Making Avatars Interact: Towards Text-Driven Human-Object Interaction for Controllable Talking Avatars

Youliang Zhang[1,2†], Zhengguang Zhou[2†], Zhentao Yu[2†], Ziyao Huang[2], Teng Hu[3,2], Sen Liang[4,2], Guozhen Zhang[5,2], Ziqiao Peng[6,2], Shunkai Li[2], Yi Chen[2], Zixiang Zhou[2], Yuan Zhou[2], Qinglin Lu[2⋆], and Xiu Li[1⋆]

[1] Tsinghua University
[2] Tencent Hunyuan
[3] Shanghai Jiao Tong University
[4] University of Science and Technology of China
[5] Nanjing University
[6] Renmin University of China

Fig. 1: Input text, audio, and reference image, our method can generate human videos that can talk, act, and interact with objects.

Abstract. Generating talking avatars is a fundamental task in video generation. Although existing methods can generate full-body talking avatars with simple human motion, extending this task to grounded human-object interaction (GHOI) remains an open challenge, requiring the avatar to perform text-aligned interactions with surrounding objects. This challenge stems from the need for environmental perception and the control-quality dilemma in GHOI generation. To address this, we propose a novel dual-stream framework, **InteractAvatar**, which decouples perception and planning from video synthesis for grounded human-object interaction. Leveraging detection to enhance environmental perception, we introduce a Perception and Interaction Module (PIM) to generate text-aligned interaction motions. Additionally, an Audio-Interaction Aware

† Equal contribution. ⋆ Corresponding author.

Generation Module (AIM) is proposed to synthesize vivid talking avatars performing object interactions. With a specially designed motion-to-video aligner, PIM and AIM share a similar network structure and enable parallel co-generation of motions and plausible videos, effectively mitigating the control-quality dilemma. Finally, we establish a benchmark, **GroundInter**, for evaluating GHOI video generation. Extensive experiments and comparisons demonstrate the effectiveness of our method in generating grounded human-object interactions for talking avatars.

1 Introduction

Human-centric video generation has achieved remarkable progress, with significant strides made in tasks like photorealistic talking head synthesis, focusing primarily on accurate lip-synchronization and facial expression generation [4, 27, 32, 35, 40]. However, to construct vivid and practical digital humans, we must move beyond facial modeling and endow them with the ability to perform complex actions and interact meaningfully with their surroundings.

We introduce **Grounded Human-Object Interaction** for talking avatars, aiming to enhance current digital humans with environmental perception and text-driven human-object interaction generation ability. Different from the current digital humans, GHOI features: 1) Environment-aware, perceives its environment and performs reasonable actions within it. 2) Initial frame consistency, maintaining the integrity of the initial scene frame; 3) Textual Guidance, facilitating high-level control without requiring granular pose or trajectory specifications; 4) Grounded Interaction. While grounding involves locating target objects via text (similar to Grounding DINO [25]), GHOI extends this by requiring the generation of plausible human-object interactions rooted in this spatial context.

Despite recent advancements, existing paradigms still fall short of enabling such nuanced, scene-aware interaction due to fundamental architectural limitations: ***(1)*** **Audio-driven methods**, through several pioneering works on simple full-body animation [4, 9, 22, 36], typically learn a direct mapping from acoustic features to pixel space, lacking explicit modeling of objects and environments, thus making complex human-object interactions difficult to control; ***(2)*** **Pose-driven approaches** [6, 31, 33, 38] offer explicit control but offload the planning burden to the user. These methods require pre-defined skeletal sequences as input, which are not only difficult and expensive to obtain but also often misaligned with the specific context of a reference image; ***(3)*** **Subject-consistent methods** [3, 15, 24] excel in preserving subject identity and achieving coherent integration, but lack mechanisms for grounded interaction. They are designed to synthesize new videos from text rather than to carry out contextual commands in an interactive environment. Therefore, enabling a speaking avatar to perform stable, text-driven object interactions remains a significant and unresolved challenge, especially in multi-object scenes. The core difficulties are twofold.

(1) **Scene-Action Grounding**: Instead of merely generating a video featuring human-object interaction, GHOI demands the interaction to occur within

a specific environment involving designated objects. This requires the model to interpret the spatial layout and semantic content, and to ground textual commands within this visual context. ***(2)*** **Control-Quality Dilemma**: Owing to the complexity of GHOI, existing methods often face a trade-off between controllability and visual quality when controlled solely on sparse textual input (audio just for sync). They either generate high-fidelity videos that struggle to follow instructions or achieve reasonable responses at the expense of video fidelity.

To address these challenges, we propose InteractAvatar, a novel dual-stream Diffusion Transformer (DiT) framework designed to generate grounded human-object interactions in talking avatars. This framework explicitly decouples perception and planning from video synthesis. For clarity, we define **motion** as the rendered skeletal poses and object boxes (RGB format). We introduce a Perception and Interaction Module (PIM) to tackle the **Scene-Action Grounding Challenge**, which grounds the right object and generates text-aligned motion sequences based on the ref image. The PIM is jointly trained on detection and motion generation tasks, ensuring the production of scene-aware and text-aligned motions. Conditioned on both audio and the motion latent features, an Audio-Interaction Aware Generation Module (AIM) is proposed to synthesize the final video. PIM and AIM share a similar network architecture and achieve the parallel co-generation of motions and videos. With a specially designed motion-to-video (M2V) aligner, our symmetrical yet decoupled framework effectively alleviates the **control-quality dilemma** in complex GHOI video generation, achieving vivid and plausible grounded human-object interactions for talking avatars.

To rigorously evaluate our approach, particularly its ability to model grounded human-object interactions, we constructed the GroundInter benchmark. GroundInter comprises 600 distinct test cases, each containing a reference image, a structured textual description of the interaction, and the corresponding speech audio. Extensive experiments conducted on this benchmark demonstrate that our InteractAvatar significantly outperforms existing approaches in generating plausible, controllable, and high-quality grounded human-object interaction videos.

Our main contributions are summarized as follows:

- We propose **InteractAvatar**, a novel dual-stream DiT framework that explicitly decouples perception planning from video synthesis, achieving text-driven grounded human-object interaction generation for talking avatars.
- We introduce a **Perception and Interaction Module** (PIM), trained on both detection and motion generation tasks, to generate scene-aware and text-aligned motions. In addition, we propose an **Audio-Interaction Aware Generation Module** (AIM) with an M2V aligner to produce videos with controllability and high visual quality.
- Beyond the core GHOI video generation, our framework provides **unified and flexible multimodal control**, accepting any combination of text, audio, and motion as input.
- We establish **GroundInter**, a benchmark for fine-grained evaluation of GHOI. Experiments demonstrate that our method significantly outperforms

existing approaches in producing controllable high-quality talking avatars with grounded human-object interactions.

2 Related Work

2.1 Audio-driven human animation.

Given audio and human reference images as input, audio-driven human animation aims to generate videos with well-aligned lip movements. Initial research predominantly focused on head-and-shoulder animations, SadTalker [46] employing 3D motion coefficients to improve realism, Hallo [40] using a hierarchical diffusion approach for lip-sync accuracy, EchoMimic [5] combining audio with facial landmarks to enhance stability, and Loopy [17] leveraging temporal modules to generate natural motion without manual templates. Subsequently, research has progressively extended this capability toward half-body and full-body digital human synthesis with body action modeling. Hallo3 [9] first applied a pre-trained DiT model for portrait animation, HY-Video-Avatar [4] achieves dynamic, emotion-controllable, and multi-character dialogue video generation. FantasyTalking [36] proposes a dual-level optimization for global motion and lip synchronization, and OmniHuman [22] scales up training data with hybrid motion conditions to generate more realistic full-body videos. However, these methods lack the capacity to generate more complex GHOI videos. Our work extends audio-driven animation to incorporate human-object interaction, thereby enhancing the realism and practical applicability of the generated avatars.

2.2 Human-Object Interaction Video Generation

Recently, Human-Object Interaction (HOI) video generation has garnered increasing attention. Some works focus on editing existing footage. MIMO [26] and AnimateAnyone2 [14] substitute humans within dynamic scenes, whereas HOI-Swap [42] and ReHoLD [11] replace handheld objects. However, a primary limitation of these methods is their inability to animate a static reference image from scratch. Another works seek fine-grained control via complex conditional inputs. AnchorCrafter [41] utilizes pose and depth maps for interaction-aware human motion, ManiVideo [28] employs precise 3D models of both the human and object for accurate control, and HunyuanVideo-HOMA [16] explores weak HOI conditions with a novel appearance injection method. Nevertheless, acquiring and aligning these complex conditions with an arbitrary ref image is often impractical, hindering their use in digital human applications. In contrast, our method overcomes these challenges by generating realistic human-object interaction videos conditioned solely on text, thereby eliminating the need for source videos or other complex control signals.

Recently, subject-consistent video generation also exhibited the capacity to synthesize HOI video. Phantom [24] leverages the inherent consistency of DiT-based models by concatenating reference image latents with video latents. Concurrent works such as InterActHuman [39] and HunyuanCustom [15] integrate

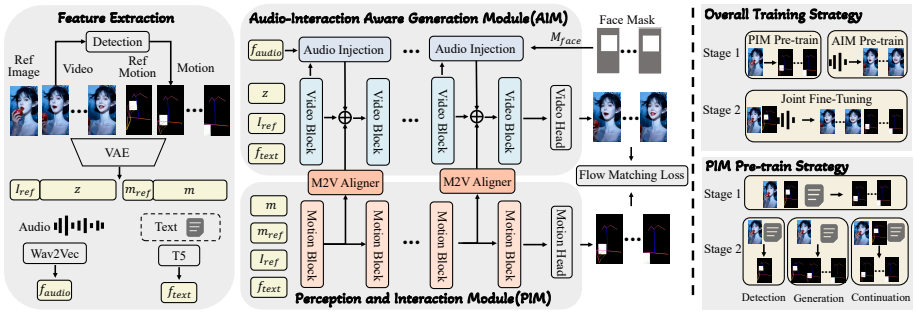

Fig. 2: Overview of our dual-stream InteractAvatar framework and multi-modal conditioned training strategy.

subject-consistent video generation with audio-driven animation, while HuMo [3] proposes a progressive multimodal training paradigm to enable flexible control. However, a critical distinction between GHOI and subject-consistent video generation is that these methods do not preserve the world defined by the scene of the ref image. Instead, they typically generate a new scene based on text and subject appearance, rather than situating actions within the provided scene.

3 Method

As shown in Fig. 2, we introduce a dual-stream Diffusion Transformer (DiT) for generating text-controlled grounded human-object interactions. Our architecture decomposes this complex task into two sub-problems, perceptual planning and video rendering, handled by the Perception and Interaction Module (PIM) and the Audio-Interaction aware Generation Module (AIM), respectively. We define **motion** as the rendered human skeletal poses and object boxes. A key aspect of our design is the parallel co-generation of structural motion sequences and corresponding video frames. The PIM acts as the **planning brain**, focusing on high-level structure by generating scene-aware and text-aligned motions. The AIM serves as the **rendering engine**, synthesizing high-fidelity videos with precise audio lip-sync. Rather than a sequential pipeline, AIM is informed throughout the generation process via the M2V aligner, which employs a layer-wise residual injection mechanism to ensure that the generated video consistently adheres to the evolving structural motion from PIM in lockstep.

Both modules are trained under a unified Flow Matching paradigm. A pre-trained Variational Autoencoder (VAE) [34] encodes all visual inputs (videos, motions, ref images) into latent space, while a T5 [30] encoder embeds the textual commands. The model u_θ is optimized to predict a vector field v_t with the following objective function:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_0,c}\left[\|v_t(z_t) - u_\theta(z_t, t, c)\|_2^2\right], \tag{1}$$

where z_0 is the clean latent from VAE, z_t is its noised version at timestep t, $v_t(z_t)$ is the ground-truth vector field, and c represents all conditioning information.

3.1 Perception and Interaction Module

Works like VideoJAM [2] show the benefits of motion co-generation in video synthesis, we further explore generating motion in GHOI and avatar tasks. The PIM is responsible for parsing the environmental context from a reference image I_{ref} and generating a spatio-temporally reasonable interaction motion with the guidance of the text prompt T. We define the interaction motion m as a composite RGB representation, comprising a sequence of human skeletal keypoints, $P = \{p_t\}_{t=0}^{N-1}$, and object bounding box trajectories, $O = \{o_t\}_{t=0}^{N-1}$, where N is the total number of frames. This motion serves as an intermediate representation that is injected at the feature level into the video generation module AIM, thereby enabling disentangled control of the human-object interaction.

Interaction Motion Generation. To precisely control the generation process, we explicitly partition the motion generation into two distinct task types: pure action generation, involving only the human pose sequence P, and human-object interaction generation, including both the pose sequence P and object trajectories O. We inject the task-specific information by concatenating a task embedding vector f_{task} corresponding to the task type $c_{task} \in \{\text{ACTION}, \text{HOI}\}$ with the text prompt embedding f_{text}. This creates a unified conditioning vector for the cross-attention layers of the model:

$$f_{text} = \text{Concat}(f_{text}, f_{task}). \tag{2}$$

The reference image I_{ref} is temporally prepended to the first motion frame m_{ref}, the model takes m_{ref} as first frame to generates overall motion sequence $\{m_t\}_{t=1}^{N-1}$ conditioned on I_{ref}. To ensure the reference image provides effective global guidance without conflicting with the positional encodings of the motion sequence, we employ a custom mapping for its Rotary Position Embeddings (RoPE). For a given patch z_I from the reference image with an original 3D position index (l, i, j), its position is remapped as:

$$\text{RoPE}_{z_I}(l, i, j) = \text{RoPE}(-1, i + w, j + h). \tag{3}$$

This strategy positions the reference image as a special environment frame at a virtual timestep of -1, allowing it to condition the entire sequence generation without disrupting the temporal integrity of the video frames themselves. Given that motions encode structure rather than fine-grained texture, we therefore constrain the shorter side of the motions to 256 pixels, preserving essential structural information while significantly reducing computational overhead.

Environment Perception Training. Naive conditioning on the reference image is insufficient for the model to understand the complex scene layouts and object geometries. To address this, we develop an environment perception training curriculum that encourages the model to engage more deeply with the reference image. Our training curriculum dynamically alternates between two task modes: **(1) Conditional Continuation.** In this mode, the model receives the reference image I_{ref}, the first frame motion m_{ref}, and the text prompt T as conditions. Its objective is to generate the subsequent motion sequence

$\{m_t\}_{t=1}^{N-1}$. **(2) Perception as Generation.** In this mode, the initial motion m_{ref} is masked. The model must predict the entire motion sequence $\{m_t\}_{t=0}^{N-1}$ from scratch based on ref image I_{ref} and text T. To strictly enforce grounding, we randomly truncate the target length to 1. This converts the generation task into a pure scene-text grounding problem, ensuring the model learns to locate the target objects defined in T within I_{ref} without relying on any motion priors.

This detection-like task is also optimized using the same flow matching objective, with the loss computation simply confined to the first frame. This unified framework seamlessly trains the PIM to leverage a single set of network parameters for two seemingly disparate yet fundamentally related tasks: parsing scene structure and generating temporal motions from perception. This strategy obviates the need for heterogeneous loss functions and their associated hyperparameters, which may introduce training instabilities.

3.2 Audio-Interaction Aware Video Generation

Audio Injection. Precise lip synchronization with the input audio is paramount for generating realistic avatars. We employ a pre-trained Wav2Vec model [1] as our audio feature extractor. To capture the co-articulation and temporal dynamics of speech, rather than relying on isolated phonemes, we extract features from a contextual window around each timestep. For the i-th video frame, its corresponding audio feature $f_{\text{audio},i}$ is computed by encoding and aggregation:

$$f_{\text{audio},i} = \text{Aggregate}\left(\text{Wav2Vec}(\mathbf{a}_{[c_i-w:c_i+w]})\right), \tag{4}$$

where $\mathbf{a}$ is the raw audio, c_i is the center sample aligned with frame i, w denotes the half-width of the contextual window, and Aggregate$(\cdot)$ is a feature aggregation operation, such as concatenation, for temporal video audio align.

To achieve frame-level audio-visual alignment, the extracted audio feature sequence $\{f_{\text{audio},i}\}$ is injected into the AIM via cross-attention. To accelerate the learning of audio-driven capabilities in early AIM training, we introduce a spatial face mask $\mathbf{M}_{\text{face}}$, which is not employed during later training stages or inference. This mask spatially weights the output of the audio injection module, concentrating its influence on the facial region for stable training:

$$\mathbf{h}'_{v,l} = \mathbf{h}_{v,l} + \text{CrossAttention}(\mathbf{h}_{v,l}, f_{\text{audio}}) \odot \mathbf{M}_{\text{face}}, \tag{5}$$

where $\mathbf{h}_{v,l}$ represents the visual features at layer l of the AIM and $\odot$ denotes element-wise multiplication. This strategy effectively channels the impact of the audio signal to the target facial area, leading to a more stable learning process.

Motion to Video Aligner. Leveraging the isomorphic architecture of PIM and AIM, we introduce the M2V aligner with a layer-wise residual injection for motion control. Specifically, we inject the feature residual, the difference between the output of each DiT block in the PIM, into the corresponding layer of the AIM. For given layer $l > 0$ in the PIM, its residual $\Delta\mathbf{h}_{m,l}$ is computed as:

$$\Delta\mathbf{h}_{m,l} = \mathbf{h}^{\text{out}}_{m,l} - \mathbf{h}^{\text{out}}_{m,l-1}. \tag{6}$$

For the base layer ($l = 0$), we use $\mathbf{h}^{\text{out}}_{m,0}$ directly.

To reconcile the discrepancy of the low-resolution motion and the high-resolution video, we first upsample the residual feature $\Delta\mathbf{h}_{m,l}$ to match the spatial dimensions of video features $\mathbf{h}_{v,l}$ using simple bilinear interpolation. This is followed by a projection through a zero-initialized linear layer. The final output of an AIM block is computed as:

$$\mathbf{h}^{\text{out}}_{v,l} = \text{Block}_v(\mathbf{h}^{\text{in}}_{v,l} + \text{Linear}_{\text{zero}}(\text{Interp}(\Delta\mathbf{h}_{m,l}))). \tag{7}$$

This residual injection strategy significantly reduces loss fluctuations during training. Furthermore, the zero-initialization of the linear layer ensures stability in the early stages of training by allowing the model to gradually learn the guidance of the motion. This effectively prevents visual artifacts, such as the ghosting of the motion's skeletal lines, from leaking into the generated video.

Unified Training for Generative and Driven Control. A truly versatile digital human system should not only generate motions from textual prompts but also faithfully follow explicit driving signals, such as skeletal sequences, interaction motions, or object trajectories. To achieve this, we co-train the model on two complementary tasks: Joint Generation and External Driving. For External Driving, the model is provided with a clean driving signal, such as a skeleton sequence, which is fed directly to the PIM. Here, the PIM actually acts as a feature encoder, which is elegantly achieved by setting the diffusion timestep t to 0. Driven signals are also extracted as layer-wise feature residuals $\{\Delta\mathbf{h}_{m,l}\}$, which are then injected into the AIM to guide the generation process. To balance these dual capabilities, we employ a 4:1 training data ratio between the Joint Generation and External Driving tasks. This co-training strategy endows our model to function both as a creative, text-guided motion video generator and as a precise, controllable animator, all within a single, unified framework.

3.3 Multimodal Conditioned Training Strategy

A three-stage curriculum was designed to progressively integrate multimodal conditions, ensuring training stability and mitigating inter-modal conflicts.

PIM Pre-training. The initial stage is to train the PIM independently. The goal is to establish its core ability to perceive scenes from the reference image and generate coherent motion plans based on text. The training curriculum blends multiple tasks: (1) a object detection task to build scene perception, (2) a continuation task, motion continuation from a given first motion m_{ref}, and (3) a crucial hybrid generation task that requires generating the motion sequence from only the image I_{ref} and text. We found that this mixed-task approach, which forces a synergy between perception and generation, is superior to the pure generation pipeline. The training strategy of those tasks is shown in Fig. 2.

AIM Pre-training. The second stage imbues AIM with the ability for audio-driven lip synchronization while preserving its underlying image-to-video (I2V) quality. Crucially, we found that the training order is paramount. The model should be trained with audio before being exposed to interaction motion. We attribute this to the distinct nature of condition signals: audio is a soft,

localized, and heterogeneous modality injected through cross-attention, whereas motion serves as a hard, global, and homogeneous structural constraint introduced via residual injection. Introducing the dominant structural motion condition only after the weaker local audio conditioning has stabilized effectively prevents the audio signal from being suppressed during training optimization.

End-to-End Joint Fine-tuning. In the final stage, the pre-trained PIM and AIM are jointly fine-tuned, where the PIM's outputs are injected into the AIM as layer-wise residuals. The training strategy adopts a probabilistic mix of tasks and conditions to unify the model's capability to accommodate any combination of text, audio, and motion inputs. Specifically, 30% of the samples are conditioned on audio to preserve lip sync ability, 15% on ground-truth motions for driven ability, and 60% on joint video-motion generation. Retaining a considerable portion of pure image-to-video samples serves as a critical regularizer, improving identity preservation, video dynamics, and generalization.

4 Experimental Results and Analysis

4.1 Experimental Details

Both our PIM and AIM are initialized from the wan2.2-5B [34]. The PIM is trained for 30,000 steps on 3-10 second clips of skeleton sequences and object trajectories (short side 256px), while the AIM is trained for 5,000 steps on 3-6 second video clips (short side 704px). In the initial pre-training phases for PIM and AIM, the learning rate is fixed at 1e-5. This is followed by a joint training phase of 4,000 steps where the learning rate is lowered to 2e-6. Same VAE is used for both video and motion, the downsample rate of motion is 2.75.

4.2 Datasets and Benchmark Construction

Our model is trained on three open-source datasets: SpeakerVid-5M [47], OpenHumanVid [20], and HOIGen-1M [23]. We utilize Gemini-2.5 Pro [8] for detailed human motion and object interaction caption, DWPose [43] for human skeletons and DINO [45] for object detection. To facilitate a rigorous evaluation for human object interaction generation, we constructed a benchmark with images generated by jimeng4.0, named GroundInter. GroundInter benchmark includes 400 images depicting actual or potential human-object interactions with 1-3 objects, with 100 different common objects. Each image is richly annotated with object types, 1-3 corresponding action descriptions (detailing interaction type, objects, and temporal segments), and matching dialogue scripts synthesized into audio using CosyVoice [10], resulting in 600 test cases. Additionally, we provide ground-truth object segmentation masks [19], detection results [45], and DWPose keypoints [43]. We introduce a suite of metrics focused on interaction quality: (1) VLM-QA, we design 30 questions structured around three categories (object, human, interaction), and the VLM [37] assigns a score of 1 or 0 to each based on the provided reference image and video. (2) Semantic Consistency

(CLIP_{re}), the CLIP-score [29] between the original text prompt and a VLM-generated caption of the generated video; (3) Hand Quality (HQ), assessed by the product of hand dynamics score and hand Laplacian sharpness score [21]; (4) Object Quality (OQ), assessed by the product of object dynamics score [47] and object DINO consistency [45]; and (5) Pixel-Level Interaction (PI), which verifies contact between DINO object boxes and DW-Pose keypoints. Our evaluation is further complemented by standard metrics. Ref Consistency with DINO for human appearance ($\text{DINO}_{subject}$) and ref image appearance (DINO_{ref}). Sync confidence (Sync_{conf}) for audio-visual consistency [7], CLIP-B-T (CLIP_{text}) for text-video alignment, temporal consistency (Temp-C) and dynamic degree (DD) from VBench [44]. Detailed protocols for training data processing and test benchmark construction are provided in the Supplementary Materials.

Table 1: Quantitative comparison on GroundInter. With different input signals, our single model can infer with various modes. TA2V means audio-driven, TAM2V means audio and motion-driven, T2MV represents the use of self-generated motion.

Task	Method	Human-Object Interaction					Ref Consistency		Text-Video Align		Audio and Video	
		VLM-QA ↑	HQ ↑	OQ ↑	DD ↑	PI ↑	$\text{DINO}_{subject}$ ↑	DINO_{ref} ↑	CLIP_{text} ↑	CLIP_{re} ↑	Temp-C ↑	Sync_{conf} ↑
Audio-Driven	Wan-S2V [13]	24.65	0.336	0.063	0.095	0.619	0.670	**0.870**	0.281	0.885	**0.955**	5.43
	HY-Video-Avatar [4]	26.88	0.745	0.112	0.321	0.666	0.661	0.841	0.285	0.873	0.934	5.98
	Fantasy-Talking [36]	25.35	0.416	0.098	0.127	0.593	**0.673**	0.855	0.276	0.889	0.941	5.79
	OminiAvatar [12]	26.76	0.732	0.118	0.305	0.732	**0.639**	0.847	0.282	0.885	0.940	5.95
	Ours (TA2V)	**27.32**	**0.931**	**0.133**	**0.765**	**0.803**	0.671	0.857	**0.286**	**0.899**	0.945	**6.04**
Pose-Driven	UniAnimate-DiT [38]	24.65	0.862	0.082	0.441	0.669	0.621	0.782	0.277	0.875	0.938	–
	Ours (TAM2V)	**28.47**	**0.957**	**0.141**	**0.773**	**0.832**	**0.648**	**0.842**	**0.287**	**0.902**	**0.940**	5.73
Text-Driven	VACE [18]	26.74	0.908	0.118	0.719	0.705	0.629	0.817	0.285	0.893	0.940	–
	HuMo [3]	24.12	0.910	0.101	0.380	0.491	0.566	0.726	0.281	0.880	0.935	5.15
	Ours (T2MV)	29.05	**0.975**	**0.150**	**0.792**	**0.852**	0.637	0.835	0.289	**0.904**	0.932	–
	Ours (TA2MV)	**29.07**	0.973	0.147	0.783	0.850	**0.641**	**0.839**	**0.290**	**0.904**	**0.942**	**5.92**

4.3 Comparison with the State-of-the-Art

As there are currently no methods specifically designed for Grounded Human-Object Interaction Video Generation, we compare our method against state-of-the-art (SOTA) approaches from several adjacent domains. For audio-driven methods, we select HY-Video-Avatar [4], Wan-S2V [13], OmniAvatar [12], and Fantasy-Talking [36] for comparison. For subject-consistent video generation, we compare against HuMo [3]. For pose-driven animation, we benchmark against UniAnimate-DiT [38]. VACE [18] is also utilized for comparison. As shown in Tab. 1, the **audio-driven methods** demonstrate an advantage in ref consistency metrics but perform poorly on human-object interaction, particularly on Hand Quality (HQ) and Object Quality (OQ) scores. This is attributable to their primary focus on facial dynamics, which only allows for the generation of simple, low-dynamic actions and largely avoids interactive behaviors, also confirmed by their low dynamic degree (DD) scores. While this low dynamism facilitates the preservation of background and subject consistency, the absence of meaningful action is clearly reflected in the poor interaction scores. In contrast, our TA2V variant achieves a substantial lead in interaction metrics, representing an improvement of approximately 180% in HQ, 111% in OQ compared to the strong baseline Wan-S2V. Crucially, this is achieved while maintaining

comparable reference preservation in terms of $\text{DINO}_{subject}$ and DINO_{ref}. On lip synchronization performance, our method even outperforms all audio-driven method in the Human-Object Interaction scenario. For **pose-driven methods**, UniAnimate-DiT takes motion sequences generated by our PIM as input. Even with this high-quality guidance, our TAM2V variant outperforms it across all metrics, achieving a particularly notable margin on the OQ score. This discrepancy arises because pose-driven methods do not account for object morphology, leading to interaction misalignments and unnatural object deformations. Their inherent limitations also include the difficulty of obtaining and aligning pose sequences with the reference image, as well as the limited speaking ability. Regarding the **subject-consistent method**, HuMo generates an entirely new video with subjects borrowed from the reference image, rather than continuing from the original scene, resulting in lower scores on reference consistency. Furthermore, our method significantly surpasses HuMo in terms of interaction quality. This is because interactions generated by subject-consistent models tend to be stereotypical and fail to produce smooth transitions from a non-interactive to an interactive state. Notably, all our variants share a single model with different inference modes, highlighting its versatility and practical utility.

4.4 Qualitative Evaluation

In Fig. 3, we present a qualitative comparison between our method and SOTA approaches from related domains for the Grounded Human-Object Interaction task. VACE exhibits responsiveness to text but suffers from severe artifacts during interaction, failing to preserve object integrity or accurately execute the interaction instructions. Moreover, it lacks the crucial audio-driven capability. Wan-S2V successfully preserves the interaction environment specified by the reference image. However, its dynamism is largely confined to the audio-driven facial region, limiting its ability to respond to broader interaction instructions. HuMo excels at generating fine-grained interaction details and facial dynamics. However, it fails to maintain the original interaction environment, instead using the subject from the reference image as an appearance template to synthesize an entirely new video, which is unsuitable for the GHOI task. In contrast, our method accurately follows textual instructions, enabling the human to interact stably and plausibly with objects within the reference image environment. This is facilitated by the explicit decoupling of perception and planning from video synthesis. The PIM plans plausible scene-aware interaction motions, while the AIM is responsible for synthesizing the photorealistic details of the interactions.

4.5 Ablation Studies

Effect of Environmental Perception Training. We conduct an ablation study to evaluate different configurations of our environmental perception training process, with the results summarized in Tab. 2. By default, the reference image is prepended to the video sequence. **Last** indicates that the reference image is appended after the last frame. **RoPE** specifies whether the modified

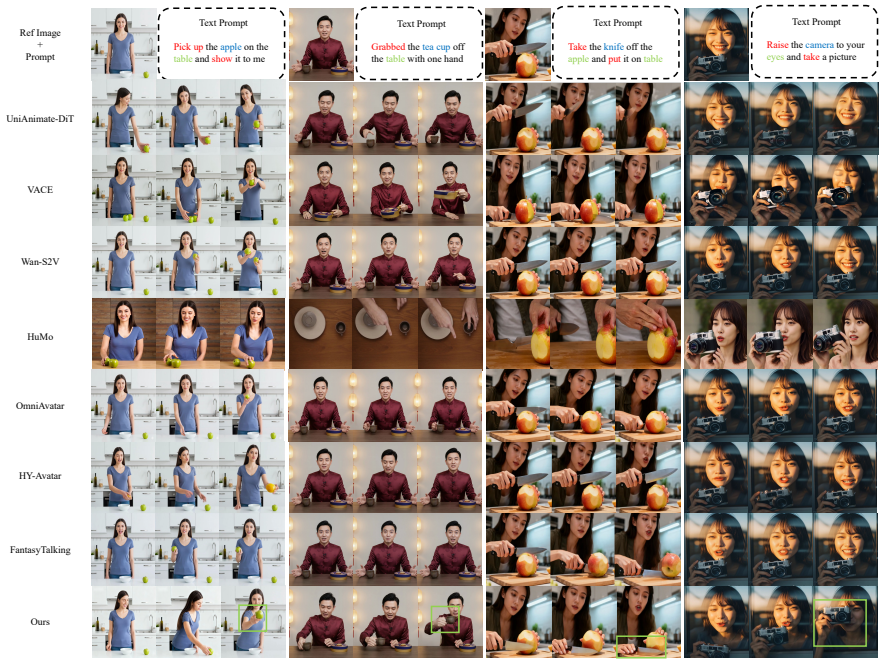

Fig. 3: Qualitative comparison with SOTA methods. Refer to the supplementary materials for the results of HY-Video-Avatar and Fantasy-Talking.

Rotary Positional Embeddings are applied to the reference image. **Det.** denotes the use of detection data during PIM pre-training, which involves generating a single motion frame based on the reference image. **Det.+Cont.** represents generating a full motion sequence without reference motion. Our findings indicate that prepending the reference image yields superior performance compared to appending it, due to the structural similarity between the reference image and motion frames. However, this advantage is contingent upon modifying RoPE to mitigate the excessive influence of the reference image on the generated frames. Furthermore, we observe that detection data is crucial for establishing environmental perception abilities, while Det.+Cont. data serves as a vital bridge between perception and generation. Both types of data benefit the PIM training.

Table 2: Effect of Environmental Perception Training.

RoPE	Det.	Det.+Cont.	Last	VLM-QA ↑	$CLIP_{re}$ ↑	HQ ↑	OQ ↑	PI ↑
				26.36	0.895	0.826	0.118	0.711
			✓	26.71	0.897	0.818	0.115	0.710
✓				27.15	0.902	0.845	0.123	0.730
✓	✓			27.87	0.902	0.917	0.132	0.751
✓	✓	✓		**28.89**	**0.904**	**0.925**	**0.144**	**0.780**

Effect of PIM. The homologous architecture of our PIM and AIM establishes intrinsic alignment in their feature representations, which we leverage for effective information fusion. As detailed in Tab. 3 (top), we conduct an ablation study on the injection of interaction motion features. w/o PIM refers to training a single AIM as a text-image-audio-driven video generator, which exhibits relatively poor text alignment and human-object interaction quality. This

Table 3: Ablation of PIM design and effect (top) and the Multimodal Training Strategy for unified and flexible multimodal control (bottom).

Method	VLM-QA ↑	HQ ↑	OQ ↑	DD ↑	PI ↓	$DINO_{subject}$ ↑	$DINO_{ref}$ ↑	$CLIP_{text}$ ↑	$CLIP_{re}$ ↑	Temp-C ↑	$Sync_{conf}$ ↑
w/o PIM	26.19	0.711	0.104	0.587	0.685	0.638	0.840	0.285	0.892	0.936	5.41
Cascade	28.19	0.876	0.137	0.730	0.746	0.642	0.841	0.289	0.901	0.940	5.27
Last-layer	28.31	0.880	0.141	0.752	0.769	0.643	**0.849**	0.289	0.901	0.938	5.32
Addition	28.46	0.903	0.134	0.751	0.777	0.630	0.846	0.286	0.887	0.935	5.14
Ours	**28.89**	**0.925**	**0.144**	**0.761**	**0.780**	**0.646**	0.846	**0.290**	**0.904**	**0.942**	**5.43**
AM	28.23	0.912	0.133	0.745	0.770	0.639	0.838	0.289	0.900	0.933	4.23
$M \rightarrow AM$	28.84	0.920	0.141	**0.764**	0.778	0.636	0.837	**0.290**	0.903	0.934	3.98
$A \rightarrow AM$	28.02	0.906	0.134	0.738	0.755	0.640	0.842	0.287	0.897	0.940	**5.49**

Fig. 4: Visualization of the PIM effect.

results in the absence of environmental perception capabilities and exacerbates the control-quality dilemma in GHOI. With PIM, we observe that a unified joint generation significantly outperforms a cascaded pipeline, in which a separately trained (frozen) PIM generates a motion sequence to drive the AIM. The cascaded approach creates an information bottleneck. Its reliance on explicit motion conditions is too restrictive to capture object shape changes, which in turn leads to artifacts and reduces the plausibility of the generated hands. In contrast, joint training allows for a richer flow of information. Within the joint training paradigm, injecting only the last-layer feature from PIM into all AIM block layers mirrors the performance of the cascaded model, suggesting that globally applied motion is less effective than adaptive, layer-wise guidance. Our layer-wise residual injection mechanism, augmented with a zero-initialized linear layer, outperforms the layer-wise addition, showing the advantages of our design.

As shown in Fig. 4, we visualize the effects of our PIM. With frozen PIM, the cascaded variant can be regarded as a pose-driven model, while w/o PIM corresponds to a TIA2V model. The pose-driven variant can generate text-aligned interaction videos but is prone to abnormal object deformations. The TIA2V variant often struggles with semantic comprehension, resulting in inaccurate execution of requested interactions, whereas our method successfully achieves reasonable and text-aligned interaction videos. While PIM increases parameters, the computational cost for 256-pixel (short side) motion remains low. Inference time comparisons and ablations are in the Supplementary Material.

Model Architecture. We also explore the video-motion modeling in a single DiT, including video and motion concatenation via both the token sequence length (with modality-specific norms, modified RoPE, and type embedding) and channel dimensions (new patch embeddings). Results in Tab. 4 (top) show the ef-

Table 4: Motion Representation and Model Architecture Ablation.

Method	VLM-QA ↑	HQ ↑	OQ ↑	DD ↑	PI ↓	$DINO_{sub}$ ↑	$Sync_{conf}$ ↑
Sequence-Cat	26.52	0.796	0.113	0.452	0.692	0.547	3.85
Channel-Cat	26.23	0.758	0.104	0.531	0.698	0.596	3.41
2D (x,y) Rep.	26.07	0.705	0.094	0.698	0.655	0.623	4.72
2D (x,y) Cascade	25.58	0.884	0.121	0.659	0.672	0.625	5.04
Ours	**28.89**	**0.925**	**0.144**	**0.761**	**0.780**	**0.646**	**5.43**

Table 5: Real Scene Comparison.

Method	VLM-QA ↑	HQ ↑	OQ ↑	PI ↓	$DINO_{sub}$ ↑	$CLIP_{text}$ ↑	$Sync_{conf}$ ↑
Wan-S2V	24.97	0.851	0.115	0.740	0.646	0.285	5.36
HY-Avatar	25.29	0.817	0.108	0.724	0.644	0.285	5.45
HuMo	26.23	0.840	0.116	0.781	0.635	0.288	5.33
Ours	**28.49**	**0.910**	**0.135**	**0.794**	**0.665**	**0.289**	**5.51**

fectiveness of our method. While these designs inherently align motion and video, we found they degrade audio sync and yield weaker dynamics than dual-stream design. Sequence-Cat suffers from color consistency issues; Channel-Cat leads to more frequent artifacts. Furthermore, the lack of explicit scene-perception reduces the accuracy of motion and its harmony with environments.

Motion Representation. Table 4 (middle) compares our RGB motion representation against 2D coordinate baseline. We observe that coordinate-based models lack the rich semantic priors inherent in video foundation models and remain context-unaware; they cannot perceive the environment of the initial frame and ground objects from it. Consequently, these models exhibit poor generalization and require manual initialization (e.g., object locations). Furthermore, the modality gap between coordinates and pixels necessitates a cascaded "generate-then-render" pipeline, which underperforms compared to our joint generation approach. By operating in the RGB space, our method leverages established video priors to achieve superior text-to-motion alignment and GHOI (General Human-Object Interaction) synthesis with significantly lower data requirements.

Real Scene Evaluation. For better SOTA comparison, we collect 50 cases from real scenes to build an additional test set. Quantitative experimental results on Tab. 5 demonstrate that our method performs well in real scenes.

Multimodal Training Strategy. This section investigates the optimal strategy for sequencing multimodal conditions during AIM training. As detailed in Tab. 3 (bottom), $A \rightarrow AM$ denotes first training on audio, followed by joint training on audio and motion. AM represents training both modalities from the beginning, without any AIM audio pre-training. The $A \rightarrow AM$ strategy achieves substantially better performance on audio metrics than either joint training from the beginning (AM) or pre-training on the motion modality first ($M \rightarrow AM$). This outcome stems from the asymmetric nature of the conditioning signals. Consistent with the conclusions of Omnihuman [22], audio is a comparatively weaker signal than interaction motion and requires a dedicated pre-training stage for effective learning; otherwise, its influence is overshadowed by the dominant motion signal. Additionally, we augment the audio-driven pre-training with Text-to-Image-to-Video (TI2V) data, even in the absence of paired audio. This strategy ensures the model retains robust motion dynamics and general generative priors, preventing the network from over-fitting to simplistic audio-driven patterns.

User Study. We conduct a user study with 20 participants in 50 generation samples (each method). Given two videos generated from different methods, participants are asked to choose the better one. Results in Fig. 5.

5 Conclusion

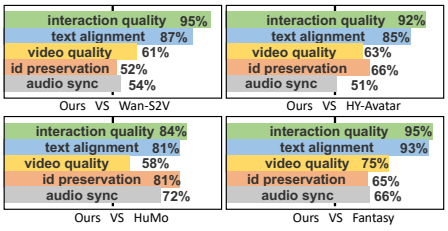

Fig. 5: User Study Results.

This paper introduces a dual-stream talking avatar video generation framework that enables the generation of grounded human-object interactions. Our method decouples perception and planning from video synthesis, enabling the generation of plausible, controllable, and high-quality GHOI videos. The PIM is carefully designed to perceive the context of ref image and plan text-aligned interaction motion, while AIM generates vivid talking and interacting avatars with an M2V aligner. A benchmark is proposed for evaluating GHOI generation. Comprehensive experiments showcase our model's performance and highlight each module's contributions and impacts. Our work provides valuable insights for future research in this field. The main limitation of our work is that it can only handle a single-person scene and is not designed to generate multi-person involved human-object interaction.

Acknowledgement

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404 and Shenzhen Key Laboratory of New Generation Interactive Media Technology Innovation(ZDSYS20210623092001004).

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
2. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. arXiv preprint arXiv:2502.02492 (2025)
3. Chen, L., Ma, T., Liu, J., Li, B., Chen, Z., Liu, L., He, X., Li, G., He, Q., Wu, Z.: Humo: Human-centric video generation via collaborative multi-modal conditioning (2025), `https://arxiv.org/abs/2509.08519`
4. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025)
5. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2403–2410 (2025)
6. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., et al.: Wan-animate: Unified character animation and replacement with holistic replication. arXiv preprint arXiv:2509.14055 (2025)
7. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)

8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
10. Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., et al.: Cosyvoice 2: Scalable streaming speech synthesis with large language models. arXiv preprint arXiv:2412.10117 (2024)
11. Fan, Y., Yang, Q., Wang, K., Zhou, H., Li, Y., Feng, H., Ding, E., Wu, Y., Wang, J.: Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17550–17560 (2025)
12. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
13. Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., et al.: Wan-s2v: Audio-driven cinematic video generation. arXiv preprint arXiv:2508.18621 (2025)
14. Hu, L., Wang, G., Shen, Z., Gao, X., Meng, D., Zhuo, L., Zhang, P., Zhang, B., Bo, L.: Animate anyone 2: High-fidelity character image animation with environment affordance. arXiv preprint arXiv:2502.06145 (2025)
15. Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., Lu, Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation (2025), https://arxiv.org/abs/2505.04512
16. Huang, Z., Zhou, Z., Cao, J., Ma, Y., Chen, Y., Rao, Z., Xu, Z., Wang, H., Lin, Q., Zhou, Y., et al.: Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. arXiv preprint arXiv:2506.08797 (2025)
17. Jiang, J., Liang, C., Yang, J., Lin, G., Zhong, T., Zheng, Y.: Loopy: Taming audio-driven portrait avatar with long-term motion dependency. arXiv preprint arXiv:2409.02634 (2024)
18. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
20. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7752–7762 (2025)
21. Lin, G., Jiang, J., Liang, C., Zhong, T., Yang, J., Zheng, Y.: Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. arXiv preprint arXiv:2409.01876 (2024)
22. Lin, G., Jiang, J., Yang, J., Zheng, Z., Liang, C.: Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv preprint arXiv:2502.01061 (2025)
23. Liu, K., Liu, Q., Liu, X., Li, J., Zhang, Y., Luo, J., He, X., Liu, W.: Hoigen-1m: A large-scale dataset for human-object interaction video generation. In: Proceedings

of the Computer Vision and Pattern Recognition Conference. pp. 24001–24010 (2025)
24. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025)
25. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
26. Men, Y., Yao, Y., Cui, M., Bo, L.: Mimo: Controllable character video synthesis with spatial decomposed modeling. In: Computer Vision and Pattern Recognition (CVPR), 2025 IEEE Conference on (2025)
27. Meng, R., Zhang, X., Li, Y., Ma, C.: Echomimicv2: Towards striking, simplified, and semi-body human animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5489–5498 (2025)
28. Pang, Y., Shao, R., Zhang, J., Tu, H., Liu, Y., Zhou, B., Zhang, H., Liu, Y.: Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12209–12219 (2025)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
31. Tan, S., Gong, B., Wang, X., Zhang, S., Zheng, D., Zheng, R., Zheng, K., Chen, J., Yang, M.: Animate-x: Universal character image animation with enhanced motion representation. arXiv preprint arXiv:2410.10306 (2024)
32. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: European Conference on Computer Vision. pp. 244–260. Springer (2024)
33. Tu, S., Xing, Z., Han, X., Cheng, Z.Q., Dai, Q., Luo, C., Wu, Z.: Stableanimator: High-quality identity-preserving human image animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21096–21106 (2025)
34. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, C., Tian, K., Zhang, J., Guan, Y., Luo, F., Shen, F., Jiang, Z., Gu, Q., Han, X., Yang, W.: V-express: Conditional dropout for progressive training of portrait video generation. arXiv preprint arXiv:2406.02511 (2024)
36. Wang, M., Wang, Q., Jiang, F., Fan, Y., Zhang, Y., Qi, Y., Zhao, K., Xu, M.: Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. arXiv preprint arXiv:2504.04842 (2025)

37. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
38. Wang, X., Zhang, S., Tang, L., Zhang, Y., Gao, C., Wang, Y., Sang, N.: Unianimate-dit: Human image animation with large-scale video diffusion transformer. arXiv preprint arXiv:2504.11289 (2025)
39. Wang, Z., Yang, J., Jiang, J., Liang, C., Lin, G., Zheng, Z., Yang, C., Lin, D.: Interacthuman: Multi-concept human animation with layout-aligned audio conditions. arXiv preprint arXiv:2506.09984 (2025)
40. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)
41. Xu, Z., Huang, Z., Cao, J., Zhang, Y., Cun, X., Shuai, Q., Wang, Y., Bao, L., Li, J., Tang, F.: Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation. arXiv preprint arXiv:2411.17383 (2024)
42. Xue, Z.S., Luo, R., Chen, C., Grauman, K.: Hoi-swap: Swapping objects in videos with hand-object interaction awareness. Advances in Neural Information Processing Systems **37**, 77132–77164 (2024)
43. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)
44. Zhang, F., Tian, S., Huang, Z., Qiao, Y., Liu, Z.: Evaluation agent: Efficient and promptable evaluation framework for visual generative models. arXiv preprint arXiv:2412.09645 (2024)
45. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection (2022)
46. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: 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. pp. 8652–8661 (2023)
47. Zhang, Y., Li, Z., Wang, D., Zhang, J., Zhou, D., Yin, Z., Dai, X., Yu, G., Li, X.: Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. arXiv preprint arXiv:2507.09862 (2025)