. Proc.*, 2023, pp. 1–10.
- [37] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, “Imos: Intent-driven full-body motion synthesis for human-object interactions,” in *Computer Graphics Forum (CGF)*, vol. 42. Hoboken, NJ, USA: Wiley, 2023, pp. 1–12.
- [38] S. Fan, B. Du, X. Cai, B. Peng, and L. Sun, “TextiM: Part-aware interactive motion synthesis from text,” *Comput. Graph. Forum*, vol. 44, no. 2, 2025, pp. 1–9.
- [39] H. Hu, W. Wang, W. Zhou, and H. Li, “Hand-object interaction image generation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23805–23817.
- [40] Y. Men, Y. Yao, M. Cui, and L. Bo, “MIMO: Controllable character video synthesis with spatial decomposed modeling,” in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 21181–21191.
- [41] K. Liu et al., “Hoigen-1M: A large-scale dataset for human-object interaction video generation,” in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 24001–24010.
- [42] Z. Yang, A. Zeng, C. Yuan, and Y. Li, “Effective whole-body pose estimation with two-stages distillation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4210–4220.
- [43] K. Liu et al., “Motions as queries: One-stage multi-person holistic human motion capture,” in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 17529–17539.
- [44] Y. Jafarian and H. S. Park, “Learning high fidelity depths of dressed humans by watching social media dance videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12753–12762.
- [45] Z. Wang et al., “HumanVid: Demystifying training data for camera-controllable human image animation,” in *Proc. 38th Conf. Neural Inf. Process. Syst. Datasets Benchmarks*, 2024, pp. 20111–20131.
- [46] C. Li et al., “MVHumanNet: A large-scale dataset of multi-view daily dressing human captures with richer annotations for 3D human digitization,” 2025, *arXiv:2505.01838*.
- [47] Y. Pang et al., “Dreamdance: Animating human images by enriching 3D geometry cues from 2D poses,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 14039–14050.
- [48] A. Blattmann et al., “Stable video diffusion: Scaling latent video diffusion models to large datasets,” 2023, *arXiv:2311.15127*.
- [49] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, Munich, Germany, 2015, pp. 234–241.
- [50] D. P. Kingma, “Auto-encoding variational bayes,” 2013, *arXiv:1312.6114*.
- [51] J. Xing et al., “DynamCrafter: Animating open-domain images with video diffusion priors,” in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 399–417.
- [52] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [53] M. Oquab et al., “DINOv2: Learning robust visual features without supervision,” 2023, *arXiv:2304.07193*.
- [54] S. Tang et al., “Mvdiffusion: A dense high-resolution multi-view diffusion model for single or sparse-view 3D object reconstruction,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 175–191.
- [55] Y. Yue, A. Das, F. Engelmann, S. Tang, and J. E. Lenssen, “Improving 2D feature representations by 3D-aware fine-tuning,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 57–74.
- [56] M. B. S. Millán, A. Dai, and M. Nießner, “Animating the uncaptured: Humanoid mesh animation with video diffusion models,” 2025, *arXiv:2503.15996*.
- [57] L. Li, W. Wang, and Y. Yang, “Human-object interaction detection collaborated with large relation-driven diffusion models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 23655–23678.
- [58] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, “Reconstructing hands in 3D with transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9826–9836.
- [59] N. Ravi et al., “SAM2: Segment anything in images and videos,” 2024, *arXiv:2408.00714*.
- [60] K. Xian, J. Peng, Z. Cao, J. Zhang, and G. Lin, “ViTa: Video transformer adaptor for robust video depth estimation,” *IEEE Trans. Multimedia*, vol. 26, pp. 3302–3316, 2024.
- [61] X. Ji et al., “Sonic: Shifting focus to global audio perception in portrait animation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 193–203.
- [62] W. Chai, X. Guo, G. Wang, and Y. Lu, “Stablevideo: Text-driven consistency-aware diffusion video editing,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 23040–230 50.
- [63] S. Zhang, J. Zhuang, Z. Zhang, Y. Shan, and Y. Tang, “FlexiAct: Towards flexible action control in heterogeneous scenarios,” in *Proc. ACM SIGGRAPH 2025 Conf.*, 2025, pp. 1–11.
- [64] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.
- [65] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” 2018, *arXiv:1812.01717*.
- [66] Z. Huang et al., “VBench: Comprehensive benchmark suite for video generative models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 21807–21818.
- [67] M. Kim, A. K. Jain, and X. Liu, “Adaface: Quality adaptive margin for face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18750–18 759.
- [68] S. Bai et al., “Qwen2. 5-VL technical report,” 2025, *arXiv:2502.13923*.
- [69] L. Wang et al., “Dreamactor-H1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers,” 2025, *arXiv:2506.10568*.



Ziyi Xu received the BE degree from the Zhejiang University of Technology, in 2019. She is currently working toward the MS degree with the University of Chinese Academy of Sciences (UCAS), under the supervision of the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her research interests lie in digital human and video generation.



Ziyao Huang received the BE degree from Beihang University, in 2017, and the MS degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2022. He is currently working toward the PhD degree with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests lie in digital human and video generation.



Juan Cao received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2008. She is a professor with the Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include multimedia content analysis and fake multimedia detection. She has more than 90 publications in international journals and conferences, including *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Image Processing*, KDD, CVPR, etc.



Yong Zhang received the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2018. From 2015 to 2017, he was a visiting scholar with the Rensselaer Polytechnic Institute. He was with the Tencent AI Lab and now with Meituan. His research interests include computer vision and machine learning.



Yuchen Wang received the BE degree from Hebei Agricultural University, in 2019. He is currently working toward the MS degree with the University of Chinese Academy of Sciences (UCAS), under the supervision of the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests lie in image generation, digital human, and video generation.



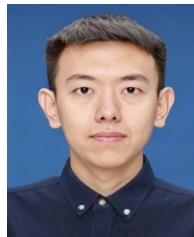
Xiaodong Cun (Member, IEEE) received the BSc degree in computer science from Xidian University, the MS and PhD degrees from the Department of Computer and Information Science, University of Macau, in 2018 and 2021 respectively. He is currently an assistant professor in Great Bay University. Before that, he was a senior researcher in the Visual Computing Center of Tencent AI Lab from 2021 to 2024. His current research interests include image/video generation, translation and editing.



Linchao Bao received the MS degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, and the PhD degree in computer science from the City University of Hong Kong, in 2015. He was a research intern with Adobe Research from 2013 to 2014 and worked for DJI as an Algorithm Engineer from 2015 to 2016. He is currently a Principal Research Scientist with Tencent AI Lab. His research interests include computer vision and graphics.



Qing Shuai received the PhD degree in computer science from Zhejiang University, Hangzhou, China, in 2024, under the supervision of Xiaowei Zhou. He is currently working with Tencent (2024.7–), where his focus is on human motion capture and generation under multimodal inputs. His research interests lie at the intersection of computer vision and computer graphics, with a particular emphasis on 3D human pose estimation and generation, 3D reconstruction, and novel view synthesis.



Fan Tang (Member, IEEE) received the BSc degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.