LongCat-Video-Avatar Technical Report

Meituan LongCat Team

ABSTRACT

Recent advances in audio-driven human video synthesis have significantly enhanced the realism of both half-body and full-body generation. However, generating long-duration sequences remains a persistent challenge, as existing methods often suffer from error accumulation and identity drift over time. Reference image injection strategies, such as InfiniteTalk, have been proposed to address these issues, but they frequently introduce a rigid "copy-paste" effect and limit motion diversity due to conditional image leakage. Moreover, current models tend to over-rely on speech signals, resulting in unnaturally static behaviors. To overcome these limitations, we present LongCat-Video-Avatar, a unified architecture designed for super-realistic, lip-synchronized long video generation with natural dynamics and consistent identity. Building upon the powerful unified video foundation model LongCat-Video, this model inherits its multiple generation modes, including audio-text-to-video, audio-text-image-to-video, and audio-driven video continuation. We begin by analyzing the coupling between speech and human motion and propose a Disentangled Unconditional Guidance strategy that separates audio signals from motion dynamics, ensuring natural behavior even in the absence of speech. To mitigate identity degradation during long video generation, we inject a reference image to preserve human identity, supporting flexible placement by specifying its corresponding index in the RoPE. We systematically analyze the impact of reference image placement within video latents and introduce a Reference Skip Attention mechanism to balance the trade-off between guidance strength and conditional image leakage. This design effectively reduces the "copy–paste" artifact while maintaining high visual fidelity and rich motion dynamics. Additionally, to address error accumulation caused by redundant VAE decode-encode cycles in autoregressive generation, we propose a Cross-Chunk Latent Stitching strategy. Extensive evaluations demonstrate the effectiveness of our approach in generating super-realistic, long-duration human videos.

GitHub: https://github.com/meituan-longcat/LongCat-Video

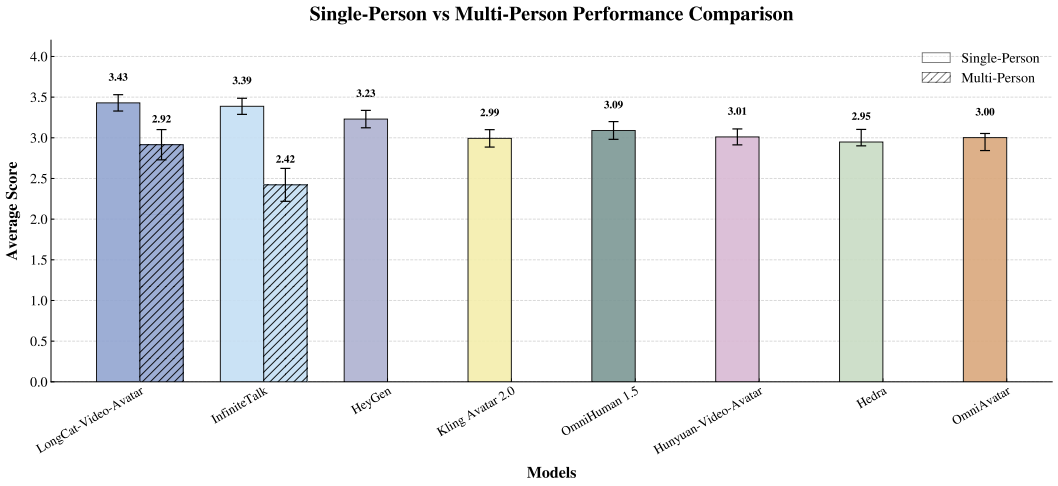

Figure 1: Human evaluation on naturalness and realism of the synthesized videos. The benchmark Zhou et al. [2025] contains more than 400 testing samples with different difficulty levels, different scenarios, and different languages for evaluating the performance of single and multiple human video generation.

Contents

1 Introduction

Audio-driven human animation aims to synthesize lifelike avatar videos in which lip motions, facial expressions, and body movements are coherently synchronized with input audio signals. Recent advances in video diffusion models have greatly accelerated progress in this field, enabling more realistic and temporally consistent animations. And audio-driven human animation has been increasingly applied in advertising, gaming, film and television production, education, and human–computer interaction.

Figure 2: Limitations of existing methods. Existing approaches commonly suffer from error accumulation over long duration and exhibit "copy–paste" artifacts, which degrade motion diversity and visual quality.

Recently, Diffusion Transformer (DiT) based audio-driven methods Meng et al. [2024], Cui et al. [2024], Wang et al. [2025a], Kong et al. [2025], Jiang et al. [2025] have achieved notable progress, particularly in enhancing lip synchronization and enabling the generation of half-body and full-body animations. However, despite these advancements, the duration of videos produced by such models remains limited to only a few seconds. When extended to longer sequences, these approaches tend to suffer from error accumulation. As shown in Fig.2 (a), the results generated by OmniAvatar Gan et al. [2025] exhibit identity drift and a noticeable decline in visual quality over time.

Figure 3: Limitations of existing methods: It suffers from rigid coupling between speech and body movements.

To address this issue, InfiniteTalk Yang et al. [2025] introduces a keyframe injection strategy that employs a ReferenceNet to insert reference frames for preserving identity, characteristic gestures, and camera trajectories. Although this approach effectively mitigates identity drift and suppresses cumulative errors, the rigid "copy–paste" behavior of the injected keyframes leads to conditional image leakage, which restricts motion richness and diversity. As illustrated in Fig.2(b), excessive reliance on reference images often causes the generated frames to overly mimic the keyframe content, resulting in repetitive and unnatural motions.

Furthermore, we observe a strong coupling between speech and body movements in many audio-driven human animation methods Wang et al. [2025a], Gan et al. [2025], Kong et al. [2025], Yang et al. [2025], which limits the naturalness of the generated results. As illustrated in Fig. 3, taking OmniAvatarGan et al. [2025] and InfiniteTalk Yang et al. [2025] as examples, when the input audio contains no human speech, they produce an almost static video, which is unrealistic. This phenomenon indicates an excessive reliance on speech signals to drive global human motion.

In light of this, we propose LongCat-Video-Avatar, a unified architecture for audio-driven human animation that supports multiple condition inputs, including audio-text-to-video (AT2V), audio-text-image-to-video (ATI2V), and audio-driven video continuation. We rethink the phenomenon where humans appear unnaturally static when the input audio does not contain speech, revealing an undesirable coupling between speech signals and global body motion. To address this,

we propose a Disentangled Unconditional Guidance strategy that effectively disentangles audio speech from motion dynamics. Moreover, to preserve identity information and avoid errors accumulating, we inject an additional reference image during long video generation. Our trained model allows flexible placement of the reference image by specifying its corresponding index in the RoPE. Through systematic analysis, we find that the placement of the reference image plays a critical role in generation quality: inserting it at the prefix or suffix of the generated chunk results in insufficient guidance and detail loss, whereas placing it in the middle induces a strong "copy–paste" effect. To balance the trade-off between inject intensity and conditional image leakage, we introduce a Reference Skip Attention mechanism that selectively incorporates reference cues while avoiding conditional image leakage. Finally, we observe that during long video generation, repetitive VAE decoding and re-encoding of motion frames introduce pixel degradation. To overcome this issue, we design a Cross-Chunk Latent Stitching strategy for training that eliminates redundant VAE operations and effectively bridges the train–test gap. Extensive quantitative and qualitative experiments across multiple public datasets demonstrate the effectiveness and robustness of our proposed framework.

In summary, our main contributions are as follows.

- We propose **LongCat-Video-Avatar**, a unified architecture for audio-driven human animation that supports multiple conditioning modalities, including audio–text-to-video, audio–text-image-to-video, and audio-driven video continuation.
- We identify and address the unnatural static posture phenomenon in silent-audio scenarios by introducing a **Disentangled Unconditional Guidance** strategy, which disentangles the strong coupling between speech signals and global body motion.
- Our trained model injects a reference image to preserve human identity and allows flexible placement of the reference image by specifying its corresponding index in the RoPE. We conduct a systematic analysis of the impact of reference image placement in long video generation and propose a Reference Skip Attention mechanism that balances the trade-off between guidance strength and conditional image leakage, effectively mitigating the "copy–paste" artifact.
- We design a novel **Cross-Chunk Latent Stitching** training strategy that eliminates redundant VAE decoding and encoding cycles, thereby reducing pixel degradation and narrowing the train–test gap in long video generation.
- Extensive experiments on multiple public datasets demonstrate that our unified framework achieves superior performance and robustness across diverse audio-driven human animation tasks.

2 Related Work

2.1 Video Generation

Video generation has advanced rapidly in recent years, driven by the availability of large-scale, high-quality video datasets and the success of diffusion models Ho et al. [2020]. Early video diffusion approaches Chen et al. [2023, 2024], Blattmann et al. [2023], Guo et al. [2023] predominantly adopt a U-Net–based architecture, extending 2D U-Nets pretrained on text-to-image tasks into 3D variants to capture temporal dynamics. These models leverage strong image priors and are subsequently fine-tuned on video datasets to generate coherent frame sequences. More recently, some works Yang et al. [2024], Kong et al. [2024], Wang et al. [2025b], Gao et al. [2025a], Team et al. [2025] have replaced the conventional U-Net backbone with a Diffusion-in-Transformers (DiT) architecture, achieving substantial improvements in video generation quality and temporal consistency. By modeling video as a sequence of spatiotemporal patches, DiT-based methods offer enhanced scalability and flexibility, enabling efficient generation of high-resolution videos with variable durations and aspect ratios.

In parallel, a growing number of studies have employed video diffusion models to advance world-model research He et al. [2025], Yu et al. [2025], positioning video generation as a critical pathway toward building predictive models of the physical world. Efficient long-horizon video generation is especially important in this context. LongCat-Video Team et al. [2025] demonstrates strong performance in video continuation and long-range temporal modeling. To fully exploit the priors of video diffusion models for physical simulation and extended video generation, we adopt LongCat-Video as our backbone.

2.2 Audio-driven Human Animation

Early audio-driven human animation methods Guan et al. [2023], Zhang et al. [2023], Cheng et al. [2022], Pang et al. [2023], Yin et al. [2022], Gong et al. [2023], Wang et al. [2024] typically rely on two-stage pipelines. These approaches first employ an audio-to-motion model to convert speech signals into intermediate representations such as 3DMM or FLAME parameters, and then use motion-to-video rendering techniques, such as GANs, to synthesize dynamic portrait

animations. More recently, end-to-end approaches Tian et al. [2024], Wei et al. [2024], Xu et al. [2024], Chen et al. [2025a], Cui et al. [2024], Ji et al. [2024], Li et al. [2024], Jiang et al. [2024], Gao et al. [2025b] have emerged, utilizing a single diffusion model to directly integrate audio cues with facial dynamics. To overcome the limitation of generating only head movements, subsequent works Lin et al., Tian et al. [2025], Meng et al. [2024], Jiang et al. [2025], Gan et al. [2025], Wang et al. [2025a] further incorporate body motion and human–object interactions. Another line of research Wei et al. [2025], Kong et al. [2025], Chen et al. [2025b], Huang et al. [2025], Ma et al. [2025], Wang et al. [2025c], Zhong et al. [2025] explores multi-identity video generation by binding multiple speakers to multiple audio streams, enabling conversational multi-human animation.

Although these approaches achieve impressive results for short clips, they struggle when extended to long sequences due to significant error accumulation. InfiniteTalk Yang et al. [2025] attempts to mitigate this issue by introducing a sparse-frame video dubbing strategy that injects keyframes into each generated video chunk through ReferenceNet. However, its "copy-paste" constraint often leads to repetitive and unnatural motions. StableAvatar Tu et al. [2025] proposes a time-step-aware audio adapter to reduce error accumulation, but its ability to maintain accurate lip–audio synchronization remains limited.

3 Method

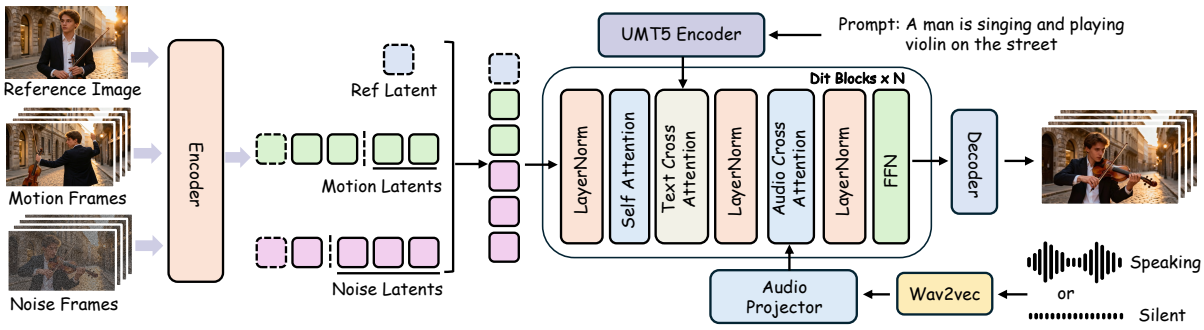

Figure 4: The overall pipeline of LongCat-Video-Avatar.

The overall pipeline of our framework is illustrated in Fig. 4. Our unified architecture supports multiple audio-driven human animation tasks, including audio–text-to-video, audio–text-image-to-video, and audio-driven video continuation. Section 3.1 outlines the network design of the proposed framework. Section 3.2 introduces the **Disentangled Unconditional Guidance** to effectively decouple speech signals from motion dynamics for natural behavior. Section 3.3 investigates the effect of reference image placement in long video generation and introduces the **Reference Skip Attention** mechanism to mitigate the "copy–paste" artifact. Subsequently, Section 3.4 presents the **Cross-Chunk Latent Stitching** training strategy to reduce error accumulation and bridge the train–test gap. Finally, Section 3.5 describes the training strategy for multi-person animation.

3.1 Model Architecture

In this work, we adopt a DiT-based video diffusion model as our foundational architecture. The model is built upon a 3D Variational Autoencoder (VAE), and each DiT block comprises 3D self-attention, text cross-attention, and a Feed-Forward Network (FFN). Text embeddings are encoded using a UMT5 encoder, while 3D Rotary Position Embeddings (RoPE) are applied to the visual tokens to encode spatiotemporal positional information.

The network accepts three types of latent sequences as input: a reference latent, motion latents, and noise latents. As the video foundation model utilizes a unified framework that supports multiple tasks, the same architecture is employed with different input configurations. For text-to-video generation, only noise latents are provided. For text-image-to-video generation, the reference latent is temporally concatenated with the noise latents. For video continuation, the motion latents are temporally concatenated with the noise latents and then fed into the model as additional conditioning signals.

However, the original video foundation model does not include a native mechanism for audio conditioning. To enable audio-driven generation, we modify each DiT block by inserting an additional audio cross-attention layer after the text cross-attention module, allowing audio cues to be integrated into the visual generation process. Yet, directly adding this layer causes training instability and prevents the model from effectively aligning audio signals with corresponding mouth movements. To mitigate this issue, we introduce an Adaptive Layer Normalization (adaLN) module before

each audio cross-attention layer. This module functions as a gating mechanism that preserves the pre-trained visual priors while progressively incorporating audio control, thereby stabilizing optimization and facilitating the learning of accurate audio-to-lip motion mappings.

We employ the widely used audio feature extractor Wav2Vec Baevski et al. [2020] to obtain frame-level audio embeddings. Because each audio frame is influenced by its surrounding context, we follow Tian et al. [2024] and concatenate neighboring audio embeddings within a temporal window around the current frame. During training, we adopt a mixed frame rate strategy using both 16 FPS and 25 FPS. We observe that applying the same window size across different frame rates leads to suboptimal alignment between audio and mouth movements. Therefore, for 16 FPS training, we proportionally reduce the window size to improve the temporal correspondence between audio cues and lip shapes. Since the video VAE applies a 4 rate temporal downsampling when converting video from pixel space into the latent space, a corresponding temporal compression is required for audio input to maintain alignment between audio and visual representations. Following Kong et al. [2025], we introduce an audio projector to temporally compress the audio embeddings before injecting them into the audio cross-attention.

During training, in the t timestep, the key objective of the diffusion model is to predict the noise ϵ utilizing a denoising DiT ϵ_θ conditioned on the text input c_{text}, audio input c_{audio} and concatenated latent sequence z_t defined as:

$$\epsilon = \epsilon_\theta(z_t, t, c_{text}, c_{audio}) \tag{1}$$

3.2 Disentangled Unconditional Guidance

For audio-driven human animation, most existing methods are trained using a single audio–image-to-video (AI2V) task, as illustrated in Fig. 5(a). While this setting allows the model to learn reasonable audio lip synchronization, it inadvertently induces a strong coupling between speech and body movements. Consequently, when the input audio contains no human speech, the model tends to synthesize an almost static video, which is unrealistic.

To alleviate this issue, Kong et al. [2025], Yang et al. [2025] introduce a multi-task training paradigm, shown in Fig. 5(b). In addition to the AI2V task, an extra image-to-video (I2V) task is incorporated. Since the I2V task lacks audio input, a zero audio embedding is fed into the audio cross-attention layers. Although this multi-task design enhances instruction-following capability, it still fails to resolve the strong coupling problem.

Motivated by this, we further extend the multi-task framework by not only training audio cross-attention but also injecting trainable parameters into the self-attention layers using LoRA, as shown in Fig. 5(c). However, this attempt still does not sufficiently mitigate the undesired correlation between speech signals and global body motion.

Upon revisiting the problem, we identify the root cause: the use of zero audio embeddings during multi-task training causes the model to conflate I2V tasks with silent-audio conditions during inference. This ambiguity makes the model incorrectly interpret silent audio as an unconditional input, thus reinforcing the unnatural static behavior.

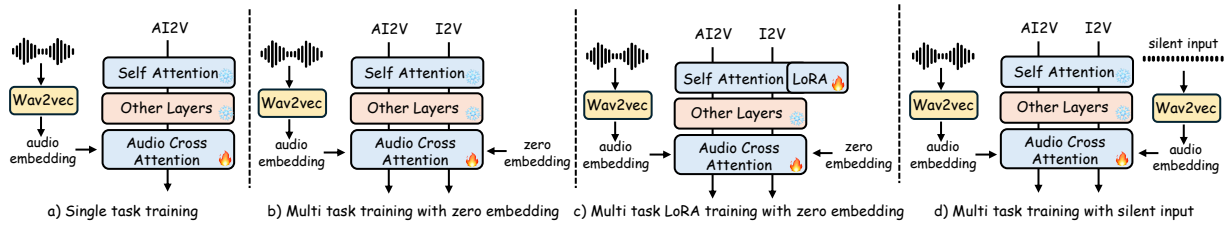

Figure 5: Comparison of different training strategies.

During inference, classifier-free guidance (CFG) Ho and Salimans [2022] is commonly employed to balance conditional fidelity and generative diversity, and is formulated as:

$$\begin{aligned}\epsilon = &\epsilon_\theta(x_t, t, c_{un\text{-}text}, c_{un\text{-}audio}) \\ &+ w_{text}\big[\epsilon_\theta(x_t, t, c_{text}, c_{audio}) - \epsilon_\theta(x_t, t, c_{un\text{-}text}, c_{audio})\big] \\ &+ w_{audio}\big[\epsilon_\theta(x_t, t, c_{un\text{-}text}, c_{audio}) - \epsilon_\theta(x_t, t, c_{un\text{-}text}, c_{un\text{-}audio})\big]\end{aligned} \tag{2}$$

where w_{text} and w_{audio} denote the guidance scales for text and audio conditions, respectively. By omitting the text condition for simplicity, this formula reduces to:

$$\begin{aligned}\epsilon = &\epsilon_\theta(x_t, t, c_{un\text{-}audio}) \\ &+ w_{audio}\big[\epsilon_\theta(x_t, t, c_{audio}) - \epsilon_\theta(x_t, t, c_{un\text{-}audio})\big]\end{aligned} \tag{3}$$

where $c_{un\text{-}audio}$ typically represents a zero embedding (null condition).

Recent works Kong et al. [2025], Yang et al. [2025] employ a multi-task hybrid training strategy that combines audio-image-to-video and image-to-video tasks. In the I2V setting, due to the absence of audio signals, the audio input is conventionally set to $c_{un\text{-}audio}$ (i.e., a zero embedding). However, we observe that simply replacing the missing audio with a zero embedding creates an ambiguity: the model fails to distinguish between explicit I2V inputs and truly unconditioned inputs used for CFG.

In standard CFG, the unconditional branch is expected to serve as a generic baseline. However, since the model is trained to generate high-quality videos from zero embeddings (via the I2V task), the unconditional predictions in CFG become overly coherent and similar to the conditional predictions. This similarity diminishes the discrepancy between the two branches, effectively weakening the guidance signal and leading to suboptimal generation quality.

To address this, we propose **Disentangled Unconditional Guidance**. The core idea is to explicitly differentiate the representation of "silent audio" (for I2V) from the "null audio" (for unconditional generation). Specifically, during I2V training, we replace the zero embedding with an encoded representation of silence:

$$c_{\text{silent}} = \text{Adapter}(\text{Wav2Vec}(audio_{\text{silent}})) \tag{4}$$

where $audio_{\text{silent}}$ denotes a silent audio waveform. Since the embedding derived from the Wav2Vec encoder and adapter is distinct from the zero embedding, this design ensures that the model can effectively distinguish I2V inputs from unconditioned inputs. Consequently, the unconditional branch remains a proper baseline, preserving the effectiveness of CFG and improving overall generation quality.

3.3 Reference Skip Attention

The proposed method supports both audio-text-to-video generation and audio-image-to-video generation. However, these two modes typically produce videos of only a few seconds, which is still insufficient for real-world applications. To enable long video synthesis, we introduce an audio-driven video continuation mechanism.

Specifically, during video continuation, the final latents of the previously generated video frames, denoted as context latents $z_{\text{context}} \in \mathbb{R}^{B\times N_{\text{context}}\times H\times W\times C}$, are temporally concatenated with the latent noise $z_{\text{noise}} \in \mathbb{R}^{B\times N_{\text{noise}}\times H\times W\times C}$. The concatenated sequence serves as the input latent for the diffusion model, where N_{context} and N_{noise} represent the lengths of the context and noise latent sequences, respectively. Formally, the input can be expressed as

$$z = \text{concat}([z_{\text{context}}, z_{\text{noise}}]). \tag{5}$$

Although this strategy enables temporally coherent video continuation, error accumulation and identity degradation inevitably emerge as the generated video length increases. To address this issue, following Yang et al. [2025], we incorporate an additional reference latent $z_{\text{ref}} \in \mathbb{R}^{B\times 1\times H\times W\times C}$ to preserve key visual attributes, such as identity, background appearance, and stylistic details, which effectively suppresses long term drift. The complete latent input is thus defined as

$$z = \text{concat}([z_{\text{ref}}, z_{\text{context}}, z_{\text{noise}}]), \tag{6}$$

with

$$z \in \mathbb{R}^{B\times(1+N_{\text{context}}+N_{\text{noise}})\times H\times W\times C}. \tag{7}$$

During training, we randomly sample $((N_{\text{context}} + N_{\text{noise}} - 1) \times 4 + 1)$ consecutive frames from each video clip. The first $((N_{\text{context}} - 1) \times 4 + 1)$ frames serve as context frames, while the remaining $(N_{\text{noise}} \times 4)$ frames serve as noise frames. We additionally sample one extra frame from the same video clip as the reference image. To encode the spatiotemporal positions of the input latents, we employ rotary position embeddings (RoPE). Since the context and noise frames correspond to temporally consecutive frames in pixel space, we assign monotonically increasing temporal indices to their latent frames. The temporal position of the reference latent is computed based on its relative distance to the first context frame in pixel space. This distance is converted into latent-space coordinates by dividing it by the VAE's temporal compression rate, and the resulting value is used as the temporal position of the reference latent in RoPE during training.

During inference, the position of the reference image can be flexibly specified by assigning the corresponding index in RoPE. We therefore explore three distinct placements of the reference image: inserting it at the prefix, middle, or suffix of each generated video chunk. These different placements yield noticeably different results.

Placing the reference image at the prefix or suffix results in insufficient inject intensity, leading to detail loss. In contrast, although placing the reference image in the middle of the chunk can avoid this insufficient guidance problem, it introduces another "copy–paste" issue similar to InfiniteTalk Yang et al. [2025]. Specifically, the generated video

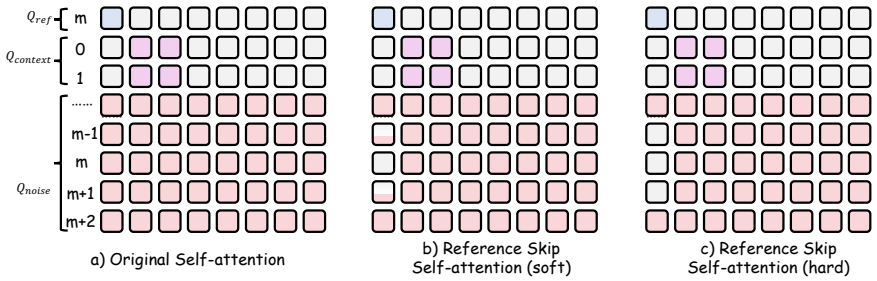

Figure 6: Illustration of the Reference Skip Attention mechanism.

chunk tends to replicate the pose of the reference image, producing repetitive and unnatural motions due to conditional image leakage.

To alleviate this "copy–paste" effect and balance the trade-off between injection intensity and conditional image leakage, we propose a **Reference Skip Attention** mechanism that selectively incorporates reference cues while avoiding excessive reliance on the reference latent.

As shown in Fig. 6(a), the original self-attention is computed as follows:

$$x_{\text{ref}} = \text{Attention}(Q_{\text{ref}}, K_{\text{ref}}, V_{\text{ref}}), \tag{8}$$

$$x_{\text{context}} = \text{Attention}(Q_{\text{context}}, K_{\text{context}}, V_{\text{context}}), \tag{9}$$

$$x_{\text{noise}} = \text{Attention}\Big(Q_{\text{noise}}, [K_{\text{ref}}, K_{\text{context}}, K_{\text{noise}}], [V_{\text{ref}}, V_{\text{context}}, V_{\text{noise}}]\Big), \tag{10}$$

where Q_{ref}, K_{ref}, and V_{ref} denote the query, key, and value for reference latent; Q_{context}, K_{context}, and V_{context} denote the query, key, and value for context latents; and Q_{noise}, K_{noise}, and V_{noise} denote the query, key, and value for noise latents.

During the attention computation for x_{noise}, the queries are computed from noise latents, while the keys and values come from the concatenation of reference latent, context latents, and noise latents. When we place the reference latent in the middle of the chunk by assigning it a temporal index m in RoPE, the noisy latent sequence z_{noise} also contains a latent with the same positional index m, as illustrated in Fig. 6. We denote this latent as the *anchor latent*. Due to the shared positional index, the anchor latent and the reference latent occupy the same temporal position within the video latents, resulting in excessively strong identity injection. This over-conditioning manifests as a "copy–paste" effect, characterized by repetitive motions that closely mimic the reference image.

To mitigate excessive identity injection, we propose Reference Skip Attention, which excludes the reference latent from the attention computation of the anchor latent and its neighboring latents, while preserving the original attention pattern for all remaining latents. Specifically, as shown in Fig. 6(c), for the latents at indices $\{m-1, m, m+1\}$ in the noisy latent sequence, we remove the reference latent from the key value set during attention calculation; these latents therefore attend only to the context and noisy latents. All other noise latents outside $\{m-1, m, m+1\}$ still attend to the reference latent following the original formulation. This selective masking enables reference cues to be incorporated in a selective manner, striking a balance between identity guidance strength and conditional image leakage. It not only alleviates the repetitive motion artifacts caused by the "copy–paste" effect, but also avoids the error accumulation that would occur if the reference latent were entirely removed.

For Reference Skip Attention, we explore two implementation variants: a soft-mask strategy, as shown in Fig. 6(b), and the hard-mask strategy described above and shown in Fig. 6(c). In the soft-mask version, only the latents at index m are fully masked, while latents at adjacent positions $\{m-1, m+1\}$ are not removed but instead downweighted by reducing their corresponding values in the self-attention map. We find that this soft-mask approach achieves performance comparable to the hard-mask strategy. However, for simplicity and implementation efficiency, we adopt the hard-mask formulation as our final design. Our Reference Skip Attention strategy is training-free and is applied only during inference. In future work, we plan to investigate integrating this strategy into both the training and inference stages.

3.4 Cross-Chunk Latent Stitching

For long video generation, existing methods typically rely on multi-stage pipelines that combine AT2V or ATI2V generation with audio-driven video continuation. The initial video chunk is generally produced using AT2V or ATI2V. As illustrated in Fig. 7(a), take ATI2V as an example: the first chunk is generated by predicting noise conditioned on a

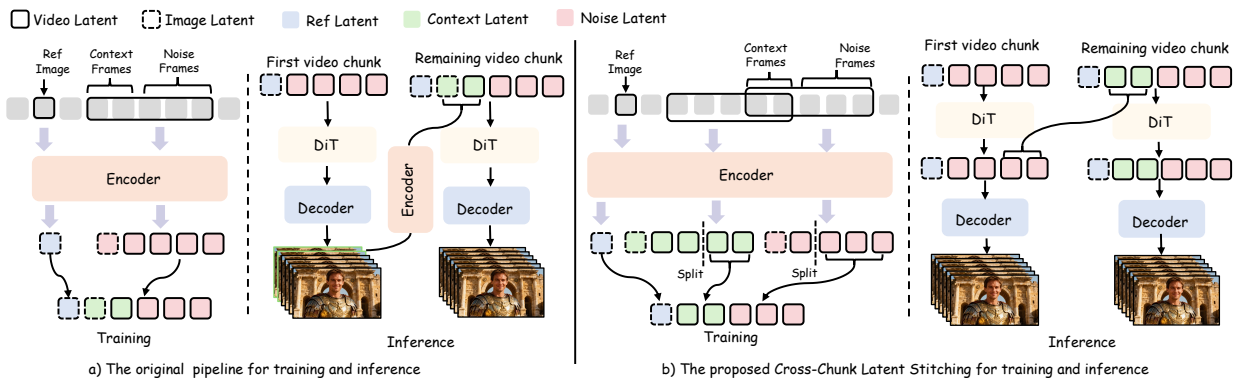

Figure 7: Illustration of long video processing during training and inference. (a) The original pipeline for long video training and inference. (b) The improved process utilizes the proposed Cross-Chunk Latent Stitching strategy.

reference image, followed by VAE decoding to obtain pixel-space video frames. Subsequent chunks are synthesized via audio-driven video continuation. Specifically, the last several frames of the preceding chunk are extracted as context frames. These frames are passed through the VAE encoder to obtain context latents, which are then concatenated with the reference latent and noise latents to form the input sequence z. This sequence serves as the input for synthesizing the current video chunk, and the process is repeated for all subsequent chunks.

However, during the generation of each continuation, obtaining the context latents requires first decoding the previously predicted chunk via the VAE and then re-encoding the selected context frames. This repetitive VAE decode-encode cycle introduces two major issues. First, the repeated encoding and decoding can result in information loss and error accumulation within each chunk. Second, it significantly reduces inference efficiency. Moreover, the context latents $z_{context}$ used during training follow the format:

$$\{\, l_I + l_V \times (N_{\text{context}} - 1) \,\}, \tag{11}$$

where l_I and l_V denote the image latent and video latent, respectively, and N_{context} is the number of context latents. However, directly reusing the last several video latents from the previously generated chunk without VAE decoding and re-encoding changes the context format to:

$$\{\, l_V \times N_{\text{context}} \,\}, \tag{12}$$

thereby replacing the first image latent with a video latent. This discrepancy introduces a train–test gap, leading to error accumulation and degraded visual quality during long video generation.

To address this challenge, we introduce a **Cross-Chunk Latent Stitching** strategy that mitigates error accumulation, enhances inference efficiency and effectively bridges the train–test gap. As illustrated in Fig. 7(b), during training, we sample two consecutive video chunks with a temporal overlap, where the number of overlapping frames matches the context frames required for continuation. Each chunk is independently encoded by the VAE to obtain its corresponding video latents. We then extract the context latents from the earlier video latent and replace the corresponding context latents in the later chunk with these extracted features. This ensures that the model learns video continuation under conditions consistent with inference, eliminating the need for repetitive VAE decode encode cycles. Finally, we concatenate the reference latent along the temporal dimension to construct the full input sequence for training.

3.5 Multi-Person Animation

To enable multi-person animation, we extend our model with L-RoPE Kong et al. [2025] to associate multiple audio streams with their corresponding individuals within the audio cross-attention layers. In multi-person scenarios, each person is provided with an independent audio stream. We first partition the video latents into person 1, person 2, and background regions by computing the similarity between the reference latent and each video latent frame using the self-attention map. In audio cross-attention, video tokens serve as queries, while keys and values are derived from all audio streams. Then, we assign the same RoPE label to the audio and video tokens belonging to the same person. This labeling enlarges the activation of the corresponding region within the audio cross-attention map, thereby enabling accurate audio–visual binding for multiple individuals.

4 Evaluation

4.1 Settings

Datasets. For training data, we use our collected video dataset containing 2K hours of videos, covering both head and full body movements. For multi-person generation, we additionally curate 200K video clips featuring diverse events and rich human object and human environment interactions. All data used in our experiments are collected from publicly available online sources. The data collection and curation process strictly follows the best practices established in prior works Wang et al. [2025d], Nan et al. [2024], Ju et al. [2024], ensuring consistency with widely accepted community standards. Moreover, all data sources are released under the CC BY 4.0 International license.

To evaluate the effectiveness of our method, we employ both talking head and talking body datasets. For talking-head evaluation, we use two publicly available benchmarks: HDTF Zhang et al. [2021][1] and CelebV-HQ Zhu et al. [2022][2]. For talking-body evaluation, we adopt the EMTD dataset Meng et al. [2024][3]. In addition, we use EvalTalker Zhou et al. [2025], a dataset with varying difficulty levels that provides more than 400 samples.

Dataset	Model	Metrics				
		FID↓	FVD ↓	Sync-C↑	Sync-D↓	CSIM ↑
HDTF	Hallo3	58.95	217.88	7.15	8.59	0.686
	OmniAvatar	60.36	202.88	7.09	8.52	0.685
	MultiTalk	64.55	201.90	9.07	6.86	0.631
	InfiniteTalk	51.76	**196.28**	9.05	6.79	0.730
	Ours	**51.63**	206.46	**9.23**	**6.51**	**0.754**
CelebV-HQ	Hallo3	73.11	379.45	5.62	9.11	0.529
	OmniAvatar	72.32	367.59	5.90	8.64	0.590
	MultiTalk	74.31	**352.94**	7.29	7.66	0.579
	InfiniteTalk	**65.37**	420.06	7.36	7.42	**0.653**
	Ours	67.60	362.30	**7.70**	**7.06**	0.628
EMTD	Hallo3	110.92	803.98	5.62	9.58	0.460
	OmniAvatar	116.14	810.82	6.97	8.49	0.565
	MultiTalk	112.83	897.02	8.13	7.51	0.581
	InfiniteTalk	113.43	1000.73	8.51	7.21	0.605
	Ours	**65.05**	**433.36**	**8.71**	**6.93**	**0.672**

Table 1: Quantitative comparisons between our method and other competitive models on HDTF, CelebV-HQ, and EMTD datasets.

Evaluation Metrics. Following Kong et al. [2025], we adopt widely used evaluation metrics to assess our method. Frechet Inception Distance (FID) Heusel et al. [2017] and Fréchet Video Distance (FVD) Unterthiner et al. [2019] are employed to evaluate the overall visual quality of the generated results. Cosine Similarity (CSIM) measures the identity preservation ability. To quantify audio–lip synchronization, we report Sync-C and Sync-D scores Chung and Zisserman [2017]. To evaluate long-video generation performance, we extract 16 frames from the tail of each generated video when computing FID and FVD. For CSIM, we uniformly sample 16 frames from the final 2 seconds of the generated video to assess identity preservation over long temporal horizons.

[1] https://github.com/MRzzm/HDTF
[2] https://github.com/CelebV-HQ/CelebV-HQ
[3] https://github.com/antgroup/echomimic_v2/tree/main/EMTD_dataset

	FID↓	FVD ↓	Sync-C↑	Sync-D↓	CSIM ↑
HunyuanVideo-Avatar*	86.52	620.60	5.08	9.04	0.598
InfiniteTalk	88.37	627.35	6.06	8.36	0.629
OmniAvatar	89.41	641.83	6.24	8.21	0.546
Kling Avatar 2.0	86.21	617.26	4.55	9.58	**0.670**
Hedra	87.51	**582.45**	5.09	8.99	0.530
HeyGen	87.78	639.13	6.13	8.61	0.657
OmniHuman 1.5	92.35	648.14	5.92	8.55	0.630
Ours	**86.12**	611.38	**6.28**	**8.06**	0.620

Table 2: Quantitative comparisons between our method and other competitive models on the EvalTalker dataset Zhou et al. [2025]. The videos generated by HunyuanVideo-Avatar (*) are only 5 seconds in length, whereas those generated by other methods range from 5 to 90 seconds.

Implementation Details. We adopt Longcat-Video as the foundational video diffusion model for all experiments. The model is optimized using AdamW with a constant learning rate of $1e-5$ and a warm-up schedule. During training, we fine-tune the audio cross-attention layers, self-attention layers, text cross-attention, FFN, and the audio adapter, while keeping all remaining network parameters frozen.

4.2 Comparisons with Competing Methods

Quantitative Evaluation. We compare Longcat-Video-Avatar with several state-of-the-art audio-driven human animation methods that support long video generation, including Hallo3 Cui et al. [2024], OmniAvatar Gan et al. [2025], MultiTalk Kong et al. [2025], and InfiniteTalk Yang et al. [2025], on the HDTF, CelebV-HQ, and EMTD datasets. The quantitative results are reported in Table 1. Our method achieves substantially better lip-synchronization performance than all competing approaches, demonstrating its superior ability to align audio and visual modalities. In terms of visual quality, our method outperforms competing approaches across most evaluation metrics, demonstrating its strong visual generation capability.

In addition, we evaluate our method on the EvalTalker dataset against both open-source methods, including HunyuanVideo-Avatar Chen et al. [2025b], InfiniteTalk Yang et al. [2025], and OmniAvatar Gan et al. [2025], as well as commercial methods such as Kling Avatar 2.0 Kling [2025], Hedra Hedra [2025], HeyGen HeyGen [2025], and OmniHuman1.5 Jiang et al. [2025]. The quantitative comparisons are summarized in Table 2. Our approach outperforms all compared methods in terms of audio lip synchronization, while achieving competitive performance in visual quality.

Qualitative Evaluation. To demonstrate the visual effectiveness of our method, we conduct a comprehensive comparison against several competitive approaches, as shown in Fig. 8. Most existing methods produce satisfactory results when generating short video clips. However, as the length of the input audio increases, these methods suffer from severe error accumulation. This manifests as a noticeable degradation in visual quality and substantial identity drift throughout the generated sequence. Although InfiniteTalk Yang et al. [2025] mitigates long-term error accumulation, it often exhibits a "copy–paste" artifact, where the generated frames repeatedly replicate the pose or expression of the reference image, resulting in limited motion diversity. In contrast, our method avoids both error accumulation and "copy–paste" issues, consistently maintaining visual fidelity and identity over long durations, thereby demonstrating its superior performance.

We additionally provide a comparison with InfiniteTalk Yang et al. [2025] regarding identity preservation in long video generation, as shown in Fig.9. InfiniteTalk exhibits noticeable identity degradation over time, primarily reflected in the gradual disappearance of fine-grained details such as wrinkles and moles. In contrast, our method effectively preserves these identity-specific features throughout long sequences, demonstrating superior identity retention compared with InfiniteTalk.

Multi-human Animation Leveraging the L-RoPE Kong et al. [2025] strategy enables our method to support multi-human animation. The qualitative results for multi-person conversational video generation are presented in Fig. 10.

Figure 8: Visual comparison with other competing methods.

Figure 9: Visual comparison with InfiniteTalk.

Long Video Generation The proposed method demonstrates strong capability in long video generation. As shown in Fig. 11, our approach produces high-quality results for audio-driven video continuation, successfully generating videos exceeding 5 minutes in duration without noticeable error accumulation. These results highlight the robustness and effectiveness of our method for long-horizon video synthesis.

4.3 Ablation Study

To assess the effectiveness of various components within our method, we conduct an ablation study encompassing the following elements: Reference Skip Attention, Disentangled Unconditional Guidance, and Cross-Chunk Latent Stitching.

Figure 10: Visual results of multi-human animation.

Figure 11: Visual results of audio-driven video continuation exceeding 5 minutes in duration.

Reference Skip Attention. In our proposed framework, the position of the reference image can be flexibly specified. We further observe that the placement of the reference image plays a crucial role in determining the overall generation quality. As illustrated in Fig. 12, positioning the reference image at the prefix or suffix of the generated video chunk leads to information loss. For instance, text or patterns on clothing may become distorted or lose fine-grained details, resulting in insufficient identity guidance. Placing the reference image in the middle of the video chunk alleviates this information loss; however, it introduces a severe "copy–paste" problem, where the generated frames tend to replicate the exact pose or expression of the reference image rather than producing natural motion. By incorporating the proposed Reference Skip Attention (RSA), we address both issues simultaneously. When the reference image is placed in the

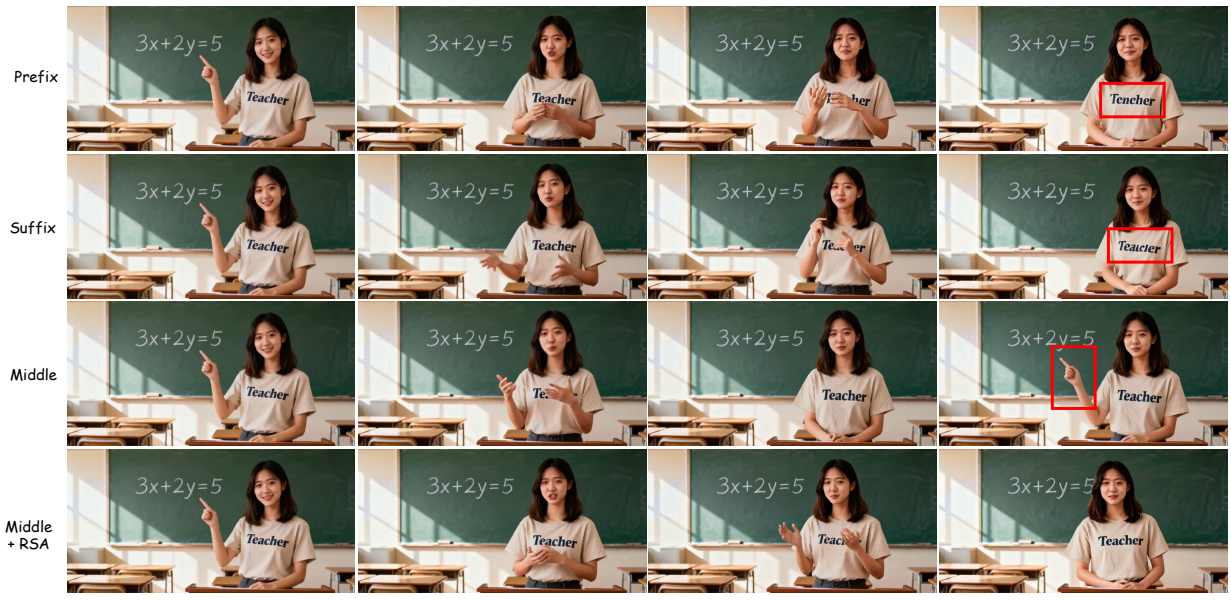

Figure 12: Qualitative ablation study of the position of the reference image and Reference Skip Attention.

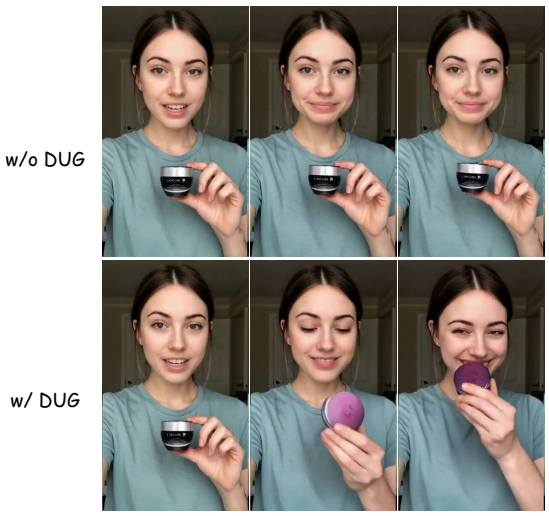

Figure 13: Qualitative ablation study of Disentangled Unconditional Guidance.

Figure 14: Qualitative ablation study of Cross-Chunk Latent Stitching(CLS).

middle, our mechanism preserves identity cues without causing motion imitation, effectively preventing information loss while eliminating the “copy–paste" artifact. This demonstrates the effectiveness of the proposed method.

Disentangled Unconditional Guidance. Next, we investigate the effectiveness of Disentangled Unconditional Guidance (DUG). We train two models, one with DUG and one without DUG, and compare their performance under silent audio input, as illustrated in Fig. 13. Without DUG, the model produces almost static character motions when given silent audio, revealing an excessively strong coupling between speech and body movement, which leads to unnatural behavior. In contrast, the model trained with DUG generates more natural motions and subtle reactions even under silent audio. This demonstrates that DUG successfully decouples speech content from general motion dynamics and improves the realism of generated videos.

Cross-Chunk Latent Stitching. Finally, we evaluate the effectiveness of the proposed Cross-Chunk Latent Stitching (CLS) strategy. We compare visual results under three settings: (1) training and inference without CLS, (2) training without CLS but applying CLS only during inference, i.e., directly using the last few denoised latents of the previous chunk as the context latents for the next chunk without repeated VAE decoding and re-encoding (denoted as Train w/o CLS, Inference w/ CLS), and (3) our full method with CLS applied during both training and inference. The comparison results are presented in Fig. 14.

The Train w/o CLS, Inference w/ CLS setting exhibits the most severe error accumulation, causing rapid degradation in visual quality. Training and inference entirely without CLS also leads to noticeable error accumulation over time. In

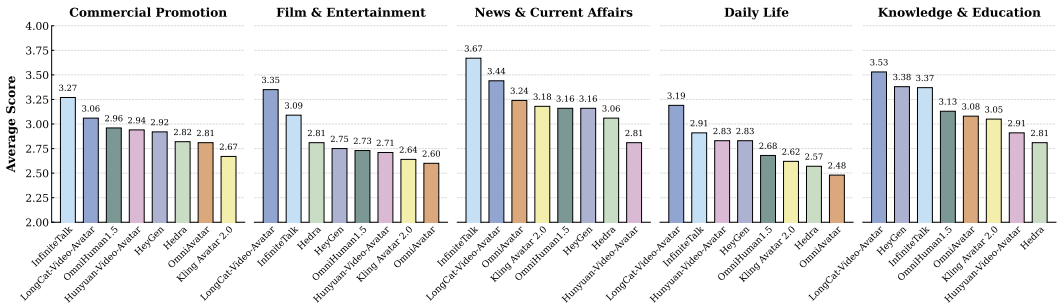

Figure 15: Performance comparison in different application scenarios.

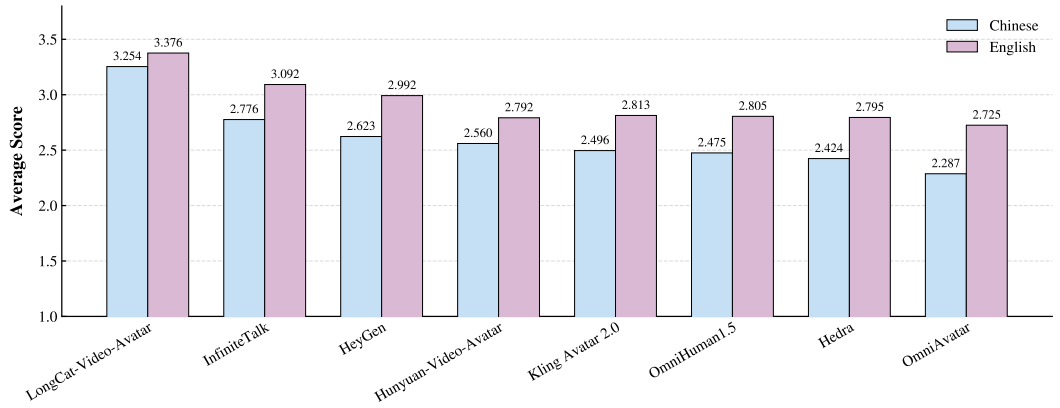

Figure 16: Performance comparison in different languages.

contrast, our full CLS strategy effectively stabilizes long-sequence generation and preserves visual fidelity even for extended videos, demonstrating the importance and effectiveness of Cross-Chunk Latent Stitching.

4.4 User Study

It is challenging to design an objective metric that accurately measures the naturalness and realism of generated videos. Therefore, we conduct a user study to directly evaluate these perceptual qualities. We compare our method against several competitive methods, including InfiniteTalk Yang et al. [2025], HeyGen HeyGen [2025], Kling Avatar 2.0Kling

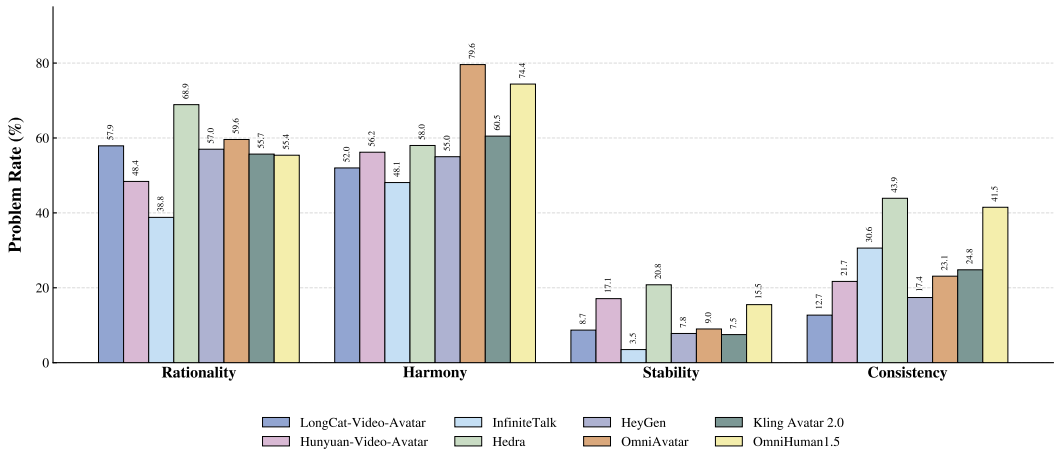

Figure 17: Statistical analysis of generation artifacts.

[2025], OmniHuman1.5 Jiang et al. [2025], HunyuanVideo-Avatar Chen et al. [2025b], Hedra Hedra [2025], and OmniAvatar Gan et al. [2025], using the public benchmark EvalTalker Zhou et al. [2025]. InfiniteTalk, HunyuanVideo-Avatar, and OmniAvatar are open-sourced models. HeyGen, OmniHuman1.5, and Hedra are commercial models that we use through their APIs or websites. For OmniHuman1.5, we use the master mode. A total of 492 participants took part in the evaluation, with each video rated independently by three users. To ensure diverse participation and avoid user fatigue, each participant was limited to evaluating at most three videos. The number of evaluators per model is kept approximately consistent.

Participants rated each video on a scale from 1 to 5, where higher scores indicate greater naturalness and realism. We evaluate audio-driven human animation methods from multiple perspectives, including scene type, language, and failure characteristics. In addition, we analyze the frequency of common generation issues along four dimensions: rationality, consistency, harmony, and stability.

We first categorize the test samples into five scene types based on their semantic characteristics: Commercial Promotion, Film & Entertainment, News & Current Affairs, Daily Life, and Knowledge & Education. We then evaluate model performance separately within each category. As shown in Fig. 15, our method achieves superior performance in Film & Entertainment, Daily Life, and Knowledge & Education scenarios, demonstrating strong generalization across diverse application settings.

We further compare different methods across multiple languages, including Chinese and English. The results are presented in Fig. 16. Our method consistently outperforms all competing approaches in both languages, indicating robust cross-lingual performance and accurate audio–lip synchronization.

We identify several common failure modes in generated videos, including deficiencies in rationality, consistency, harmony, and stability, and compute the proportion of samples exhibiting each issue. Specifically, rationality refers to the logical coherence and naturalness of human motions and interactions in multi-subject scenarios. Harmony describes smooth synchronization and coordinated relationships among multiple subjects and their surrounding environment. Stability captures visual steadiness, characterized by the absence of flickering artifacts or abrupt temporal changes throughout the video sequence. Consistency evaluates identity preservation and temporal continuity across frames, as well as coherence of background and environmental elements. The statistical results are presented in Fig. 17. While most existing methods struggle with rationality and inter-subject coordination, our approach demonstrates strong performance in terms of consistency, with rationality remaining the primary limitation.

Finally, we report the average scores for all methods in the single-person setting. For the multi-person scenario, we compare against InfiniteTalk, the only existing method that supports multi-person audio-driven generation. As shown in Fig. 1, our method consistently outperforms the baselines in both settings, demonstrating superior naturalness and realism in the generated videos.

5 Conclusion

In this work, we present LongCat-Video-Avatar, a unified framework for generating realistic, long-duration, audio-driven human animations. The proposed method addresses key challenges in existing approaches—including identity drift, motion rigidity, and unnatural silence handling. Our contributions are threefold: (1) a Reference Skip Attention mechanism that balances identity preservation and motion diversity; (2) Disentangled Unconditional Guidance, which maintains natural human expressiveness regardless of audio activity; and (3) Cross-Chunk Latent Stitching that bridges the train-test gap by eliminating redundant VAE decode-encode cycles. Extensive experiments on multiple benchmarks demonstrate that our framework consistently produces higher-quality and more stable long-form videos, setting a new state-of-the-art for audio-driven avatar animation.

6 Contributors and Acknowledgments

All people are cataloged alphabetically by last name. (†) indicates the project leader and (‡) indicates the sponsors.

Contributors

Xunliang Cai‡	Meng Cheng	Feng Gao	Zhe Kong	Shijun Liang
Siyu Ren	Xiaoming Wei‡	Tianyu Yang	Yong Zhang†	Ziyi Zhao

Acknowledgments

Fengjiao Chen	Tianye Dai	Sitao Fu	Anqing He	Xuanhua He
Qilong Huang	Zhuoliang Kang	Hongyu Li	Shengxi Li	Jiajun Liu
Hao Lu	Wenhan Luo	Liya Ma	Xin Pan	Chao Wang
Shushi Wang	Rixu Xie	Xiaoming Xu	Tong Zhang	

References

Yingjie Zhou, Xilei Zhu, Siyu Ren, Ziyi Zhao, Ziwen Wang, Farong Wen, Yu Zhou, Jiezhang Cao, Xiongkuo Min, Fengjiao Chen, et al. Evaltalker: Learning to evaluate real-portrait-driven multi-subject talking humans. *arXiv preprint arXiv:2512.01340*, 2025.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. *arXiv preprint arXiv:2412.00733*, 2024.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In *ACM MM*, pages 9891–9900, 2025a.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *NeurIPS*, 2025.

Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint arXiv:2508.19209*, 2025.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025b.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025a.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.

Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *CVPR*, pages 1505–1515, 2023.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, pages 8652–8661, 2023.

Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *ACM SIGGRAPH Asia*, pages 1–9, 2022.

Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *CVPR*, pages 427–436, 2023.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *ECCV*, pages 85–101. Springer, 2022.

Yuan Gong, Yong Zhang, Xiaodong Cun, Fei Yin, Yanbo Fan, Xuan Wang, Baoyuan Wu, and Yujiu Yang. Toontalker: Cross-domain face reenactment. In *ICCV*, pages 7690–7700, 2023.

Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *ECCV*, pages 244–260. Springer, 2024.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *AAAI*, volume 39, pages 2403–2410, 2025a.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024.

Chunyu Li, Chao Zhang, Weikai Xu, Jinghui Xie, Weiguo Feng, Bingyue Peng, and Weiwei Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. *arXiv preprint arXiv:2412.09262*, 2024.

Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025b.

Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, Zerong Zheng, and Yanbo Zheng. Cyberhost: A one-stage diffusion framework for audio-driven talking body generation. In *ICLR*.

Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation. *arXiv preprint arXiv:2501.10687*, 2025.

Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025b.

Yubo Huang, Weiqiang Wang, Sirui Zhao, Tong Xu, Lin Liu, and Enhong Chen. Bind-your-avatar: Multi-talking-character video generation with dynamic 3d-mask-based embedding router. *arXiv preprint arXiv:2506.19833*, 2025.

Xingpei Ma, Shenneng Huang, Jiaran Cai, Yuansheng Guan, Shen Zheng, Hanfeng Zhao, Qiang Zhang, and Shunsi Zhang. Playmate2: Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. *arXiv preprint arXiv:2510.12089*, 2025.

Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. *arXiv preprint arXiv:2506.09984*, 2025c.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, et al. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *CVPR*, pages 8428–8437, 2025d.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *NeurIPS*, 37:48955–48970, 2024.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, pages 3661–3670, 2021.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, pages 251–263. Springer, 2017.

Kling. Kling. https://app.klingai.com, 2025.

Hedra. Hedra. https://www.hedra.com, 2025.

HeyGen. Heygen. https://www.heygen.com, 2025.