

HOMA: Towards Generic Human-Object Interaction in Multimodal Driven Human Animation with Weak Conditions

ZIYAO HUANG*, University of Chinese Academy of Sciences, China

ZIXIANG ZHOU*, Tencent, China

JUAN CAO, University of Chinese Academy of Sciences, China

YIFENG MA, Tencent, China

YI CHEN, Tencent, China

ZEJING RAO, University of Chinese Academy of Sciences, China

ZHIYONG XU, Tencent, China

HONGMEI WANG, Tencent, China

QIN LIN, Tencent, China

YUAN ZHOU, Tencent, China

QINGLIN LU[†], Tencent, China

FAN TANG[†], University of Chinese Academy of Sciences, China

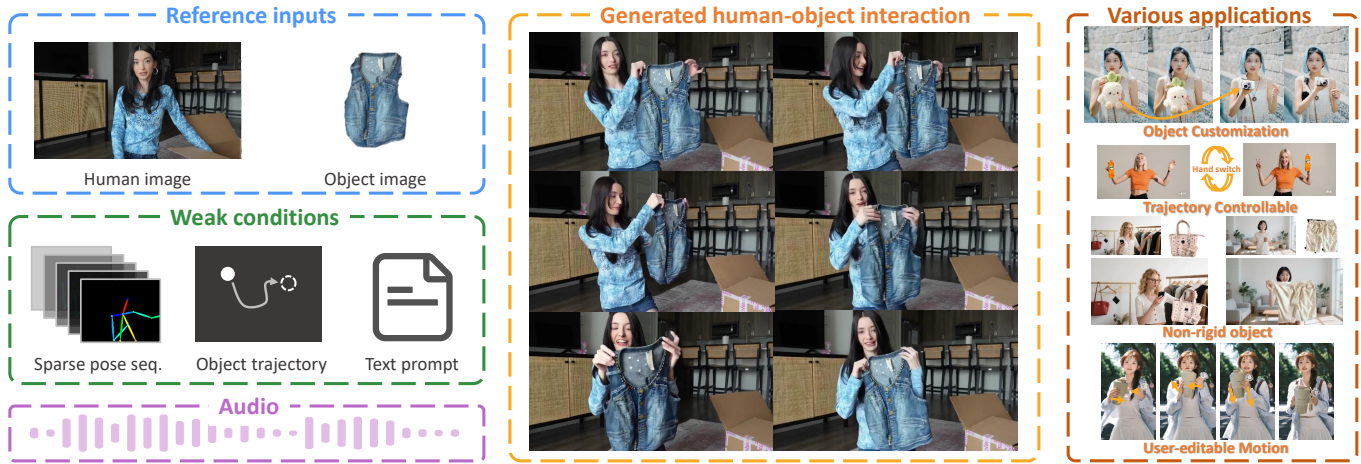


Fig. 1. HOMA is a multimodal-driven human animation framework for human-object interaction (HOI) generation, utilizing weak conditions to enable HOI video generation, delivering high-fidelity results from a single object input while maintaining strong generalizability across applications.

While recent advances in human-object interaction (HOI) video generation showcase promising capabilities for synthesizing coordinated human-object

*Both authors contributed equally to this research.

[†] Corresponding authors.

Authors' Contact Information: Ziyao Huang, University of Chinese Academy of Sciences, Beijing, China, huangziyao19f@ict.ac.cn; Zixiang Zhou, Tencent, Beijing, China, zxxzhou0916@163.com; Juan Cao, University of Chinese Academy of Sciences, Beijing, China, caojuan@ict.ac.cn; Yifeng Ma, Tencent, Beijing, China, mayf18@mails.tsinghua.edu.cn; Yi Chen, Tencent, Beijing, China, yi_chen@sjtu.edu.cn; Zejing Rao, University of Chinese Academy of Sciences, Beijing, China, zjr24@163.com; Zhiyong Xu, Tencent, Beijing, China, xuzhy16@qq.com; Hongmei Wang, Tencent, Beijing, China, whm712@tju.edu.cn; Qin Lin, Tencent, Shenzhen, China, angelqin@tencent.com; Yuan Zhou, Tencent, Beijing, China, zy1651722481@163.com; Qinglin Lu, Tencent, Shenzhen, China, qinglinlu@tencent.com; Fan Tang, University of Chinese Academy of Sciences, Beijing, China, tfan.108@gmail.com.



This work is licensed under a Creative Commons Attribution 4.0 International License. SA Conference Papers '25, Hong Kong, Hong Kong

dynamics, existing methods remain constrained by their reliance on meticulously curated motion sequences and actor-specific data, thereby limiting practical scalability and user accessibility. Furthermore, generalization to novel object appearances and interaction scenarios remains understudied. To address these limitations, we propose HOMA, a weakly conditioned multimodal-driven HOI video generation framework that introduces sparse, decoupled motion guidance to enhance controllability and reduce dependency on stringent input conditions. Our approach encodes appearance and motion signals into the dual input space of a multimodal diffusion transformer (MMDiT), fusing them within a shared context space to enable temporally consistent and physically plausible interactions. To optimize learning efficiency and feature injection accuracy, we introduce a parameter-space HOI adapter initialized with pretrained MMDiT weights to preserve prior knowledge while enabling efficient adaptation. Additionally, we design a facial cross-attention adapter for audio-driven lip synchronization,

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2137-3/25/12

<https://doi.org/10.1145/3757377.3763861>

ensuring anatomically accurate speech animation. Extensive experiments demonstrate that HOMA achieves state-of-the-art performance in interaction naturalness and generalization under weak supervision, outperforming existing methods by significant margins. We further illustrate HOMA’s versatility through diverse applications, including text-conditioned generation and interactive object manipulation, facilitated by a user-friendly demo interface. The project page is <https://bone-11.github.io/homa-page/>.

CCS Concepts: • **Computing methodologies** → **Computer vision; Animation**.

Additional Key Words and Phrases: Video generation, human-object interaction, weak condition

ACM Reference Format:

Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, Qinglin Lu, and Fan Tang. 2025. HOMA: Towards Generic Human-Object Interaction in Multimodal Driven Human Animation with Weak Conditions. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers ’25)*, December 15–18, 2025, Hong Kong, Hong Kong. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3757377.3763861>

1 Introduction

Human image animation for generating human-centric videos [Chang et al. 2023; Hu 2024; Wang et al. 2024b; Xu et al. 2024b] has been extensively studied in recent years. Advancements in video diffusion models have markedly enhanced the quality of synthesized human motion, while concurrently enabling diverse driving modalities such as pose [Hu 2024; Zhang et al. 2024] and audio [Ji et al. 2024; Lin et al. 2025; Meng et al. 2024; Zhang et al. 2023]. Building upon single-human-driven methodologies, recent breakthroughs have integrated human-object interaction (HOI) into these frameworks [Fan et al. 2025; Pang et al. 2024; Xu et al. 2024a; Xue et al. 2024], enabling the generation of interactive objects alongside their corresponding motion dynamics. These approaches offer considerable potential for generating realistic human-object interactions in videos, a capability with broad applications across diverse domains, including online retail, digital advertising, and film production.

Despite the progress in HOI video generation, significant challenges persist in achieving robust generalization and applicability. A major limitation stems from the reliance on strong motion control mechanisms, such as actor-captured motion sequences [Fan et al. 2025; Xu et al. 2024a; Xue et al. 2024] or computationally intensive 3D modeling [Pang et al. 2024]. These requirements impose constraints on real-world usability. Furthermore, while existing methods demonstrate proficiency in handling simple rigid objects, their generalization to diverse object categories—including rigid objects with intricate geometries and surface properties, as well as deformable or articulated objects—remains understudied and inadequately addressed.

To address these challenges, we propose HOMA, a novel framework built upon HunyuanVideo [Kong et al. 2024], designed to provide a more practical and flexible solution for HOI video generation through a multimodal-driven human animation pipeline. To mitigate the reliance on strongly supervised motion sequences, we explore a weak motion control paradigm that enhances both user editability and model adaptability. Specifically, we introduce a spatially sparse and decoupled motion condition, where only a

single arm skeleton and the trajectory of the object’s center point are employed as minimal representations of human and object motion. This formulation captures essential interaction dynamics with the weakest possible control signal. Built upon this weak condition model, our framework enables the animation of a reference human and the seamless injection of interactive objects into the generated video, achieving controllable motion and natural interactions. Additionally, we incorporate text and audio inputs as complementary modalities, offering further control flexibility while maintaining the framework’s sparse input philosophy.

Under weak condition control, the model is expected to produce reasonable outputs despite missing substantial guidance. In such scenarios of weak guidance and strong inference, we argue that the model should be encouraged to leverage generative priors to fill in the uncontrolled components while still adhering to the given controls. To address this issue, we introduce a novel approach built upon the input context space of multi-modal diffusion transformer (MMDiT) [Esser et al. 2024] for controllable HOI generation. Specifically, we encode appearance and motion conditions into the context space of the MMDiT framework, ensuring precise and coherent integration of both visual and kinematic information. To enhance object appearance consistency and improve the plausibility of human-object interactions, we introduce an HOI Adapter, which is also in MMDiT architecture and initialized with the pretrained parameters from HunyuanVideo [Kong et al. 2024], leveraging prior knowledge to accelerate convergence and achieve superior consistency and interaction realism. For lip-syncing control, we propose an Audio Cross-attention Adapter that injects speech signals into facial regions of generated videos, enabling temporally aligned and anatomically accurate lip movements. By exploiting the model’s generative priors, our approach effectively enhances the consistency of object injection. The data pipeline and training strategies are designed to optimize the performance of the HOMA framework, including a depth-aware approach to filter out real-world video data without human-object interaction.

Experimental results demonstrate that HOMA outperforms existing methods, with improved generalization capabilities of object appearance preservation and interaction naturalness. Various applications under different weak conditions are also demonstrated, including motion generation guided by different types of weak pose control, text-conditioned generation, and diverse modifications to object appearance and trajectory. Additionally, we present an interactive demo to support our weak-condition framework and facilitate user-guided refinement. In summary, our contributions are as follows:

- We propose HOMA, a novel framework that introduces weak motion guidance to reduce the dependency on strong motion conditions for HOI video generation, achieving both flexibility and precision for real-world applications.
- Under limited weak conditions, we explore a novel solution to inject HOI appearance and motion in a manner that aligns with the model’s generative prior and enhances consistency in conditioning and naturalness in interaction generation.
- We conduct comprehensive quantitative and qualitative evaluations to validate the effectiveness of HOMA, benchmarking it

against state-of-the-art diffusion-based methods for HOI video generation and showcasing multiple application scenarios.

2 Related Work

2.1 Motion-controllable Human Video Generation

With the advancement of video generation models [Blattmann et al. 2023; Kong et al. 2024; Wang et al. 2025a; Zhou et al. 2024], the controllability of human subjects has become a central research focus. Key challenges include accurately controlling human motions using multimodal inputs such as pose and audio while preserving identity and appearance. For pose-driven control, the research trajectory includes methods from UNet-based approaches [Hu 2024; Huang et al. 2024; Tu et al. 2024; Wang et al. 2024d,c; Zhang et al. 2024; Zhu et al. 2024] to DiT-based methods [Shao et al. 2024; Wang et al. 2025b], which progressively improves motion fidelity and temporal coherence. For audio-driven control, representative works from head animation [Chen et al. 2025; Ji et al. 2024; Tian et al. 2024; Zhang et al. 2023], semi-body animation [Corona et al. 2024; Lin et al. 2024; Meng et al. 2024] to full-body animation [Lin et al. 2025]. However, these methods primarily focus on single-subject animation. Even as model capacities improve, enabling plausible motion generation from images depicting well-structured human interactions, the lack of attention to non-human subjects (e.g., objects, environments) still constrains the ability of current systems to achieve controllable human-object interaction video generation.

2.2 Human-object Interaction Video Generation

As controllable human video generation continues to improve, a key emerging agenda is how to enable human-object interaction grounded on controllable human motion. Outside of video generation, researchers have focused on generating human-object and human-scene motion [Ghosh et al. 2023; Jiang et al. 2024a,b; Li et al. 2023; Peng et al. 2023; Xu et al. 2023]. Within video generation, some approaches aim to retain real human-object interactions in existing videos while replacing either the human or the object. For example, MIMO [Men et al. 2024] and AnimateAnyone2 [Hu et al. 2025] attempt to substitute human subjects in complex dynamic interaction scenes, whereas HOI-Swap [Xue et al. 2024] and Re-HoLD [Fan et al. 2025] focus on replacing handheld objects within a real video. However, these approaches rely on real videos as a foundation and are thus limited to editing existing content, lacking the ability to generate human-object interactions from scratch.

Other works have explored object injection in human animation to human-object interaction video generation from scratch. For instance, AnchorCrafter [Xu et al. 2024a] fine-tunes target objects and utilizes multi-conditional inputs such as pose and depth to represent interaction-aware human motion. ManiVideo [Pang et al. 2024] achieves accurate control of interaction motion through precise 3D modeling of both the human and the object. Nevertheless, these methods rely on strong conditional inputs, making them difficult to apply in unconstrained scenarios. Our proposed HOMA explores a weak HOI condition for user-friendly generation and accomplishes high generalizability of the objects.

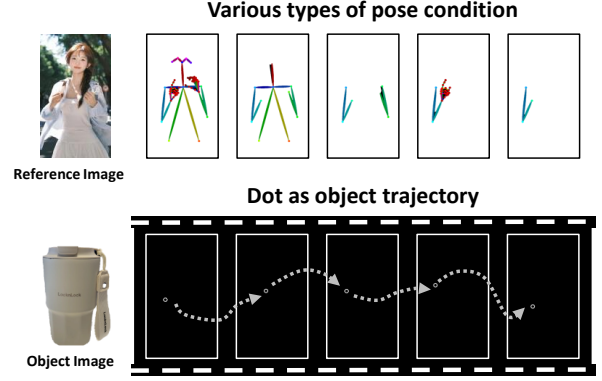


Fig. 2. We explore a weak HOI condition, where each component of the human pose can be optionally removed, and the object’s motion is a sequence of dots indicating the center points of its trajectory.

3 Method

In this section, we propose the weak condition for HOI video generation in Sec. 3.1, our weak condition driven network architecture in Sec. 3.2, and the training strategies and data curation in Sec. 3.3.

3.1 Weak HOI Condition Setting

A core challenge in HOI video generation is acquiring expressive yet editable motion inputs. Existing methods rely on actor performances or just swapping objects, limiting scalability and user controllability. We propose a sparse, user-editable input framework that leverages minimal but semantically meaningful conditions to generate realistic, physically plausible interactions.

We formulate weakly conditioned HOI video generation as animating a reference human image I_H with an inserted object image I_O , producing an output video $Y = \{y^{1:n}\}$ with human-object interaction of n frames. Two separate motion sequences guide the human-object interactions under intentionally weak controls (Fig. 2). Human motion is driven by sparse skeletal sequences $P = \{p^{1:n}\}$, selectively removing all body parts except the arms in contact with the object; finger skeletons are optionally excluded to reduce difficulty for users to specify, with finger motion inferred contextually. Object trajectory is represented by a dot sequence $D = \{d^{1:n}\}$ indicating the object’s center point, enabling autonomous orientation inference and avoiding the unnatural constraints of naive bounding-box methods.

Our framework also supports text T and audio $A = \{a^{1:n}\}$ as complementary control signals. The text offers intuitive semantic guidance, while audio enables the animated human not only to act but also to talk as expected in an end-to-end manner.

3.2 Architecture

3.2.1 Preliminaries. We propose our framework to accomplish the weak-conditioned HOI video generation. Our network architecture is modified from HunyuanVideo-T2V [Kong et al. 2024], including an MMDiT-based [Esser et al. 2024] video diffusion model and a 3D VAE model [Kingma 2013] with encoder Enc and decoder Dec .

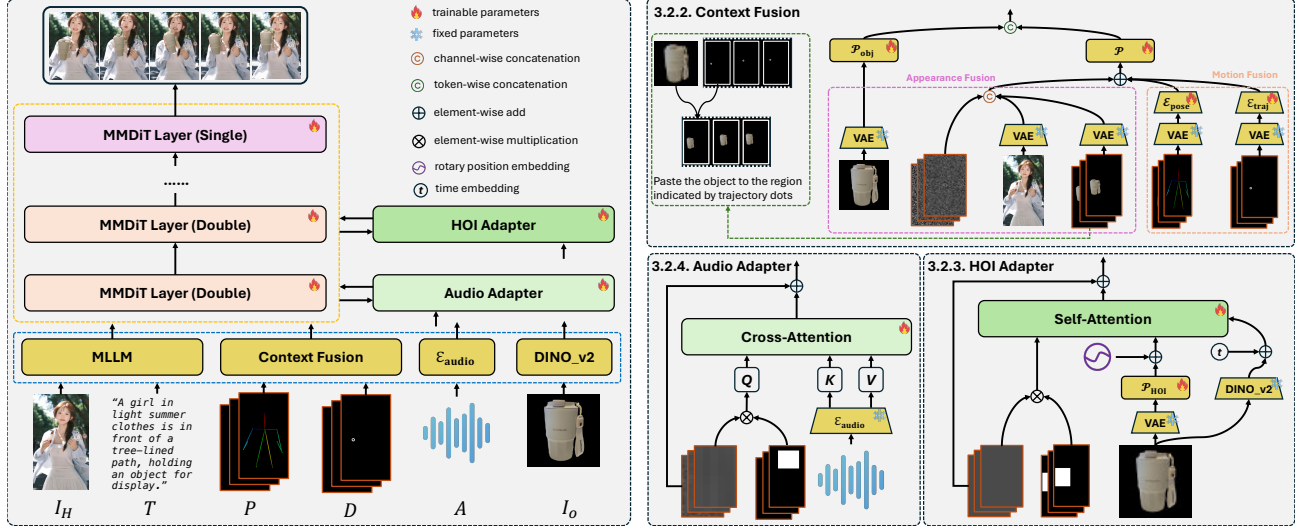


Fig. 3. Architecture of HOMA. HOMA is built upon the multi-modal diffusion transformer and is composed of context fusion, HOI adapter, and audio adapter.

During training, the video sequence X is encoded into the latent space as $Z_0 = \text{Enc}(X) = \{z_0^{1:n}\}$ and flow matching [Lipman et al. 2022] is used as the training objective. During inference, a set of initial latent variables $U = \{u^{1:n}\}$ is sampled from standard Gaussian noise and progressively denoised over T timesteps to produce an estimate $\hat{Z}_0$. The final output video is then obtained via decoding, $Y = \text{Dec}(\hat{Z}_0) = \{y^{1:n}\}$.

3.2.2 Context Fusion. To effectively leverage the model’s generative prior, we propose a context fusion strategy for object and human injection in the input spaces. The overall architecture is illustrated in Fig. 3. Context fusion involves two spaces, namely, the latent space and token space. The latent space is encoded by the VAE encoder as $Z = \text{Enc}(X)$, RoPE [Su et al. 2024] is added to provide position information. Subsequently, a patch embedding layer $\mathcal{P}$ maps the latent features into the token space, yielding $H = \mathcal{P}(Z)$. In the following, we introduce our context fusion strategy from two perspectives: appearance and motion.

Appearance Context Fusion. The main purpose of this module is to maintain the appearance of human image I_H and injected object I_O during animation. Due to the different requirements of humans and objects, we propose a separate manner to inject their appearance. For the human image I_H to be animated, we propose a channel-concat style injection approach performed on input latent space:

$$Z_{cat} = \text{concat}_c(Z, Z_{ref}), \quad (1)$$

where Z is the input noise, Z_{ref} is the VAE feature extracted from I_H , $Z, Z_{ref} \in [b, f, h, w, c]$, $Z_{cat} \in [b, f, h, w, 3c]$, and concat_c is the channel-wise concat operation. The output embedding Z_{cat} is then input into the DiT blocks. The proposed channel concatenation scheme aims to maintain the structural and semantic content of the reference image. As for the object image I_O , we perform object injection from both the token space and the latent space. In token

space, we adopt a temporal-concat style scheme:

$$H = \text{concat}_f(H_{obj}, H), \quad (2)$$

where $H_{obj} \in [b, 1, h, w, c]$ is the object feature extracted from I_O as $H_{obj} = \mathcal{P}(\text{Enc}(I_O))$ and concat_f is the temporal-wise concat operation. After temporal-concat, the output H is in $[b, f+1, h, w, c]$. The RoPE is set with offset=-1 for the temporal-concat image. Unlike reference human images that aim to preserve original content, object appearances are inserted into the video content. Therefore, we adopt an injection strategy that aligns more closely with the video generation context—specifically, the object is injected at frame -1 and emerges in the subsequent frames. In addition, given the effective injection of the human representation in the latent space, we introduce object appearance features into the latent space to further improve stability. The object is copy-pasted to the image region indicated by trajectory to serve as strong appearance prior:

$$Z_{obj|D} = \{Z_{obj}^{i:n} | Z_{obj}^i = Z_{obj|D}^i\}, \quad (3)$$

and the features are then concatenated along the channel dimension in the latent space. The final channel concatenation is:

$$Z_{cat} = \text{concat}_c(Z, Z_{ref}, Z_{obj|D}), \quad (4)$$

where the final dimension of Z_{cat} is in $[b, f, h, w, 3c]$. Beyond improving injection consistency, the size of spatially aligned and pasted object $Z_{obj|D}$ also affects its final scale in generated video. Importantly, ours does not suffer from noticeable copy-paste artifacts.

To enhance global semantics, we extract LLaVA [Liu et al. 2023] visual features from both human and object images and fuse them with the feature of text T , and feed it to MMDiT’s text branch.

Motion Context Fusion. We perform motion context fusion in latent space as well. For the injection of human pose and object trajectory, we also adopt a separate-then-merge style module, where the two motion conditions are processed independently before being

fused. For the pose sequence P and the object trajectory sequence D , we first use the 3D VAE to map them into feature representations $Z_{pose} = Enc(P)$ and $Z_{obj} = Enc(D)$, and then use separate one-layer CNN $\mathcal{E}_{pose}$ and $\mathcal{E}_{traj}$ to project them into latent space as $Z_{pose} = \mathcal{E}_{pose}(Z_{pose})$ and $Z_{obj} = \mathcal{E}_{traj}(Z_{obj})$. Human pose and object trajectory features are then added to the appearance channel-concat latent to get the final input:

$$Z = Z_{cat} + Z_{pose} + Z_{traj}. \quad (5)$$

We observe that the separate-then-merge design leads to better appearance preservation. Analysis is provided in the ablation study.

3.2.3 HOI adapter. We observed that the model struggled to learn sufficient information about the appearance of the object. We attribute this to weak gradient propagation when injection is limited to the input layer, due to the network’s depth. To address this, we propose the HOI Adapter—a multi-layer injection strategy that introduces object features at several points within the network. We posit that effective adapter design relies on leveraging the original model’s parameter space, rather than constructing entirely new representational branches. Therefore, we modify the structure of original MMDiT’s double-stream block, where the text token is replaced by object token H_{obj} , and the weights are initialized with pretrained parameters. The HOI Adapter is formulated as:

$$H = H + Mask_{obj} \otimes SelfAttn([H, H_{obj} + RoPE(-2)]), \quad (6)$$

where the object feature H_{obj} is token-concat with the hidden feature H , followed by a self-attention layer for feature interaction. The adapter output is added back to H_{noise} within the object region $Mask_{obj}$ to provide a precise location. We also use a DINO [Caron et al. 2021] feature of I_O combined with time embedding to provide semantic information. By injecting object information across multiple layers, the model learns object injection more effectively and accurately. To preserve compatibility with the original representation space, we initialize the adapter’s self-attention weights from the corresponding MMDiT blocks. HOI Adapters are inserted only into the even-numbered layers.

3.2.4 Audio Cross-attention Adapter. We propose a cross-attention style injection module for audio input. The input audio feature A is mapped with an MLP layer into H_{audio} , then we utilize a cross-attention layer to modify the face region:

$$H = H + Mask_{face} \otimes Softmax\left(\frac{Q_H K_A^T}{\sqrt{d}}\right) \cdot V_A, \quad (7)$$

where $Q_H = W_{QH} \cdot H$, $K_A = W_{KA} \cdot H_A$, $V_A = W_{VA} \cdot H_A$, and $Mask_{face}$ is the mask with the face-region. The cross-attention adapters are added to the odd-numbered layers of MMDiT blocks.

3.3 Training Strategy and Data Curation

We adopt a three-stage training strategy. Stage one focuses on learning human pose control using human-only datasets, with appearance injected via channel-concatenated reference images and a pose encoder $\mathcal{E}_{pose}$. In stage two, we introduce HOI data at 512×512 resolution, enabling the model to learn object control while incorporating pose and audio conditions. $\mathcal{E}_{traj}$ is initialized from $\mathcal{E}_{pose}$.

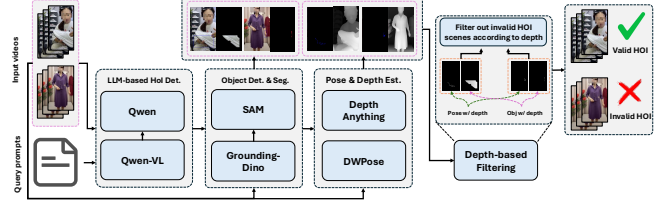


Fig. 4. Depth-aware HOI Data Curation.

Stage three increases the resolution to 512×896 to enhance visual quality and high-resolution generation capability.

To address the scarcity of HOI training data and annotations, we propose an automated pipeline for selecting human-object interaction (HOI) video clips, comprising three stages:

- **Step-I: HOI Recognition.** We prompt Qwen-VL [Wang et al. 2024a] to identify whether a video contains HOI and caption objects.
- **Step-II: HOI Extraction.** Using the generated object captions, we apply Grounding-DINO [Liu et al. 2024] and SAM2 [Ravi et al. 2024] to perform object detection and segmentation.
- **Step-III: HOI Filtering.** To reduce false positives (e.g., irrelevant background objects), we use DepthAnythingV2 [Yang et al. 2024] to estimate depth maps. We compare the mean depths of the segmented object and hand regions, retaining only those with similar depths to ensure plausible physical interaction.

4 Experiments

4.1 Experimental Settings

4.1.1 Implementation Details. We use a fixed learning rate of 1e-5 across all stages. Training is conducted on 48×96G GPUs with a total batch size of twelve. The total training consists of 16,000 steps in stage one, followed by 2,000 steps in stage two, and 5,000 steps in stage three. To improve generalization and robustness, we apply data augmentation to the object images, including random scaling, rotation, and position shifting. Human pose information is extracted using DWPose [Yang et al. 2023], while audio features (when applicable) are processed using Librosa [McFee et al. 2015]. Each generated video has a duration of five seconds.

4.1.2 Dataset. We apply depth-aware HOI(Human-Object Interaction) data curation to internet-sourced videos to support training under diverse interaction scenarios and finally collect approximately 140 hours of training data. For evaluation, we conduct experiments on two test sets: **AnchorCraft test set:** The AnchorCrafter [Xu et al. 2024a] test set includes four objects and four persons with simple object types and basic interaction motions. We reorganize the AnchorCraft test set via resampling, resulting in 80 clips. **Self-collected test set:** To evaluate more challenging scenarios, we construct a custom test set from Internet videos, containing 100 clips with complex human-object interactions and diverse object appearances. Audio inputs are randomly collected for both test sets. We conduct quantitative and qualitative evaluations on each test set to assess the model’s generalization ability across varying levels of interaction difficulty. **Due to page limit, AnchorCrafter test set results are in the supplementary materials.**

4.1.3 Comparison Methods. We compare our method with the following methods: AnchorCrafter [Xu et al. 2024a], VACE-14B [Jiang et al. 2025], MimicMotion [Zhang et al. 2024], StableAnimator [Tu et al. 2024], UniAnimate-DiT [Wang et al. 2025b], and EchoMimic-v2 [Meng et al. 2024]. AnchorCrafter [Xu et al. 2024a] is a human-object interaction video generation approach that requires fine-tuning on the target object. VACE-14B [Jiang et al. 2025] is a general video editing framework based on the 14B video model Wan [Wang et al. 2025a] and we use the pose control setting. MimicMotion [Zhang et al. 2024], and UniAnimate-DiT [Wang et al. 2025b] generate motion from a person image and a sequence of poses. EchoMimic-v2 [Meng et al. 2024] takes audio and a human image as input. For methods not originally designed for human-object interaction, we composite the object into the person image to approximate an interactive scenario. For a fair comparison, all compared methods utilize the original motion inputs, and our method uses the whole full pose for inference. AnchorCrafter requires finetuning on the target object and is compared only in the AnchorCrafter test set.

4.1.4 Metrics. We evaluate our method using several established metrics. For visual quality, we use FID [Heusel et al. 2017] and FVD [Unterthiner et al. 2018] to evaluate the image and video quality. Object CLIP (OC) [Xu et al. 2024a] is used to evaluate the object consistency by computing the CLIP [Radford et al. 2021] similarity of the objects between ground-truth and generated videos within the bounding box. Additionally, Subject Consistency (SC) [Xue et al. 2024], Background Consistency (BC) [Xue et al. 2024], and Motion Smoothness (MS) [Xue et al. 2024] are employed to measure general video quality. For hand motion quality, we calculate Hand Agreement Score (HAS) [Xue et al. 2024] and Hand fidelity (HF) [Xue et al. 2024]. To evaluate hand-object interaction quality (HOQ), we use InternVL3-78B [Chen et al. 2024; Zhu et al. 2025] to identify videos with good interaction quality. To evaluate lip synchronization, we adopt the commonly used Sync-C [Chung and Zisserman 2017], which measures the alignment confidence between visual and audio.

4.2 Comparison Results

4.2.1 Quantitative Results. As shown in Table 1, our method achieves the best FID and FVD scores, indicating superior visual quality. The Object CLIP (OC) score is notably the highest among all methods. This highlights the strong ability of our framework to preserve object appearance. The results on SC, BC, and MS are comparable to others, reflecting solid structural coherence and smooth motion. Moreover, on our self-collected test set with more complex interaction scenarios, we achieve the highest Hand Agreement Score (HAS), indicating strong generalization under challenging conditions. Given the limitations of objective metrics in evaluating fine-grained hand-object interactions, the user study provides a more comprehensive and trustworthy assessment of interaction quality. Sync-C scores are also comparable to Echomimic-v2, validating the effectiveness of our audio injection design.

4.2.2 Qualitative Results. In Fig. 5, we present the qualitative comparisons of the self-collected test set, where our method achieves better preservation of human and object appearances, along with more

Table 1. Quantitative results of our method compared with SOTAs.

Method	Self-collected test set									
	FID↓	FVD↓	OC↑	SC↑	BC↑	HAS↑	HF↑	MS↑	HOQ↑	Sync-C↑
MimicMotion	83.56	1072.02	81.70	93.48	93.82	87.51	99.16	99.08	76	-
Echomimic-v2	100.20	1332.06	79.82	94.65	92.92	86.84	99.54	99.04	66	4.12
StableAnimator	125.84	1530.48	79.26	88.19	91.36	83.98	94.97	98.42	41	-
UniAnimate-DiT	83.18	844.32	83.74	95.82	94.96	83.17	99.48	99.27	83	-
VACE-14B	77.07	968.79	82.88	94.73	95.59	75.74	99.87	99.04	90	-
HOMA (ours)	51.60	502.69	90.05	95.19	94.45	87.93	98.76	99.14	97	4.33

Table 2. Ablation results with 512×512 resolutions.

Method	FID↓	FVD↓	OC↑	SC↑	BC↑	HAS↑	HF↑	MS↑	Sync-C↑
HOMA (ours)	38.32	386.83	88.75	93.87	93.65	88.23	98.70	98.78	4.81
w/ bbox	76.73	720.24	82.23	92.86	93.86	85.73	98.36	98.89	-
w/ gaussian dot	86.54	691.56	80.74	94.46	94.60	85.69	98.88	99.05	-
w/o token concat	74.38	681.18	82.18	93.88	93.86	84.73	98.34	98.64	-
w/o channel concat	74.38	799.24	84.90	94.04	94.14	85.29	97.30	98.93	-
w/ fix-copy	58.48	588.63	85.88	93.99	94.23	85.41	97.96	98.67	-
single motion enc.	50.34	505.91	86.82	94.31	94.42	85.62	98.86	98.99	-
w/o adapter	62.09	582.38	82.01	92.04	93.60	84.87	99.48	98.90	-
w/ cross-attention	52.65	592.44	83.39	92.22	93.77	85.58	98.59	98.97	-
even-numbered audio layers	41.14	399.90	86.34	93.54	93.49	87.37	98.05	98.76	4.76

consistent and visually stable results. MimicMotion produces reasonably accurate hand poses but fails to generate the object. Echomimic-v2 suffers from inconsistent object generation and noticeable flickering artifacts on the character, while StableAnimator tends to produce flawed or anatomically incorrect hands. UniAnimate-DiT maintains good subject consistency but struggles to generate realistic and coherent interaction motions. VACE-14B produces high-quality videos overall but fails to faithfully preserve the appearance of the object. In contrast, our method generates high-quality human-object interaction videos that are visually coherent, temporally stable, and faithful to the reference inputs.

4.3 Ablation Studies and Discussion

To facilitate efficient comparison across different variants, all ablation experiments are conducted at a resolution of 512×512 with 2,000 training steps. The self-collected test set is utilized to verify the ablations studies. We focus on key components of our framework, including the object motion setting and architecture variants.

4.3.1 Analysis of Object Motion Setting. To evaluate our dot-based object motion strategy, we compare three variants: bounding box (bbox), Gaussian-colored dot (Gaussian dot), and the proposed dot trajectory (dot). As shown in Fig.6 and Table2, bbox control enforces explicit constraints on position, size, and orientation, but often results in rigid or unnatural motion during complex interactions. Gaussian-dot is used as input to the motion encoder, similar to standard dot. Its main difference is the color strategy: Gaussian-dot applies Gaussian color falloff at each joint, while the standard dot only draws the outline. The outline-only dot better matches the human pose skeleton, and when using pose encoder weights to initialize the object motion encoder, it enables faster training convergence and better adaptation to human motion dynamics.

4.3.2 Analysis of Context Fusion. We evaluate the impact of appearance and motion context in our network via ablation. For object appearance context, we compare three variants: (1) without token

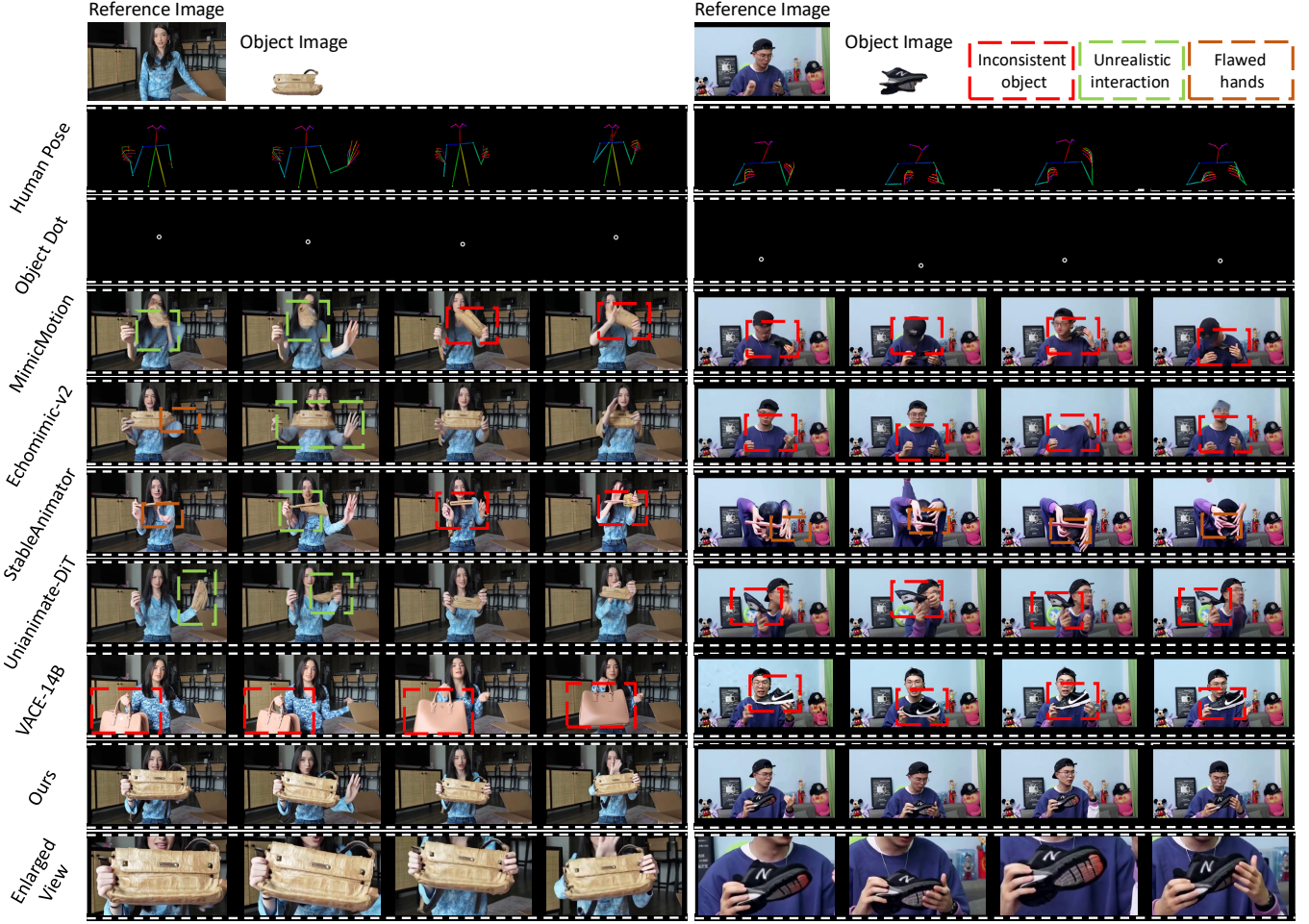


Fig. 5. Comparison with SOTAs on self-collected test set. More results can be found in the supplementary materials.

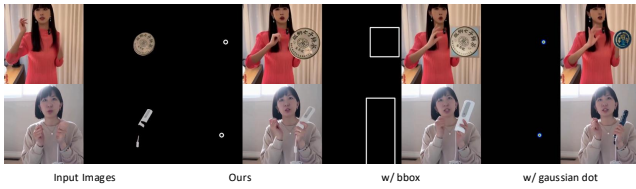


Fig. 6. Ablation results of different object motion.

concat, (2) without channel concat, and (3) with fixed-location channel concat (w/ fix-copy). Results (Table 2, Fig. 7) show that all variants reduce appearance fidelity. Removing the token concat has the most significant effect, highlighting its importance for conveying fine-grained textures. Retaining token concatenation while removing channel concatenation preserves partial appearance, though alignment issues with pose motion lead to degraded coherence. Fixing channel concat to a static location is less effective than our

alignment strategy, which avoids copy-paste artifacts and ensures semantically meaningful guidance.

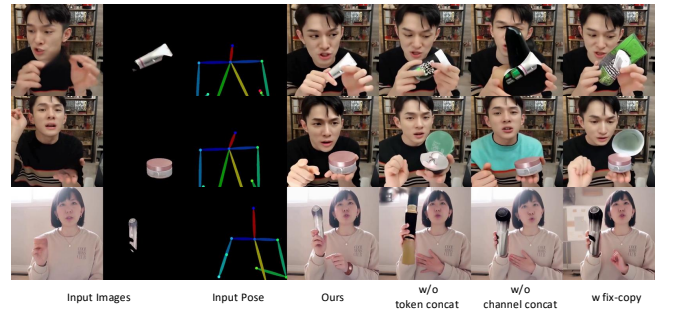


Fig. 7. Ablation results of context fusion.

For motion context fusion, we evaluate an alternative encoding strategy (Fig. 8, Table 2) in which human pose and object motion are

fused into a single composite motion map and processed by a shared motion encoder. This naive fusion leads to a noticeable drop in object appearance consistency. We attribute this to motion entanglement: merging heterogeneous motion sources without modality separation introduces semantic interference, degrading the model’s ability to encode and preserve object-related motion effectively.

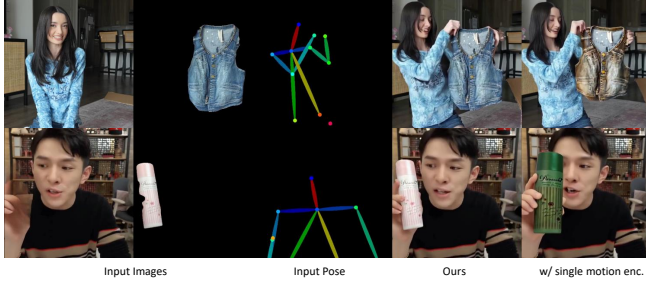


Fig. 8. Ablation results of motion encoder. The qualitative results indicate that employing the separate motion encoder enables more effective disentanglement between the appearances of the human and the object.

4.3.3 Analysis of HOI Adapter. We assess the effectiveness of our proposed HOI Adapter by comparing two variants: (1) removing the adapter entirely, and (2) replacing its self-attention with a cross-attention layer akin to the audio adapter (Sec.3.2.4). As shown in Table 2 and Fig. 9, removing the adapter leads to significantly slower convergence and failure to follow the motion dot, often resulting in the object appearing on the incorrect hand. The cross-attention variant also suffers from poor appearance preservation and inaccurate position injection, likely due to distribution mismatch. In contrast, reusing the pretrained MMDiT self-attention weights aligns more closely with the model’s internal priors, enabling more accurate and efficient HOI injection.

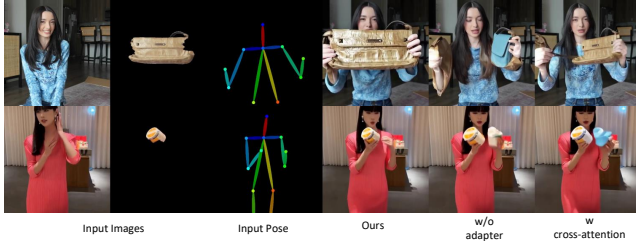


Fig. 9. Ablation results of HOI adapter. Without HOI adapter, the capability to track object motion is not well disentangled; instead, it is inherently coupled with the human pose.

4.3.4 Analysis of Audio Adapter. To verify the effectiveness of the audio injection module, we explore the location of the injection layer of the audio cross-attention adapter. We modified it to even-numbered layers for 512-resolution ablation as in Table 2. Due to the conflicts with the HOI-adapter in the same injection layers, the even-numbered audio cross-attention module performs worse.



Fig. 10. Comparisons with different types of poses.

Table 3. Quantitative results of our method compared with SOTAs.

Condition types	Self-collected test set							
	FID↓	FVD↓	OC↑	SC↑	BC↑	HAS↑	HF↑	MS↑
Full pose	51.60	502.69	90.05	95.19	94.45	87.93	98.76	99.14
Full torso w/o fingers	51.36	530.22	89.74	94.75	93.70	87.24	98.23	99.18
Arms w/ fingers	53.61	529.63	89.95	94.98	93.95	87.42	99.21	99.30
Arms w/o fingers	54.58	553.53	89.42	95.08	94.17	86.83	98.37	98.80

4.3.5 Comparison between Strong and Weak Conditions. We show the results for various conditions in Fig. 10 and Table 3. Compared with strong conditions, weak conditions show comparable performance while offering greater ease-of-use for users, such as easier acquisition and editing.

4.4 User Study

We conduct a user study to evaluate the quality of generated videos using five subjective metrics: Human consistency, object consistency, motion consistency, interaction naturalness, and video quality. The two test sets are applied in the user study. For each test case, we present the results from all methods and ask 15 users to score them from one to five. The final scores are then computed by averaging the scores given by the users. As shown in Table 4, our method consistently outperformed others across all criteria.

Table 4. User study scores.

Method	AnchorCrafter test set / Self-collected test set				
	Human Consistency ↑	Object Consistency ↑	Motion Consistency ↑	Interaction Naturalness ↑	Video Quality ↑
AnchorCrafter	3.56/-	3.46/-	2.79/-	3.38/-	3.27/-
MimicMotion	1.43/1.31	1.33/1.26	1.37/1.32	1.34/1.29	1.81/1.47
StableAnimator	1.55/1.78	1.24/1.18	1.17/1.16	1.31/1.15	1.76/1.38
Echomimic-v2	2.16/1.60	1.25/1.23	1.25/1.33	1.28/1.24	1.87/1.48
UniAnimate-DiT	3.05/2.40	2.64/2.05	2.09/2.14	2.32/2.14	2.64/2.20
VACE-14B	3.23/2.92	2.95/2.77	2.55/3.09	2.92/3.48	2.92/3.37
HOMA (ours)	3.90/3.52	3.69/3.66	3.15/3.84	3.61/4.00	3.59/3.78

4.5 More Results

We provide more results to showcase the practical applications and generalization capability of our method. To avoid copyright issues, we use the generated images as reference human input.

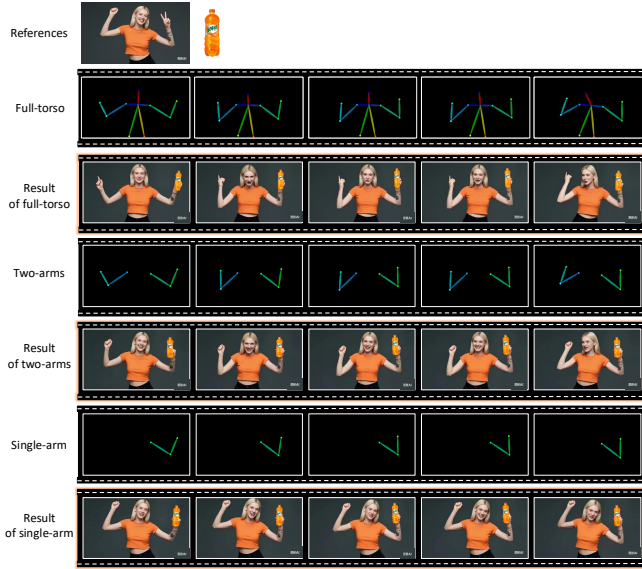


Fig. 11. Results with different types of weak poses.

Results with Different Weak Pose Types. We demonstrate the controllability of our method across a variety of weak human pose scenarios, as illustrated in Fig. 11. The results show that our framework effectively adapts to diverse poses while maintaining coherent object interactions and consistent appearance.

Influence of Text Prompt. We explore the impact of combining weak pose guidance with text-based control, as illustrated in Fig. 12. The results demonstrate that our method can effectively integrate minimal pose information with textual semantics to generate coherent and semantically aligned motions.

Results of 3D spatial Rotation of Objects. Our dot-based object trajectory condition cannot directly control object rotation; however, due to the model’s inherent capability, we also observed some rotation effects, as shown in Fig. 13.

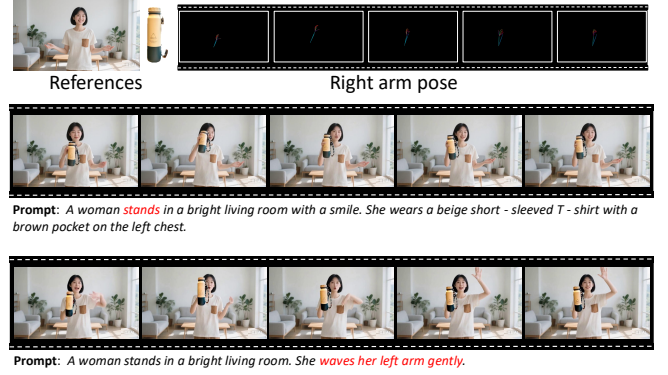


Fig. 12. Influence of Text Prompt.



Fig. 13. Examples of 3D object rotations.



Fig. 14. Results of Non-Rigid Objects.

Variations of Object Appearance and Motion. We present additional experimental results to demonstrate the versatility of our method. The results of various object appearances and motions are shown in Fig. 14 and Fig. 15, including flexible objects. The results highlight our model’s ability to adapt to varying object appearances while maintaining consistent identity and motion coherence.



Fig. 15. Results of various object appearance and motion.

4.6 Interactive Demo

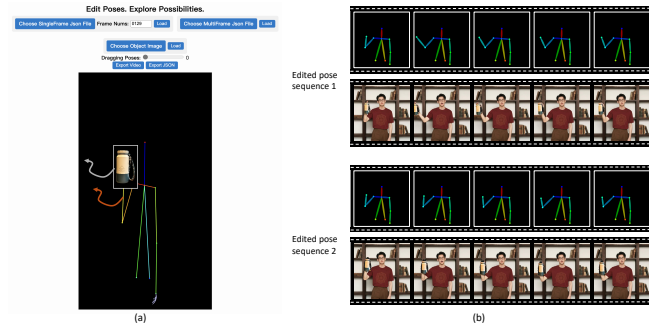


Fig. 16. Interactive Demo. Our interactive demo supports uploading either a single frame or a sequence of human poses, along with an object image. It allows users to adjust the object size on the first frame by resizing the bounding box. Users can select specific frame indices to edit, and perform drag-and-drop editing of both the human pose and the object position on each selected frame. Additionally, the demo supports interpolation between edited frames to generate smooth motion sequences.

We provide an interactive web demo that allows users to intuitively define human-object interactions by manually dragging the human pose and object position. This interactive tool grants the model adaptive freedom to resolve ambiguities, bridging the gap between artistic intent and generative feasibility. The web UI and use examples are illustrated in Fig. 16.

5 Conclusion

In this work, we propose HOMA, a novel framework for human-object interaction (HOI) video generation under weak control conditions. We explore a weak HOI condition setting to alleviate the reliance on hard-to-obtain strong HOI motion inputs. To this end, we propose a context fusion strategy and the HOI Adapter leveraging the generative prior, enabling the generation of HOI videos with high generalization and natural human-object interactions under limited input conditions. Extensive quantitative benchmarks and qualitative analyses demonstrate the superior performance of HOMA compared to existing diffusion-based approaches in human-centric video generation and editing. Furthermore, we validate the generalizability and applicability of our framework across a variety of real-world scenarios, including various object appearance and motion, different types of poses, and a user-friendly web demo.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under No. 62572458, and the Beijing Science and Technology Plan Project under No. Z231100005923033.

References

- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*. 9650–9660.
- Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. 2023. MagicPose: Realistic

- Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion. In *Forty-first International Conference on Machine Learning (ICML)*.
- Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. 2025. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 39. 2403–2410.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 24185–24198.
- Joon Son Chung and Andrew Zisserman. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II 13*. Springer, 251–263.
- Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. 2024. Vlogger: Multimodal diffusion for embodied avatar synthesis. *arXiv preprint arXiv:2403.08764* (2024).
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Yu Wu, and Jingdong Wang. 2025. Re-HOLD: Video Hand Object Interaction Reenactment via adaptive Layout-instructed Diffusion Model. *arXiv preprint arXiv:2503.16942* (2025).
- Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. 2023. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. In *Computer Graphics Forum (CGF)*, Vol. 42. Wiley Online Library, 1–12.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems (NeurIPS)* 30 (2017).
- Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8153–8163.
- Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. 2025. Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance. *arXiv preprint arXiv:2502.06145* (2025).
- Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. 2024. Make-Your-Anchor: A Diffusion-based 2D Avatar Generation Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6997–7006.
- Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. 2024. Sonic: Shifting Focus to Global Audio Perception in Portrait Animation. *arXiv preprint arXiv:2411.16331* (2024).
- Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. 2024a. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
- Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. 2024b. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1737–1747.
- Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598* (2025).
- Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).
- Jiaman Li, Jiajun Wu, and C Karen Liu. 2023. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–11.
- Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. 2024. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876* (2024).
- Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. 2025. OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human Animation Models. *arXiv preprint arXiv:2502.01061* (2025).
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems (NeurIPS)* 36 (2023), 34892–34916.
- Shilong Liu, Zhaoqiang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 38–55.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy* 2015 (2015), 18–24.
- Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. 2024. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160* (2024).
- Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. 2024. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation. *arXiv preprint arXiv:2411.10061* (2024).
- Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, and Yebin Liu. 2024. ManiVideo: Generating Hand-Object Manipulation Video with Dexterous and Generalizable Grasping. *arXiv preprint arXiv:2412.16212* (2024).
- Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2023. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553* (2023).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*. PMLR, 8748–8763.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitam Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).
- Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 2024. Human4DiT: 360-degree Human Video Generation with 4D Diffusion Transformer. *ACM Transactions on Graphics (TOG)* 43, 6 (2024).
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Reformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2024. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision (ECCV)*. Springer, 244–260.
- Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. 2024. StableAnimator: High-Quality Identity-Preserving Human Image Animation. *arXiv preprint arXiv:2411.17697* (2024).
- Thomas Unterthiner, Soerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
- Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. 2025a. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
- Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2024b. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9326–9336.
- Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. 2024d. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188* (2024).
- Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2025b. UniAnimate-DiT: Human Image Animation with Large-Scale Video Diffusion Transformer. *arXiv preprint arXiv:2504.11289* (2025).
- Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. 2024c. Humanvid: Demystifying training data for camera-controllable human image animation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. 2023. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*. 14928–14940.
- Ziyi Xu, Ziyao Huang, Juan Cao, Yong Zhang, Xiaodong Cun, Qing Shuai, Yuchen Wang, Linchao Bao, Jintao Li, and Fan Tang. 2024a. AnchorCrafter: Animate CyberAnchors Saling Your Products via Human-Object Interacting Video Generation. *arXiv preprint arXiv:2411.17383* (2024).
- Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024b. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1481–1490.

- Zihui Xue, Mi Luo, Changan Chen, and Kristen Grauman. 2024. HOI-Swap: Swapping Objects in Videos with Hand-Object Interaction Awareness. *arXiv preprint arXiv:2406.07754* (2024).
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), 21875–21911.
- Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCVW)*. 4210–4220.
- Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8652–8661.
- Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2024. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680* (2024).
- Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. 2024. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458* (2024).
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).
- Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. In *European Conference on Computer Vision (ECCV)*.