Human-Centric Video Generation via Collaborative Multi-Modal Conditioning

Liyang Chen[1], Tianxiang Ma[2], Jiawei Liu[2], Bingchuan Li[2*], Zhuowei Chen[2], Lijie Liu[2], Xu He[1], Gen Li[2], Qian He[2], Zhiyong Wu[1†]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Intelligent Creation Team, ByteDance

Abstract

Human-Centric Video Generation (HCVG) methods seek to synthesize human videos from multimodal inputs, including text, images, and audio. Existing methods struggle to effectively coordinate these heterogeneous modalities due to two challenges: the scarcity of modality-complete data and the difficulty of jointly modeling triplet conditions without performance degradation. In this work, we present **HuMo**, a unified HCVG framework for collaborative multimodal control. For the first challenge, we construct an incomplete-yet-complementary dataset for improved data utilization efficiency and training scalability. For the second challenge, we propose a two-stage progressive multimodal training paradigm with task-specific strategies at each stage. In the first stage, to balance the text-following and subject-preservation abilities, we adopt the minimal-invasive image injection strategy. In the second stage, to enhance audio-visual sync, we propose a focus-by-predicting strategy that implicitly guides the model to associate audio with facial regions. For joint learning of controllabilities across multimodal inputs, we progressively incorporate the audio-visual sync task, building on previously acquired capabilities. During inference, for flexible and fine-grained multimodal control, we design a stage-adaptive Classifier-Free Guidance strategy that dynamically adjusts guidance weights across denoising steps. Extensive experimental results demonstrate that HuMo surpasses specialized state-of-the-art methods in sub-tasks, establishing a unified framework for collaborative multimodal-conditioned HCVG.

Code — https://github.com/Phantom-video/HuMo
Extended version — https://arxiv.org/abs/2509.08519

Introduction

Recent advances in foundational video generation models have significantly accelerated progress in Human-Centric Video Generation (HCVG) (Lin et al. 2025; Hu et al. 2025; Tian et al. 2025; Liu et al. 2025b; Chen et al. 2025b; He et al. 2025), improving both the fidelity and controllability. Such progress is democratizing short video production. Traditional filmmaking, entailing scene setup, casting, styling, and scripting, is labor-intensive and demands collective expertise, incurring substantial time and financial costs. HCVG methods offers a disruptive alternative by leveraging multimodal inputs: text for describing scenes and actions, images for defining human identity and subject appearance, and audio for character speaking. It greatly reduces manual overhead and enables scalable and efficient content creation.

Figure 1: We propose **HuMo**, a multimodal HCVG framework that supports flexible input compositions of text-image, text-audio, and text-image-audio. HuMo generalizes to humans, humans with objects or animals, stylized humanoid artworks, and animations.

Prior methods (Lin et al. 2025; Wang et al. 2025a) typically adopt a two-stage pipeline: a text-to-image (T2I) model (Gao et al. 2025a) generates a subject-complete start frame containing all required elements, followed by an image-to-video (I2V)-based model for animation and audio injection. However, this approach heavily depends on the subject-complete start frame, limiting text controllability to modify the provided subject and failing in cases with missing subjects, as shown in Fig. 2(a). On the other hand, subject-consistent video generation (S2V) methods (Liu et al. 2025b; Deng et al. 2025) leverage subject reference images and support flexible video customization through textual prompts. However, these methods are unable to incor-

*Project lead.
†Corresponding author.

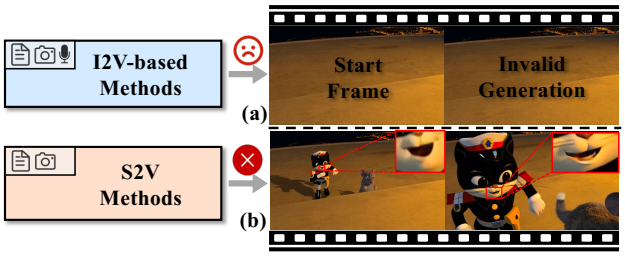


Figure 2: The I2V-based methods requiring a subject-incomplete start frame cannot generate valid or subject-consistent video, while traditional S2V methods cannot support audio input.

porate an audio modality and cannot control what a character is speaking, as shown in Fig. 2(b). This significantly limits their application in scenarios requiring audio-visual collaboration. Recent methods (Hu et al. 2025; Deng et al. 2025) attempt to integrate these two approaches mentioned above. However, they struggle to achieve effective collaboration among the triplet input modalities. For instance, emphasizing the influence of reference images often degrades audio-visual sync. Conversely, prioritizing high sync tends to compromise either the text following or the subject preservation ability.

This work focuses on the HCVG task conditioned on text, reference images, and audio, aiming to achieve collaborative multimodal control. There exist two major challenges: 1) Lack of modality-complete data. Only a small amount of data is paired with high-quality triplet conditions. Relying solely on such tri-modality complete data leads to extremely low data utilization and poor training scalability. 2) Difficulty in collaborative multimodal control. Within a multi-task training framework, it is difficult to simultaneously achieve strong text following, subject consistency with reference images, and accurate audio-visual sync, as these abilities tend to compromise or undermine one another during learning. To tackle these challenges, we propose HuMo. Our key insight lies in the joint design of the data processing pipeline and training paradigm that enables the collaborative learning of multimodal controllability with incomplete-yet-complementary multimodal data.

Specifically, to reduce reliance on fully modality-complete data and enhance training scalability, we build an incomplete-yet-complementary multimodal dataset by pairing large-scale available text-video samples with retrieved subject reference images and selectively filtered temporally-aligned audio. Building on the constructed dataset, to establish multimodal control capabilities, we propose a two-stage progressive multimodal training paradigm. 1) Subject preservation task. To enable text-image collaborative control, ensuring subject preservation without degrading the text following ability of the foundational DiT-based backbone (Wan et al. 2025), we adopt the minimal-invasive image injection strategy that avoids structural modifications of the DiT backbone and confines parameter updating to a limited subset. 2) Audio-Visual Sync. We introduce audio cross-attention layers to inject audio modality. Since our work cannot directly localize the audio's influence like previous methods (Yi et al. 2025; Wang et al. 2025a), to enhance synchronization between audio and facial motions, we propose a focus-by-predicting strategy by implicitly encouraging the model to associate audio signals with the face and upper body. To prevent audio-visual sync learning from undermining previously acquired ability, we retain the subject preservation task while progressively incorporating the audio-visual sync task. This joint optimization facilitates collaborative learning of controllability across text, image, and audio modalities.

During inference, to enable flexible and fine-grained control over different input modalities, we propose a stage-adaptive classifier-free guidance (CFG) strategy, whic dynamically adjusts the guidance strengths at different denoising steps, allowing for precise and efficient collaboration among text following, subject preservation, and audio-visual sync.

The main contributions can be summarized as follows:

- **Concept.** We attribute the imbalanced multimodal controllability in existing HCVG methods to the absence of a comprehensive design across data processing and training paradigms for handling heterogeneous inputs. We propose HuMo, a unified framework that enables collaborative control across text, image, and audio modalities within a single model. It seamlessly supports various input compositions of text-image, text-audio, and text-image-audio.
- **Methodology.** 1) We establish a multimodal data processing pipeline that reduces the dependence on modality-complete data. 2) We propose a progressive multimodal training paradigm with task-specific strategies, which facilitates joint learning of controllabilities across multiple modalities. 3) We design a stage-adaptive CFG strategy, enabling flexible, fine-grained, and collaborative multimodal control.
- **Significance.** Extensive experimental results show HuMo surpasses the specialized state-of-the-art (SOTA) methods on both subject preservation and audio-visual sync sub-tasks. We further demonstrate its effectiveness and scalability by validating the framework on models of 1.7B and 17B parameters.

Related Works

Audio-Driven Human Animation

Audio-driven human animation aims to generate videos from input human images to produce lip movements matching input speech signals. Built on the foundational video generation models (Kong et al. 2024; Wan et al. 2025; Gao et al. 2025b; Seawead et al. 2025; Chen et al. 2025c), audio-driven human animation methods have achieved impressive performance (Chen et al. 2023, 2025a,b). Hallo3 (2025) is the first full-frame-size portrait animation application on a pre-trained DiT model, using cross-attention layers for audio control. To improve the motion dynamics, FantasyTalking (2025a) proposes to establish global motion at the clip

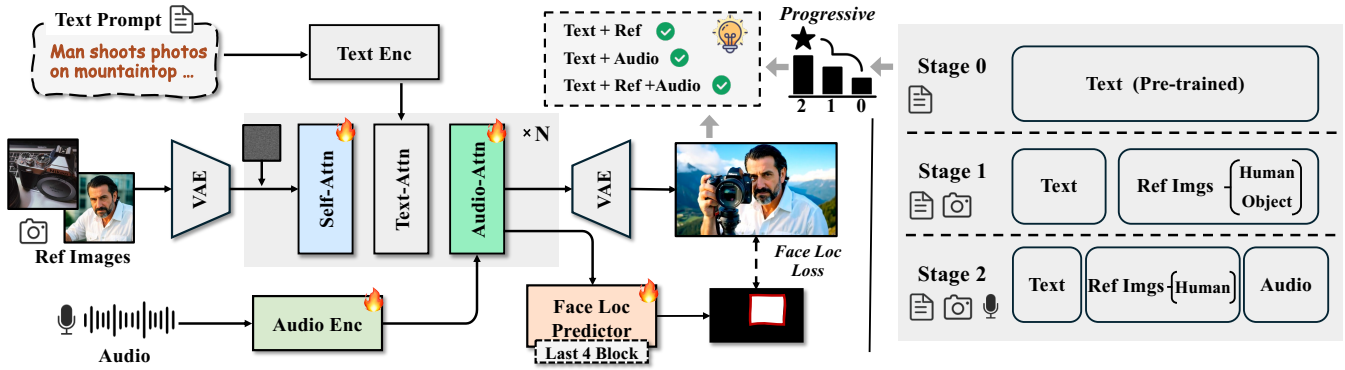


Figure 3: **Overview of our framework.** HuMo model (left) is trained based on the proposed data processing pipeline (right). Built upon a DiT-based T2V backbone from Stage 0, the model progressively learns subject preservation and audio-visual sync capabilities in Stages 1 and 2, enabling collaborative generation across different modality compositions.

level and optimize lip synchronization at the frame level. OmniHuman-1 (2025) scales the training data by incorporating hybrid motion-related conditions to generate more realistic human videos. Despite their strong performance in facial animation and body motion, existing methods still require a subject-complete start frame with visible facial features, which limits user creativity.

Subject-Consistent Video Generation

Subject-consistent video generation (S2V) aims to extract subject features from reference images and generate videos with consistent subject identity based on text prompts. Early approaches use pre-trained semantic encoders (Radford et al. 2021) to extract features from reference images, achieving identity-preserving video generation via adapters (He et al. 2024; Yuan et al. 2025b) or DiT cross-attention layers (Polyak et al. 2024). In-context methods (Liu et al. 2025b; Deng et al. 2025) leverage the inherent consistency of pre-trained DiT-based models by concatenating reference image latents with noisy video latents. These methods achieve finer-grained subject consistency but weaken textual control. To preserve the textual control of the pre-trained model in our in-context method, we freeze the text-visual cross-attention layers and fine-tune only the self-attention layers. Moreover, existing methods support only image and text modalities and cannot support human speech input, which is critical for vivid and realistic video creation. Concurrent works such as InterActHuman (2025c) and HunyuanCustom (2025) integrate subject-consistent video generation with audio-driven human animation. However, they still struggle to balance the influence of these three modalities. In contrast, we propose a progressive multimodal collaborative training paradigm and stage-adaptive CFG to enable flexible multimodal control.

Methodology

To address the challenges of modality-complete data scarcity and compromised performance in multimodal conditioning, we propose **HuMo**, a **Hu**man-Centric Video Generation framework that enables collaborative control with **Mo**ltimodal conditions. Given a textual prompt c_{txt}, reference images c_{img}, and an audio singal c_{a}, HuMo aims to generate a video where the environment, human identity, appearance, accessories, clothing, as well as facial and body motions are semantically aligned with the input conditions, while remaining spatially and temporally coherent. We begin in Sec. by describing the DiT-based T2V backbone (Wan et al. 2025). Sec. outlines the construction of the multimodal dataset. In Sec. , we present how a T2V model is extended to support triplet modalities. Finally, Sec. describes our inference strategy for flexibly modality conditioning and fine-grained and collaborative generation.

Preliminaries

In this work, we adopt a DiT-based T2V model (Wan et al. 2025) as our backbone. As shown in Fig. 3, it incorporates a 3D Variational Autoencoder (VAE) to compress video into a compact latent space. A text encoder is utilized to encode textual information and then injected into the DiT backbone with cross-attention. For training, flow matching is adopted by learning a continuous velocity field that transports samples from a simple prior to the data distribution along deterministic trajectories. HuMo inherits this DiT-based architecture and extends it by incorporating additional image and audio modalities, enabling multimodal conditioning for more controllable video generation. To effectively train the model under this multimodal setup, we formulate the flow matching objective as:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,z_0,z_1}\|v_\theta(z_t, t, c) - (z_1 - z_0)\|_2^2, \quad (1)$$

where z_0 is a sample from the simple prior, z_1 is the latent of the target video sample, and $c = \{c_{\text{txt}}, c_{\text{img}}, c_{\text{a}}\}$ denotes the multimodal condition of text, image, and audio. The DiT model v_θ takes the noisy latent $z_t = (1-t)z_0 + tz_1$, and learns to predict its velocity at any point $t \in [0, 1]$. This formulation allows HuMo to efficiently learn a deterministic transport map from noise to data under complex multimodal constraints.

Multimodal Data Processing Pipeline

High-quality, subject-consistent, and audio-visual synced multimodal data is crucial for HCVG but remains scarce and usually incomplete or misaligned. We propose a multimodal data processing pipeline to construct an incomplete-yet-complementary dataset, which forms the foundation for the training paradigm we introduce later.

As shown in Fig. 3, our pipeline unfolds in several stages. In Stage 0, we begin with a large-scale video pool (Wang et al. 2025b; Li et al. 2025b) and leverage powerful VLMs (Bai et al. 2025; Team 2025) for detailed text descriptions, ensuring basic textual modality for each video sample. In Stage 1, to avoid the common copy-paste issue (Liu et al. 2025b) in subject preservation, we follow prior works (Chen et al. 2025d; Yuan et al. 2025a) to sparsely sample video frames and retrieve cross-pair reference images from a billion-scale image corpus. For humans, we retrieve images with the same identity but varying in appearance (e.g., makeup, pose, background, clothing, age); for objects, we match the same semantic category but with varied visual attributes (e.g., color, shapes, viewpoint). This strategy improves text controllability for reference-guided generation by discouraging direct frame replication in training. In Stage 2, we supplement the data with audio modality by filtering for speech segments and creating tightly aligned audio-visual pairs with lip-sync analysis (Li et al. 2025a). It is worth noting that in Stage 2, since audio-aligned data predominantly exists in human-centric videos, we can only find human reference images. Thus, data in Stage 2 remains inherently incomplete.

Through this pipeline, we build an incomplete-yet-complementary multimodal dataset. Stage 1 includes $\mathcal{O}(1)$M video samples with text and reference images, while Stage 2 contains $\mathcal{O}(50)$K samples with temporally-aligned-audio. Compared to relying heavily on fully modality-complete data, our method offers greater data efficiency and scalability, enabling effective continue training even in the presence of missing modalities.

Progressive Multimodal Training

We structure the acquisition of multimodal control capabilities into two distinct yet progressive training stages. Stage 1 establishes text-image controllability through the subject preservation task. Stage 2 realizes the joint learning of text-image-audio controllability via progressive training of the previous task and the audio-visual sync task. Training data for each stage is sourced from the corresponding stage of our data processing pipeline.

Subject Preservation. In Stage 1, we introduce reference images as additional inputs. To preserve the strong text following and image synthesis ability of the original DiT backbone, we adopt the minimal-invasive image injection strategy, which adheres to two key principles: avoiding structural modifications of the DiT backbone and confining parameter updating to a limited subset.

Specifically, without modifying the model architecture, we concatenate the VAE latents z_{img} of the reference images c_{img} with the noisy latent z_t along the temporal dimension as inputs. c_{img} can be a single image of a human or an object, or a composition of multiple images. To prevent the model from misinterpreting the reference as the start frame and performing undesired image continuation, we always place the reference latents at the end of the video latent sequence, forming the input as $[z_t; z_{\text{img}}]$. This design encourages the model to actively extract subject identity information from the reference images via self-attention and propagate it across all frames, enabling subject-consistent generation. The training is restricted to the self-attention layers of DiT to minimally affect its inherent text alignment and visual generation capabilities. Notably, text remains as input to ensure semantic consistency and controllability in the generated videos.

Audio-Visual Sync. In stage 2, we extend the first-stage setup by incorporating audio modality. Following prior works (Wang et al. 2025a; Kong et al. 2025), we insert an audio cross-attention layer in each DiT block. Audio features are extracted via Whisper (Radford et al. 2023) for generalization across speakers and languages. Based on the observation that human motions are primarily aligned with temporally nearby audio cues (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017), we construct the final audio embedding $c_{\text{a}} \in \mathbb{R}^{f\times n\times d}$ by concatenating audio embeddings within a temporal window centered at each video frame stamp, where the window length $n = 5$ and and f same with the sequence length of the video latents. c_{a} is conducted cross-attention calculation with hidden states frame-by-frame:

$$\text{Attention}(h_z, c_{\text{a}}) = \text{softmax}(\frac{\mathbf{Q}_z\mathbf{K}_{\text{a}}^{\top}}{\sqrt{d}})\mathbf{V}_{\text{a}}, \qquad (2)$$

where $\mathbf{Q}_z$ transformed from the hidden states $h_z \in \mathbb{R}^{f\times h\times w\times d}$ of video latent in DiT net, $\mathbf{K}_a$ and $\mathbf{V}_a$ transformed from audio embedding.

The aforementioned cross-attention is computed between the audio signal and the full-frame video latent, but audio cues are usually most correlated with localized human regions (e.g., face, lip). Consequently, this coarse-grained attention can lead to weak synchronization between audio and human motion. We therefore propose a focus-by-predicting strategy to implicitly guide the model to focus more on the facial region. During training, we introduce a mask predictor $\mathcal{F}_{\text{mask}}$ to estimates the potential facial region distribution $\mathbf{M}_{\text{pred}} = \mathcal{F}_{\text{mask}}(h_z)$. $\mathbf{M}_{\text{pred}}$ is supervised by the ground-truth binary face mask $\mathbf{M}_{\text{gt}}$ using a binary cross-entropy (BCE) loss. Since early DiT blocks lack stable spatial representations, we only insert $\mathcal{F}_{\text{mask}}$ after the audio cross-attention modules in the last four DiT blocks and average these outputs as the final predicted mask. To further prevent the supervision from weakening when the face region is too small, we introduce a size-aware weight (Yi et al. 2025) by adaptively emphasizing the supervision on the face region:

$$\mathcal{L}_{\text{mask}} = \frac{hw}{\sum_{i=1}^{h}\sum_{j=1}^{w}\mathbf{M}_{\text{gt}}^{(i,j)}} \cdot \text{BCE}(\mathbf{M}_{\text{pred}}, \mathbf{M}_{\text{gt}}). \qquad (3)$$

Unlike prior methods (Yi et al. 2025) that apply hard gating on audio attention outputs, which is a truncation design

Figure 4: **Qualitative comparison for the subject preservation task.** Zoom in for details.

Method	Video Quality			Text Following	Subject Consistency			
	AES↑	IQA↑	HSP↑	TVA↑	ID-Cur↑	ID-Glink↑	CLIP-I↑	DINO-I↑
Kling 1.6 (2025)	0.645	0.714	3.792	2.564	0.470	0.501	0.639	0.394
MAGREF (2025)	0.622	0.708	3.331	2.852	0.334	0.359	0.665	0.416
HunyuanCustom (2025)	0.592	0.705	3.705	1.777	0.309	0.335	0.649	0.426
Phantom (2025b)	0.608	0.705	3.612	2.877	0.649	0.674	0.677	0.426
Ours-1.7B	0.586	0.680	3.432	3.222	0.609	0.668	0.660	0.414
Ours-17B	**0.657**	**0.717**	**3.906**	**3.939**	**0.731**	**0.757**	**0.687**	**0.447**

Table 1: **Quantitative results for the subject preservation task** with text and reference images as inputs.

we consider suboptimal. The focus-by-predicting strategy acts as a soft regularizer, which steers the model's focus without crippling its representational capacity and retains the flexibility to model full-body kinematics and complex interactions.

Progressive Training. To ensure that the acquisition of audio-visual sync ability does not degrade the subject preservation ability, we employ a progressive task-weighting curriculum in stage 2. Initially, training is dominated by the subject preservation task (80% ratio, audio input as null) to reinforce existing ability, while the audio-visual sync task constitutes the remaining 20%. As training progresses, we gradually increase the proportion of the audio-visual sync task to 50%. Throughout this stage, we still adhere to the finetuning principles in stage 1 by only updating the self-attention layers and audio-related modules. This progressive strategy facilitates a smooth transition for the model from bi-modal (text, image) to tri-modal (text, image, audio) inputs, ensuring stable training and the collaboration of multimodal learning.

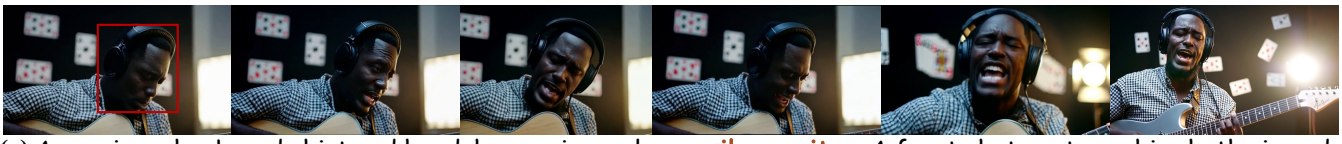
Inference Strategies

Flexible Multimodal Control. For flexible, independent control at inference, we adapt CFG with separate guidance scales (λ_{txt}, λ_{img}, λ_{a}) for each modality:

$$\begin{aligned} v_\theta(z_t, t, c) = \lambda_{\text{a}}[v_\theta(c_{\text{txt}}, c_{\text{img}}, c_{\text{a}}) - v_\theta(c_{\text{txt}}, c_{\text{img}}, \varnothing)] \\ + \lambda_{\text{img}}[v_\theta(c_{\text{txt}}, c_{\text{img}}, \varnothing) - v_\theta(c_{\text{txt}}, \varnothing, \varnothing)] \\ + \lambda_{\text{txt}}[v_\theta(c_{\text{txt}}, \varnothing, \varnothing) - v_\theta(\varnothing, \varnothing, \varnothing)] \\ + v_\theta(\varnothing, \varnothing, \varnothing). \end{aligned} \tag{4}$$

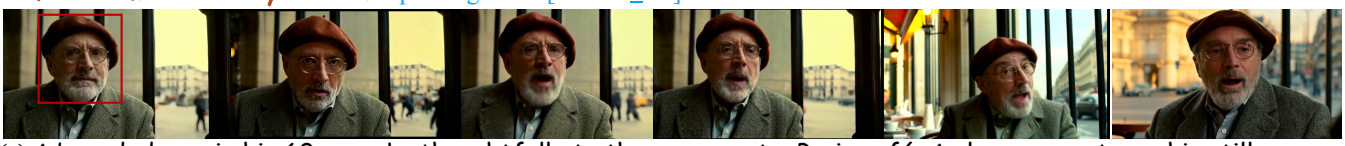
HuMo can generate coherent results even when modalities are absent, enabling condition combination of $[c_{\text{txt}}, c_{\text{img}}]$, $[c_{\text{txt}}, c_{\text{a}}]$, and $[c_{\text{txt}}, c_{\text{img}}, c_{\text{a}}]$. Missing conditions are simply replaced by a null token $\varnothing$.

Stage-Adaptive CFG. We observe that the influence of each modality shifts throughout the denoising process: early steps tend to construct the overall semantic structure and spatial layout guided by text and image, while later steps focus on fine-grained details (e.g., identity similarity, audio-visual sync). A static CFG configuration is suboptimal for the shifting modal dynamics. We therefore propose a stage-adaptive CFG that dynamically switches between two configurations of guidance scales. From timestep 1.0 to 0.98, we adopt the text/image-dominant configuration to establish a stable semantic layout (e.g., character and scene composition). From timestep 0.98 to 0, we shift to parameters that emphasize audio and image control. Empirical results demonstrate that this strategy significantly improves multimodal collaboration and enhances overall video quality.



(a) A man in a checkered shirt and headphones sings, plays a silver guitar. A front shot captures his rhythmic and emotionally engaged expression against a lit, card-decorated black wall. Speaking Word [Gas /ɡæs/]

(b) A woman in a bulky space suit with a large helmet speaks to the camera against a blurred Mars backdrop. Her face remains clearly visible. Speaking Word [One /wʌn/]



(c) A bearded man in his 60s speaks thoughtfully to the camera at a Paris café. A close-up captures his still upper body, lit by cinematic golden light with Parisian streets in the background. Speaking Word [Power /ˈpaʊər/]

Start/Ref Image **OmniHuman-1** **Hallo3** **FantasyTalking** **HunyuanCustom** **Ours**

Figure 5: **Qualitative comparison for audio-visual sync task.** For the I2V-based methods, we use the first frame generated by MoCha (Wei et al. 2025) as the input start frame. For HunyuanCustom and Ours, the cropped face from the start frame is utilized as input. The official OmniHuman-1 (2025) website API does not support text input. Zoom in for details.

Method	Video Quality			Text Following	Subject Consistency		Audio-Visual Sync	
	AES↑	IQA↑	HSP↑	TVA↑	ID-Cur↑	ID-Glink↑	Sync-C↑	Sync-D↓
OmniHuman-1 (2025)	0.545	0.682	4.503	-	0.677	0.727	**6.526**	**7.784**
Hallo3 (2025)	0.381	0.634	4.200	6.117	0.726	0.727	5.189	9.212
FantasyTalking (2025a)	0.455	0.652	4.444	6.209	0.703	0.729	3.202	10.914
HunyuanCustom (2025)	0.358	0.619	4.370	6.246	0.729	0.716	4.562	9.892
Ours-1.7B	0.322	0.661	4.350	5.865	0.721	0.729	6.005	8.648
Ours-17B	**0.589**	**0.718**	**4.537**	**6.508**	**0.747**	**0.740**	6.252	8.577

Table 2: **Quantitative results for the audio-visual sync task** on MoCha benmark.

Experiment

Implementation Details. We build our method on the backbones of Wan-2.1-1.3B and Wan-2.1-14B (Wan et al. 2025). All video samples are resampled to 25 fps with a 480×832 resolution. We employ a two-stage training strategy. Stage 1 involves training for 40k steps, with all audio-related modules disabled. Stage 2 continues for another 40k steps, where we enable the audio-related modules.

Comparative Methods. We compare HuMo with two categories of baselines. **1) S2V methods** that take text and subject reference images as input, including open-source models MAGREF (2025), HunyuanCustom (2025), Phantom-Wan-14B (2025b), and the commercial closed-source method Kling 1.6 (2025). Comparison is conducted on an in-house benchmark of 100 test cases involving humans, objects, and animals. **2) Audio-visual synced methods** that take text, image, and audio as inputs, including open-source methods Hallo3 (2025), FantasyTalking (2025a), HunyuanCustom (2025), and the commercial closed-source OmniHuman-1 (2025). Comparison is conducted on the MoCha benchmark (2025). Notably, HunyuanCustom is the only method that supports reference images, while all other baselines rely on a subject-complete start frame and follow the I2V generation paradigm. The latter methods are easier to achieve higher visual quality.

Evaluation Metrics. We conduct comprehensive evaluation covering four key aspects. **1) Video Quality.** We evaluate visual appeal and perceptual quality using aesthetics (AES) and image quality assessment (IQA) from the widely-used VBench (2024). Specifically for human videos, we leverage the SOTA VLM, Gemini-2.5-Pro (2025), to estimate the human structure plausibility (HSP). **2) Text-Video Alignment (TVA).** Semantic consistency between the input text prompt and generated video is measured via the VLM-based reward model (Liu et al. 2025a). **3) Subject Consistency.** For Human identity, we detect and crop faces from generated frames, and compute similarity with the reference image using Face-Cur (2020) and Face-Glink (2021). For non-facial

Figure 6: **Qualitative ablation study** on OpenS2V-Nexus (Yuan et al. 2025a) benchmark. Prog: progressive.

objects, we use DINO-I (2024) and CLIP-I (2021) scores to compute average embedding similarity. **4) Audio-Visual Sync.** We adopt Sync-C and Sync-D (Li et al. 2025a) to quantify alignment between input audio and facial motion.

Multimodal Conditioned Comparison

Subject Preservation Task. 1) Qualitatively. Fig. 4 shows that HuMo demonstrates superior performance in text following ability compared to other methods. For instance, in case (b), where the prompt is "step into a temple", other methods fail to generate the correspondence. Our method shows strong subject preservation and generalizes well to unseen four-subject cases, accurately maintaining four distinct human identities, while other methods suffer from missing person or human identity drift. HuMo also excels in visual aesthetics and human structure plausibility. In (a), HuMo generates a person wearing gloves without structural artifacts, whereas baselines show noticeable limb degradation. In (c), HuMo seamlessly integrates a background image into the generation, despite not being trained on any background images. **2) Quantitatively.** Tab. 1 reveals that HuMo outperforms other methods in aesthetic and overall image fidelity. HuMo achieves a strong capability in modeling human pose and body integrity with the highest HSP score. For text following and subject consistency with reference images, HuMo also achieves the SOTA performance, which is consistent with the qualitative observation. This highlights our model's capabilities to achieve strong textual editability without compromising subject consistency.

Audio-Visual Sync Task. 1) Qualitatively. Fig. 5 first reveals HuMo's superior capability in text following. In case (a) and (c), HuMo successfully synthesizes the "silver guitar" and "golden light in the background" respectively, whereas other methods fail to render these specific details. This observation underscores a fundamental weakness of I2V-based methods - their limited ability for re-editing the provided subject-complete start frame. In case (b), when provided with a dimly lit facial image, other methods are unable to generate the "clearly visible face" as required by the text prompt. HuMo, by contrast, not only synthesizes a clear face but also preserves the identity of the reference face image. **2) Quantitatively.** As shown in the Tab. 2, HuMo attains the highese scores for aesthetic quality and text follwing. In terms of HSP and identity similarity, our method outperforms HunyuanCustom. Notably, HuMo also surpasses other I2V-based methods, despite their inherent advantage with stronger priors on body layout and facial structure. For audio-visual sync, our 1.7B model already outperforms several open-source specialized models and trails only slightly behind the commercial method OmniHuman-1. It is crucial to note that for this evaluation, we only utilize a single reference face image. As illustrated in Fig. 1, HuMo supports a combined input of audio and multiple reference images (e.g., facial photos, clothing, animal), offering enhanced controllability and customization.

Variants	AES ↑	TVA ↑	ID-Cur ↑	Sync-C ↑
Full Finetune	0.529	6.157	**0.749**	6.250
w/o Progressive	0.541	6.375	0.724	6.106
w/o Face Loc	0.587	6.507	0.730	5.946
Ours-17B	**0.589**	**6.508**	0.747	**6.252**

Table 3: **Quantitative ablation study** on MoCha data.

Ablation Study

We conduct ablation studies on three key designs of our method. 1) Full Finetuning. We update all parameters of the DiT model instead of confining parameter updating to a limited subset. It leads to significant drops of AES and TVA scores in Tab. 3. Visually, Fig. 6(a) fails to generate the "phone", and case (b) exhibits noticeable artifacts. Full finetuning disrupts the DiT's pretrained capabilities for high-quality video synthesis and text-image alignment. 2) w/o Progressive Training. We train two tasks in a single stage. This change leads to a degradation across most evaluation metrics. The generated character exhibits low identity similarity to the reference image. This suggests that without progressively building different capabilities, the model struggles to achieve effective coordination across modalities. 3) w/o Face Loc. We remove the focus-by-predicting strategy. It results in a decline in Sync-C score, and the generated lip movements are misaligned with the spoken word. This highlights the importance of face location prediction for learning audio-visual correspondence. Furthermore, we observe a decline in ID-Cur score, indicating that this strategy also implicitly contributes to better facial identity consistency.

Conclusion

We propose HuMo, a novel human-centric video generation framework with multimodal conditions. HuMo establishes a multimodal data processing pipeline, which effectively alleviates the reliance on modality-complete data. The proposed progressive multimodal training paradigm successfully integrates the control capabilities of text, image, and audio modalities into a unified model. Benefiting from the proposed stage-adaptive CFG strategy, our model enables flexible, fine-grained, and collaborative control over all three modalities during inference. HuMo satisfies the multiple requirements of text prompt following, subject preservation, and audio-visual sync in human-centric video creation.

Ethical Statement

The development of HuMo for Human-Centric Video Generation may raise several ethical concerns. First, the ability to synthesize realistic human videos from multimodal inputs (text, image, and audio) may lead to misuse, such as deepfakes or non-consensual content creation. Ensuring informed consent and protecting individuals' likenesses are critical. Second, fine-grained control over generated content calls for responsible usage guidelines to prevent manipulation or misinformation. Developers and users must adhere to ethicalstandards, including transparency, data privacy, and the prevention of harm.

Acknowledgments

Special thanks to Ronggui Peng and Bingqian Yi for their assistance with the multimodal data construction. This work is supported by National Natural Science Foundation of China (62076144) and Shenzhen Science and Technology Program (JCYJ20220818101014030).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; and et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Chen, L.; Bao, W.; Lei, S.; Tang, B.; Wu, Z.; Kang, S.; Huang, H.; and Meng, H. 2025a. AdaMesh: Personalized Facial Expressions and Head Poses for Adaptive Speech-Driven 3D Facial Animation. *IEEE Transactions on Multimedia*, 27: 3598–3609.

Chen, L.; Wu, Z.; Li, R.; Bao, W.; Ling, J.; Tan, X.; and Zhao, S. 2023. VAST: Vivify Your Talking Avatar via Zero-Shot Expressive Facial Style Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, L.; Zhou, T.; He, X.; Tang, B.; Wu, Z.; Huang, Y.; Wu, Y.; Sun, Z.; Yang, W.; and Meng, H. 2025b. StableDub: Taming Diffusion Prior for Generalized and Efficient Visual Dubbing. arXiv:2509.21887.

Chen, S.; Ge, C.; Zhang, Y.; Zhang, Y.; Zhu, F.; Yang, H.; Hao, H.; Wu, H.; Lai, Z.; and et al. 2025c. Goku: Flow Based Video Generative Foundation Models. *arXiv preprint arXiv:2502.04896*.

Chen, Z.; Li, B.; Ma, T.; Liu, L.; Liu, M.; Zhang, Y.; Li, G.; Li, X.; Zhou, S.; He, Q.; and Wu, X. 2025d. Phantom-Data : Towards a General Subject-Consistent Video Generation Dataset. arXiv:2506.18851.

Cui, J.; Li, H.; Zhan, Y.; Shang, H.; Cheng, K.; Ma, Y.; Mu, S.; Zhou, H.; Wang, J.; and Zhu, S. 2025. Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21086–21095.

Deng, J.; Guo, J.; Yang, J.; Xue, N.; Cotsia, I.; and Zafeiriou, S. P. 2021. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Deng, Y.; Guo, X.; Yin, Y.; Fang, J. Z.; Yang, Y.; Wang, Y.; Yuan, S.; Wang, A.; Liu, B.; Huang, H.; and Ma, C. 2025. MAGREF: Masked Guidance for Any-Reference Video Generation. arXiv:2505.23742.

Gao, Y.; Gong, L.; Guo, Q.; Hou, X.; Lai, Z.; Li, F.; Li, L.; Lian, X.; Liao, C.; Liu, L.; and et al. 2025a. Seedream 3.0 Technical Report. arXiv:2504.11346.

Gao, Y.; Guo, H.; Hoang, T.; Huang, W.; Jiang, L.; Kong, F.; Li, H.; Li, J.; Li, L.; Li, X.; and et al. 2025b. Seedance 1.0: Exploring the Boundaries of Video Generation Models. arXiv:2506.09113.

He, X.; Liu, Q.; Qian, S.; Wang, X.; Hu, T.; Cao, K.; Yan, K.; and Zhang, J. 2024. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*.

He, X.; Wu, Z.; Li, X.; Kang, D.; Zhang, C.; Ye, J.; Chen, L.; Gao, X.; Zhang, H.; and Zhuang, H. 2025. MagicMan: Generative Novel View Synthesis of Humans with 3D-Aware Diffusion and Iterative Refinement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3): 3437–3445.

Hu, T.; Yu, Z.; Zhou, Z.; Liang, S.; Zhou, Y.; Lin, Q.; and Lu, Q. 2025. HunyuanCustom: A Multimodal-Driven Architecture for Customized Video Generation. arXiv:2505.04512.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21807–21818.

Kling. 2025. Multi-Reference Images to Video Generation Feature. https://app.klingai.com/cn/release-notes/2025-07-24.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Kong, Z.; Gao, F.; Zhang, Y.; Kang, Z.; Wei, X.; Cai, X.; Chen, G.; and Luo, W. 2025. Let Them Talk: Audio-Driven Multi-Person Conversational Video Generation. arXiv:2505.22647.

Li, C.; Zhang, C.; Xu, W.; Lin, J.; Xie, J.; Feng, W.; Peng, B.; Chen, C.; and Xing, W. 2025a. LatentSync: Taming Audio-Conditioned Latent Diffusion Models for Lip Sync with SyncNet Supervision. arXiv:2412.09262.

Li, H.; Xu, M.; Zhan, Y.; Mu, S.; Li, J.; Cheng, K.; Chen, Y.; Chen, T.; Ye, M.; Wang, J.; and Zhu, S. 2025b. OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7752–7762.

Lin, G.; Jiang, J.; Yang, J.; Zheng, Z.; and Liang, C. 2025. OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human Animation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, J.; Liu, G.; Liang, J.; Yuan, Z.; Liu, X.; Zheng, M.; Wu, X.; Wang, Q.; Qin, W.; Xia, M.; Wang, X.; Liu, X.; Yang, F.; Wan, P.; Zhang, D.; Gai, K.; Yang, Y.; and Ouyang, W. 2025a. Improving Video Generation with Human Feedback. arXiv:2501.13918.

Liu, L.; Ma, T.; Li, B.; Chen, Z.; Liu, J.; Li, G.; Zhou, S.; He, Q.; and Wu, X. 2025b. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *ECCV*, 38–55.

Polyak, A.; Zohar, A.; Brown, A.; Tjandra, A.; Sinha, A.; Lee, A.; Vyas, A.; Shi, B.; Ma, C.-Y.; Chuang, C.-Y.; et al. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 8748–8763.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; Mcleavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 28492–28518.

Seawead, T.; Yang, C.; Lin, Z.; Zhao, Y.; Lin, S.; Ma, Z.; Guo, H.; Chen, H.; Qi, L.; Wang, S.; and et al. 2025. Seaweed-7B: Cost-Effective Training of Video Generation Foundation Model. arXiv:2504.08685.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing Obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4).

Team, G. 2025. Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

Tian, L.; Hu, S.; Wang, Q.; Zhang, B.; and Bo, L. 2025. EMO2: End-Effector Guided Audio-Driven Avatar Video Generation. arXiv:2501.10687.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; and et al. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314.

Wang, M.; Wang, Q.; Jiang, F.; Fan, Y.; Zhang, Y.; Qi, Y.; Zhao, K.; and Xu, M. 2025a. FantasyTalking: Realistic Talking Portrait Generation via Coherent Motion Synthesis. In *Proceedings of the 33th ACM International Conference on Multimedia*.

Wang, Q.; Shi, Y.; Ou, J.; Chen, R.; Lin, K.; Wang, J.; Jiang, B.; Yang, H.; Zheng, M.; Tao, X.; Yang, F.; Wan, P.; and Zhang, D. 2025b. Koala-36M: A Large-scale Video Dataset Improving Consistency between Fine-grained Conditions and Video Content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8428–8437.

Wang, Z.; Yang, J.; Jiang, J.; Liang, C.; Lin, G.; Zheng, Z.; Yang, C.; and Lin, D. 2025c. InterActHuman: Multi-Concept Human Animation with Layout-Aligned Audio Conditions. arXiv:2506.09984.

Wei, C.; Sun, B.; Ma, H.; Hou, J.; Juefei-Xu, F.; He, Z.; Dai, X.; Zhang, L.; Li, K.; Hou, T.; Sinha, A.; Vajda, P.; and Chen, W. 2025. MoCha: Towards Movie-Grade Talking Character Synthesis. arXiv:2503.23307.

Yi, H.; Ye, T.; Shao, S.; Yang, X.; Zhao, J.; Guo, H.; Wang, T.; Yin, Q.; Xie, Z.; Zhu, L.; Li, W.; Lingelbach, M.; and Zhou, D. 2025. MagicInfinite: Generating Infinite Talking Videos with Your Words and Voice. arXiv:2503.05978.

Yuan, S.; He, X.; Deng, Y.; Ye, Y.; Huang, J.; Lin, B.; Luo, J.; and Yuan, L. 2025a. OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation. arXiv:2505.20292.

Yuan, S.; Huang, J.; He, X.; Ge, Y.; Shi, Y.; Chen, L.; Luo, J.; and Yuan, L. 2025b. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12978–12988.