InfiniteTalk: Audio-driven Video Generation for Sparse-Frame Video Dubbing

Shaoshu Yang* [1, 2, 3], Zhe Kong* [4, 2, 5], Feng Gao* [2], Meng Cheng* [2], Xiangyu Liu* [6, 2, 1], Yong Zhang† [2], Zhuoliang Kang[2], Wenhan Luo[5], Xunliang Cai[2], Ran He[1, 3], and Xiaoming Wei[2]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Meituan
[3]New Laboratory of Pattern Recognition (NLPR), CASIA
[4]Shenzhen Campus of Sun Yat-sen University
[5]Division of AMC and Department of ECE, HKUST
[6]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[1]https://github.com/MeiGen-AI/InfiniteTalk

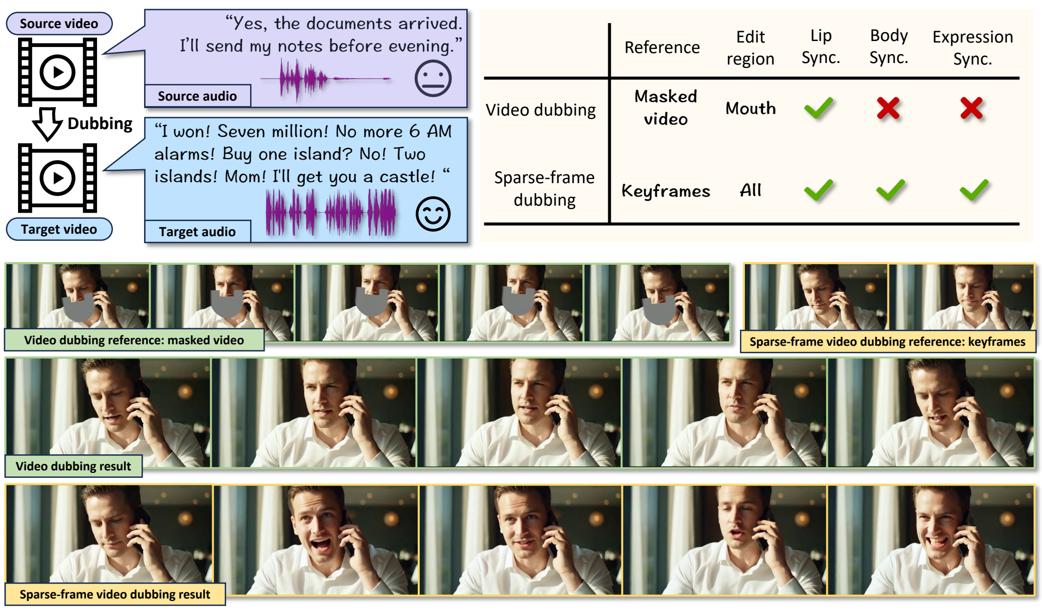

Figure 1: Compared to the traditional paradigm, sparse-frame video dubbing will not only edit mouth regions. It gives the model freedom to generate audio aligned mouth, facial, and body movements while referencing on sparse keyframes to preserve identity, emotional cadence, and iconic gestures.

Abstract

Recent breakthroughs in video AIGC have ushered in a transformative era for audio-driven human animation. However, conventional video dubbing techniques remain constrained to mouth region editing, resulting in discordant facial expressions and body gestures that compromise viewer immersion. To overcome this limitation, we introduce sparse-frame video dubbing—a novel paradigm that strategically preserves reference keyframes to maintain identity, iconic gestures, and camera trajectories while enabling holistic, audio-synchronized full-body motion editing. Through critical analysis, we identify why naive image-to-video models fail in this task, particularly their inability to achieve adaptive conditioning. Addressing this, we propose InfiniteTalk: a streaming audio-driven generator designed for infinite-length long sequence dubbing. This architecture leverages temporal

*Equal contribution
†Corresponding author

context frames for seamless inter-chunk transitions and incorporates a simple yet effective sampling strategy that optimizes control strength via fine-grained reference frame positioning. Comprehensive evaluations on HDTF, CelebV-HQ, and EMTD datasets demonstrate state-of-the-art performance. Quantitative metrics confirm superior visual realism, emotional coherence, and full-body motion synchronization.

1 Introduction

Video dubbing is an audio-driven video-to-video generation task that combines an original video with new audio to create localized content Li et al. (2024); Zhang et al. (2025); Bigata et al. (2025). This process requires editing facial movements, head rotations, and body gestures to synchronize with the dubbed speech's timing and emotional tone, while preserving the source video's visual style and camera motion. These capabilities are essential for global media distribution through streaming platforms.

Recent advances in audio-driven generative models have significantly improved lip synchronization capabilities for video dubbing Li et al. (2024). However, these methods predominantly focus on oral region inpainting, resulting in mismatched head rotations and body gestures that undermine viewer immersion. To address this limitation, we introduce sparse-frame video dubbing, a novel paradigm that preserves only reference keyframes while leveraging modern generative foundation models. As shown in Fig. 1, it references only select keyframes to preserve the original video's emotional cadence, symbolic gestures, and camera trajectories, while liberating facial expressions, head motions, and body dynamics to synchronize organically with dubbed audio. As a long video generation task, it demands robust temporal continuation capabilities. A practical solution is employing audio-conditioned video generators with initial and terminal frame guidance.

Unfortunately, naive application of audio-conditioned image-to-video generators produces unsatisfactory dubbing results, as shown in Fig. 2. These models fundamentally struggle with identity preservation during extended generation and default to strict motion copying on the conditioning frames. This results in stiff facial expressions, lip and head movements that contradicts speech dynamics. Meanwhile, simply applying initial and terminal frame conditioning creates abrupt inter-chunk transitions.

To resolve these challenges, we introduce InfiniteTalk, a audio-driven generator for long sequence sparse-frame video dubbing. It has a streaming video generation base that utilize context frames to inject momentum information that creates smooth inter-chunk transitions. To preserve human identity, background, and camera movements of the source video, it controls the output video by referencing keyframes. To achieve the soft reference mechanism in sparse-frame video dubbing, we investigate how and find the control strength is determined by the similarity between the video context and image condition. Based on our investigation, we propose a sampling strategy that balances control strength and motion alignment by fine-grained reference frame positioning, achieving high quality infinite length long sequence video dubbing with full body audio-aligned motion generation. Meanwhile, we explore methods to achieve accurate subtle camera movement preservation.

To comprehensively evaluate our method, we conduct quantitative and qualitative experiments on HDTF Zhang et al. (2021), CelebV-HQ Zhu et al. (2022), and EMTD Rang Meng (2025), including the cases for both face and full body animation. The quantitative results show InfiniteTalk achieves the state-of-the-art performance in audio synchronized motion generation and visual quality. Our human evaluation shows InfiniteTalk successfully produces plausible lip, face, and body movements that align with the speech cadence and the emotional expression. We finish an ablation experiment concerning the sampling strategy and the strength of control, showcasing the effectiveness of our algorithm design.

We summarize our contributions as follows: (1) We introduce sparse-frame video dubbing, a novel paradigm for human-centric audio-driven video-to-video generation to produce natural facial expressions, head motions, and body dynamics that synchronize organically with dubbed audio. (2) We analyze the reasons why audio-driven image-to-video generators fail to achieve satisfying performance in this task and how the reference frame positioning during training determines the control strength. With these observations, we propose propose InfiniteTalk, a streaming long video generator with soft conditioning training strategy. (3) Extensive experiments show our method achieves the state-of-the-art performance for video dubbing, especially in lip, head, body motion synchronization.

2 Related works

2.1 Video Generation

Recent advances in generative learning methods—including autoregressive models Tian et al. (2024a), diffusion models Wang et al. (2023b); Song et al. (2020); Nichol & Dhariwal (2021), and flow matching Liu et al. (2023)—have revolutionized video generation. Early efforts like the Video Diffusion Model Ho et al. (2022b) pioneered pixel-space denoising, while later works (Make-A-Video Singer et al. (2022), PYoCo Ge et al. (2023), Imagen Video Ho et al. (2022a)) integrated large language models for text-to-video synthesis. To tackle video dimensionality, research shifted toward latent space learning: VideoGPT Yan et al. (2021) combined VQ-VAE Esser et al. (2021) with transformers, establishing foundational latent modeling. Subsequent innovations employed diffusion models to approximate latent distributions, spawning latent video diffusion frameworks (He et al. (2022); Zhou et al. (2022); Xing et al. (2023); Blattmann et al. (2023b;a); Wang et al. (2023a); Chen et al. (2023; 2024)). Among these, CogVideoX Yang et al. (2024) introduced diffusion transformers and temporal-compressed VAEs, enhancing motion complexity. Leveraging expanded datasets and compute, modern large-scale generators (Team (2024); Weijie Kong & Jie Jiang (2024); Wan et al. (2025)) now achieve unprecedented quality.

2.2 Audio-driven Human Animation

Audio-driven human animation aims to generate dynamic videos from a static reference image, producing synchronized facial expressions and body movements based on audio control signals. In recent years, with the success of diffusion models, end-to-end audio-to-video synthesis methods Tian et al. (2024b); Wei et al. (2024); Xu et al. (2024); Chen et al. (2025b); Cui et al. (2024); Ji et al. (2024); Li et al. (2024); Jiang et al. (2024) have demonstrated considerable potential. These approaches eliminate the need for intermediate representations and exhibit superior performance in portrait animation for talking head generation. Another line of research extends talking head generation to talking body generation. Methods in this category Lin et al.; Tian et al. (2025); Lin et al. (2025); Meng et al. (2024); Gan et al. (2025); Wang et al. (2025) have achieved significant progress, demonstrating enhanced naturalness and consistent portrait animation capabilities, largely attributed to large-scale, high-quality training data. Recently, some works Chen et al. (2025a); Kong et al. (2025a); Huang et al. (2025) have attempted to advance from single human animation to multi-human animation.

The aforementioned human animation approaches have achieved satisfactory results in synthesizing short videos. However, when generating longer video sequences, they encounter error accumulation issues Kong et al. (2025b), such as identity degradation and color deviations. A concurrent work, named StableAvatar Tu et al. (2025), also achieves infinite-long sequence human animation. But it uses only a single image as the condition, and is not capable for the video-to-video task.

3 Method

3.1 Formulation

Conditional Flow matching for audio-driven video generation Flow matching video generative models Liu et al. (2023); Chen et al. (2025a); Wan et al. (2025) adopts a neural network to generate realistic video frames by modeling a timestep-dependent vector field that transports samples from a noise distribution to a target video distribution. Given a ground truth conditional video distribution $q(\boldsymbol{x}|\boldsymbol{c})$ where $\boldsymbol{x} \in \mathbb{R}^{t\times h\times w\times c}$ is the encoded video latent. $\boldsymbol{c} = \{y, a, \boldsymbol{x}_{\text{ref}}\}, y \in \mathbb{R}^{m\times d_{\text{text}}}, a \in \mathbb{R}^{n\times d_{\text{audio}}}, \boldsymbol{x}_{\text{ref}} \in \mathbb{R}^{t_{\text{ref}}\times h\times w\times c}$ are the conditions, including the text prompt embedding, and the audio embedding, and the reference frames latent. Conditional flow matching defines a series of distributions by interpolating $q(\boldsymbol{x}|y, a)$ with a known trivial distribution (e.g. Gaussian noise) $p(\boldsymbol{x}|y, a)$ using a continuous variable $t \in [0.1]$.

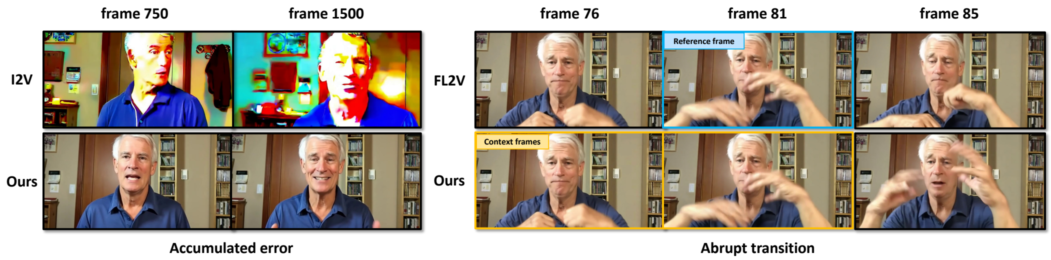

Figure 2: (**left**): I2V model accumulates error for long video sequences. (**right**): A new chunk starts from frame 82. FL2V model suffers from abrupt inter-chunk transitions.

$$q_t(\boldsymbol{x}|y,a) = (1-t)\cdot p(\boldsymbol{x}|y,a) + t\cdot q(\boldsymbol{x}|y,a). \tag{1}$$

To be specific, a random variable $\boldsymbol{x}_t \sim q_t(\boldsymbol{x}|y,a)$ can be obtained by interpolating between $\boldsymbol{x}_0 \sim p(\boldsymbol{x}|y,a)$ and $\boldsymbol{x}_1 \sim q(\boldsymbol{x}|y,a)$ via $\boldsymbol{x}_t = (1-t)\cdot\boldsymbol{x}_1 + t\cdot\boldsymbol{x}_0$. The generative model $\boldsymbol{v}_\theta(\cdot)$, parameterized by θ, is trained to match a continuous velocity field $\boldsymbol{v}_\theta(\boldsymbol{x}_t|y,a) \simeq \frac{d\boldsymbol{x}_t}{dt}$. To achieve so, we adopt the conditional flow matching objective.

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}\|\boldsymbol{v}_\theta(\boldsymbol{x}_t|y) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|_2^2. \tag{2}$$

An ODE solver can be used to sample from a flow matching generative models.

Sparse-frame video dubbing Video dubbing localizes content by replacing original audio with translated speech while preserving visual authenticity. As formalized in this work, the task transforms a source video latent $\boldsymbol{x}_0 \in \mathbb{R}^{t\times h\times w\times c}$ and a target audio $a \in \mathbb{R}^{n\times d_{\text{audio}}}$ into an output video where lip movements, facial expressions, and body dynamics synchronize organically with the new audio. Traditional video dubbing techniques focus exclusively on oral region inpainting—editing lip movements while freezing head rotations, facial expressions, and body gestures Li et al. (2024). This creates immersion-breaking mismatches, as static body language contradicts emotional speech (e.g., a rigid posture during passionate dialogue). Sparse-frame video dubbing, illustrated in Fig. 1, fundamentally redefines this process: it preserves only select keyframes $\boldsymbol{x}_{\text{ref}}$ to anchor identity, emotional cadence, symbolic gestures, and camera trajectories—critical for visual continuity—while liberating full-body dynamics (facial expressions, head motions, body gestures) to organically synchronize with dubbed audio. As Fig. 1 demonstrates, this paradigm shift enables lifelike alignment where head turns follow speech rhythm and gestures amplify emotional tone—impossible with lip-only editing. Crucially, sparse-frame dubbing operates on infinite-length sequences, demanding generative continuation beyond short clips to maintain synchronization across extended durations, a capability unattainable with traditional frame-by-frame inpainting.

3.2 Observation on naive solutions

This section investigates practical approaches for sparse-frame video dubbing using two baseline models: image-to-video (I2V) Cui et al. (2024) and first-last-frame-to-video (FL2V) Wan et al. (2025). As illustrated in Fig. 2, both methods exhibit critical limitations when generating long video sequences. The I2V approach operates by initializing the first chunk of the video from a single reference frame (e.g., the source video's starting keyframe). For subsequent chunks, it uses only the last generated frame of the preceding chunk as the new reference. While this preserves motion flexibility, the lack of persistent anchoring to the original keyframes leads to accumulated errors: subtle discrepancies in identity (e.g., facial features gradually deviating from the source actor) and color tones (e.g., background hues shifting across chunks) compound over time, resulting in visible degradation. In contrast, the FL2V method conditions each chunk on both the start and end frames of the input segment, ensuring alignment with the source video's reference poses. This eliminates accumulation errors but introduces a new problem: the model enforces rigid control by strictly replicating the reference frames at the corresponding timestamp. This contradicts the soft

Figure 3: A visual comparison between the training reference positioning strategies. All video chunks are generated using the same context frames and the same reference frame shown in below.

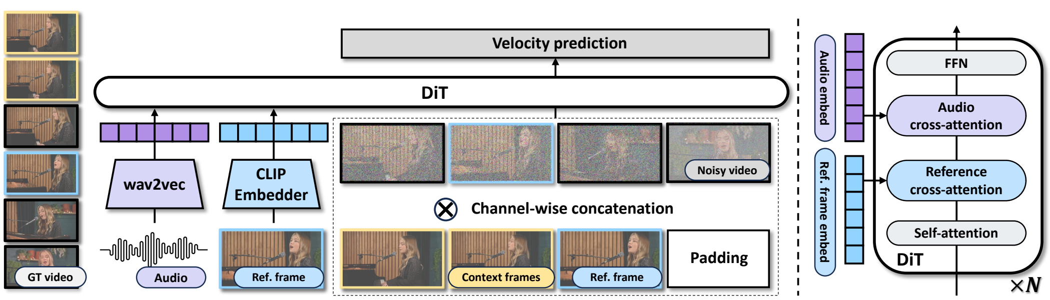

Figure 4: Visualization of InfiniteTalk pipeline. **Left**: The streaming model receives a audio, a reference frame, and context frames to denoise iteratively. **Right**: The architecture of the diffusion transformer. In addition to the traditional structures, each block includes an audio cross-attention layer and a reference cross-attention layer

conditioning required for sparse-frame dubbing, where full body motions must dynamically adapt to audio cues.

Crucially, both methods suffer from abrupt inter-chunk transitions. Since they rely solely on static image conditions (e.g., a single frame for I2V or two fixed frames for FL2V), they lack momentum information that should carry over between chunks. These observations a highlight fundamental trade-off: I2V prioritizes motion fluidity at the expense of accumulated error, while FL2V prioritizes reference fidelity at the cost of motion naturalness.

3.3 Audio-driven streaming video generator with reference frames

To resolve the accumulated accumulated error in I2V models and the abrupt transitions in FL2V models, we build a audio-driven streaming human animation architecture. This framework employs context frames, defined as the trailing segment of each previously generated chunk, to propagate kinetic momentum into subsequent segments. By processing these frames through a diffusion transformer, the model sustains motion continuity. To eliminate accumulated errors, we adopt multiple reference frames dynamically sampled from the source video, similar to FL2V's multi-frame conditioning. These keyframes preserve critical visual attributes including identity, background details, camera trajectories, and stylistic elements. Crucially, unlike FL2V's rigid replication of

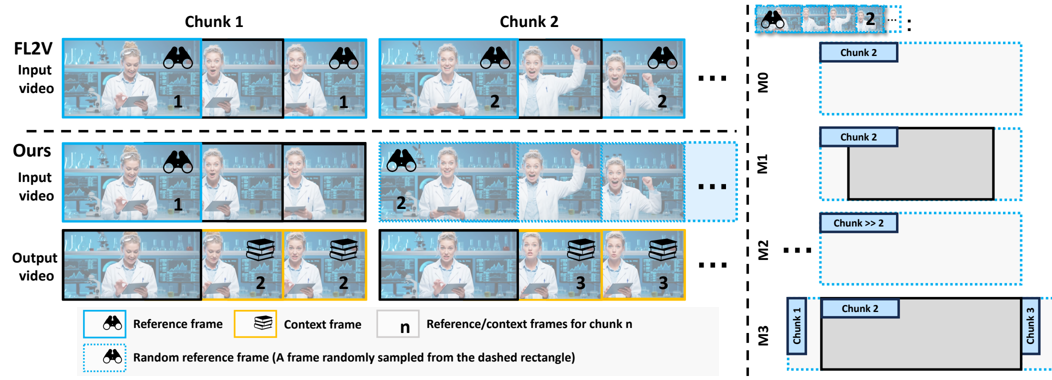

Figure 5: Visualization of reference frame conditioning strategies for video dubbing models. Top four rows: conditioning on input video frames. Bottom row: conditioning on generated video frames. **Left**: Image-to-video dubbing model with initial frame conditioning (I2V) and initial+terminal frame conditioning (IT2V). **Right**: Streaming dubbing model with four conditioning strategies. Within each category (left/right), all strategies share identical generated-video conditioning approaches.

reference frames at fixed positions, which suppresses natural motion. Our model has the potential to achieve soft conditioning as discussed in Section 3.4.

The sketch for the model is shown in Fig. 4. The model consists an audio embedder Schneider et al. (2019), a video VAE, and diffusion transformer (DiT) Peebles & Xie (2022). We first introduce how our model is trained. In the followings, let the video latent refers to embedding derived by encoding pixel space video by using the video VAE. Similarly, let audio embedding denote the embedding computed by encoding a audio sequence using the audio embedder. To train the model, we do not need a dubbed video pair, but only a video with the audio track is needed. Given a source video. The reference frame is uniformly random sampled from the video. During training, the context frames are the first $4(t_c - 1) + 1$ frames of the source video. After VAE encoding, we derive reference frame latent $\boldsymbol{x}_{\text{ref}} \in \mathbb{R}^{c \times 1 \times h \times w}$, the full source video latent $\boldsymbol{x}_{\text{full}} \in \mathbb{R}^{c \times (t+t_c) \times h \times w}$, the context frames latent $\boldsymbol{x}_{\text{context}} \in \mathbb{R}^{c \times t_c \times h \times w}$, and the subsequent frames latent $\boldsymbol{x}_0 \in \mathbb{R}^{c \times t \times h \times s}$ separately, where the full video latent is a combination of the context frames latent $\boldsymbol{x}_{\text{context}}$ and the subsequent frames latent $\boldsymbol{x}_0$, $\boldsymbol{x}_{\text{full}} = \{\boldsymbol{x}_{\text{context}}, \boldsymbol{x}_0\}$. The audio sequence in the source video is encoded to get embedding a. Without a loss of generality, we show the process to train at t in conditional flow matching in this section. Unit gaussian noise is used as the trivial distribution. The noisy latent is derived by $\boldsymbol{x}_t = (1-t) \cdot \boldsymbol{x}_1 + t \cdot \boldsymbol{x}_0, \boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ where $\boldsymbol{x}_1$ is gaussian noise that has the same dimensionality of $\boldsymbol{x}_0$. The DiT model formulates a field estimator $\boldsymbol{v}_\theta(\boldsymbol{x}_t|\boldsymbol{c})$ using condition $\boldsymbol{c} = \{y, a, \boldsymbol{x}_{\text{ref}}, \boldsymbol{x}_{\text{tran}}\}$. Specifically, we first concatenate the noisy latent $\boldsymbol{x}_t$ and the clean context frames $\boldsymbol{x}_{\text{context}}$ in the temporal dimension to get $\boldsymbol{z}_1 \in \mathbb{R}^{c \times (t+t_c) \times h \times w}$. Then we pad the reference frame to the temporal length $t_c + t$ to get $\boldsymbol{z}_2 \in \mathbb{R}^{c \times (t+t_c) \times h \times w}$. Finally, we concatenate $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and a reference frame indicating mask $\boldsymbol{m} \in \mathbb{R}^{4 \times (t+t_c) \times h \times w}$ in channel dimension. Mathematically, the process is

$$
\begin{aligned}
\boldsymbol{z}_1 &= \text{concat}((\boldsymbol{x}_{\text{context}}, \boldsymbol{x}_t), 2) \\
\boldsymbol{z}_2 &= \text{concat}((\boldsymbol{x}_{\text{ref}}, \nvdash), 2) \\
\boldsymbol{m} &= \text{concat}((\nVdash, \nvdash), 2) \\
\boldsymbol{z} &= \text{concat}((\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{m}), 1)
\end{aligned}
\tag{3}
$$

where $\text{concat}(\cdot)$ is the concatenation operator (e.g, $\text{concat}((\boldsymbol{x}_{\text{context}}, \boldsymbol{x}_t), 2)$ concatenates $\boldsymbol{x}_{\text{context}}$ and $\boldsymbol{x}_t$ in the 2-nd dimension). $\nvdash$ and $\nVdash$ are zero and one tensors that have the dimensions $\nvdash \in \mathbb{R}^{4 \times (t+t_c-1) \times h \times w}$, $\nVdash \in \mathbb{R}^{4 \times 1 \times h \times w}$. As depicted by Fig. 4 **(right)**. Within our transformer model, there is an audio cross-attention module and an image cross-attention that achieves audio and reference image conditioning. The reference frame is processed by CLIP vision model Cherti et al. (2023) to get the embedding $\boldsymbol{z}_{\text{ref}}$ before feeding to the DiT. To train this model, we adopt the conditional flow matching objective Liu et al. (2023).

$$\mathcal{L}_{\mathrm{fm}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{c}} \|\boldsymbol{v}_\theta(\boldsymbol{x}_t|\boldsymbol{c}) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|_2^2. \tag{4}$$

Then, we briefly introduce the sampling method. An illustration is presented in Fig. 5. We generate the whole long video sequence by auto-regressively generating small video chunks. In the first video chunk, we use the first frame of the input video as the reference frame and no context frames are required. In the following video chunks, we use the last $4(t_c - 1) + 1$ frames from the previous output video chunk as the context frames and use the first image of the input chunk as the reference frame.

3.4 Soft conditioning

Control strength from the Reference frame This section explores strategies for achieving soft conditioning in sparse-frame video dubbing where the model must generate audio-aligned full-body motions without rigidly replicating reference frames that may conflict with dubbed speech. Meanwhile, we expect the model to have adaptive control strength: (1) when the reference is similar to the context frames, the control strength is weak such that the model produces diverse dynamics. (2) when the reference is very distinct to the context frames, the control strength is stronger to ensure better consistency in identity and background. Below, we show the analysis on the training strategies that match the requirements.

We initiate our investigation with Model M0, which samples reference frames uniformly at random from the current source input video chunk during training. As demonstrated in Fig. 3, this approach exhibits excessive control strength. The model inappropriately duplicates reference content at arbitrary timestamps, such as replicating slapping on the forehead during emotionally neutral speech, disrupting audio-visual synchronization.

To systematically analyze how reference positioning during training governs control fidelity, we examine two granular dimensions: chunk-level positioning (selecting which temporal segment provides the reference) and frame-level positioning (choosing the specific frame within that segment). We train three model variants to isolate these effects. Model M1 samples references exclusively from the first or last frame of the input chunk. This strategy mirrors FL2V's rigidity, forcing the generated video to replicate reference poses precisely at chunk boundaries, even when they contradict the audio's emotional cadence. Model M2 samples references from temporally distant chunks (e.g., segments separated by >5 seconds). While this weakens control sufficiently to avoid replication, it introduces accumulated color and background errors over long sequences, indicating insufficient fidelity preservation. In contrast, Model M3 samples references from adjacent chunks (e.g., within 1 seconds of the input). This configuration achieves moderate control strength: references preserve identity and camera motion without exact duplication, while eliminating accumulated errors entirely.

Our experiments conclusively demonstrate that chunk-level distance is the dominant factor modulating control strength. Shorter temporal distances (as in M3) create an optimal equilibrium: they anchor visual consistency to the source while liberating facial expressions, head rotations, and body gestures to organically synchronize with audio. Longer distances (as in M2) degrade preservation capabilities, leading to destabilized outputs. Fixed boundary sampling (as in M1) prioritizes replication over expressiveness, stifling motion dynamics. Thus, M3's near-chunk positioning emerges as the foundational strategy for soft conditioning—enabling faithful yet flexible video dubbing where motions breathe in harmony with speech.

Camera control We explore ways to address the camera movement preservation in sparse-frame video dubbing in this section. The utilization of reference frames is providing a global control of camera trajectory. However, the detailed camera movement within each video chunk is not controlled and may contradict the source video. To resolve this, we use two plugin including SDEdit Meng et al. (2022) and Uni3C Cao et al. (2025). SDEdit incorporates trajectory information by adding the source video to the initialize noise at a scale t_0. The denoising sampling process starts from $t = t_0$ instead of $t = 1$. $\boldsymbol{x}_{t_0} = (1 - t_0) \cdot \boldsymbol{x}_1 + t_0 \cdot \boldsymbol{x}_0$ Uni3C injects camera movements by deploying a ControlNet-like architecture. A comparison between the methods is shown in our experiments.

Dataset	Model	Metrics				
		FID↓	FVD ↓	Sync-C↑	Sync-D↓	CSIM↑
HDTF	LatentSync	16.09	**48.45**	8.99	**6.36**	0.916
	MuseTalk	**14.20**	49.13	7.17	7.90	0.933
	Ours	26.11	131.65	**9.35**	6.67	0.775
CelebV-HQ	LatentSync	17.80	**67.97**	6.90	7.33	0.869
	MuseTalk	**17.62**	72.07	4.16	9.86	0.857
	Ours	32.29	229.67	**7.53**	**7.33**	0.726
EMTD	LatentSync	**11.43**	212.60	8.10	**6.97**	0.846
	MuseTalk	14.26	**46.07**	5.35	9.28	0.825
	Ours	32.55	312.17	**8.60**	7.16	0.713

Table 1: Quantitative comparisons our methods between the traditional video dubbing models.

4 EXPERIMENT

Implementation details We build our model based on MeiGen-MultiTalk Kong et al. (2025a), which includes a 14B parameters DiT that enables audio-driven image-to-video generation at multiple resolutions. We use wav2vec2 Baevski et al. (2020) as the audio embedder, and CLIP/H Cherti et al. (2023) as the reference image embedder. Around 2,000 hours of video containing a talking person is collected as our training data. The model is trained with a 64 NVIDIA H100 80G cluster. The context frames includes 9 images, resulting in a $t_c = 3$ context frames latent. The video chunk length is 81. Our model will produce 72 frames each time when generating a long video auto-regressively.

Test datasets and evaluation metrics To rigorously evaluate our method across diverse scenarios, we utilize three benchmark datasets: HDTF Zhang et al. (2021) and CelebV-HQ Zhu et al. (2022) (emphasizing facial dynamics) alongside EMTD Rang Meng (2025) (incorporating full-body movements). Following established dubbing evaluation protocols Li et al. (2024); Fei et al. (2025), we construct a test set of 120 videos by randomly sampling 40 videos per dataset and permuting their audio channels (replacing original audio with mismatched tracks) to simulate real dubbing conditions. We use our model to perform long sequence video dubbing at 480×480 resolution. The temporal length of generated results is the frame number of the the dubbing video input.

Performance is quantified through complementary automatic metrics and human evaluation. For objective assessment, we employ: Fréchet Inception Distance (FID) measuring per-frame visual quality; Fréchet Video Distance (FVD) evaluating inter-frame temporal coherence; SyncNet's Sync-C (confidence score) and Sync-D (lip distance) quantifying lip synchronization; and Cosine Similarity (CSIM) scoring identity preservation. To capture perceptual nuances beyond automated metrics, we conduct human studies where participants rate on 5 perspectives: gesture synchronization with audio prosody, head motion alignment to speech rhythm, lip synchronization precision, identity consistency, and overall perceptual naturalness. We receive 340 responses from 17 participants on all 40 video dubbing results from EMTD.

4.1 QUANTITATIVE EXPERIMENTS

We compare InfiniteTalk with both traditional video dubbing methods (including MuseTalk Zhang et al. (2025), FantacyTalking Wang et al. (2025), and Hallo3 Cui et al. (2024)) and audio-driven image-to-video models (including OmniAvatar Gan et al. (2025) and MultiTalk Kong et al. (2025a)). To follow the pre-processing pipeline and accurately show the performance of the counterparts, we use their open-source weights and inference scripts in this experiment.

The comparison with image-to-video models is shown in Table 2. InfiniteTalk outperforms the counterparts in lip synchronization by a large margin. Note that, there is a trade-off between synchronization metrics (Sync-C, Sync-D), visual quality matrics (FID, FVD) and identity preservation (CSIM). A trivial solution is that, if a method is copying the input as the output, it will achieve the best FID, FVD, and CSIM over all methods. When comparing InfiniteTalk with methods that have competitive synchronization performances, our method achieves very remarkable visual quality and

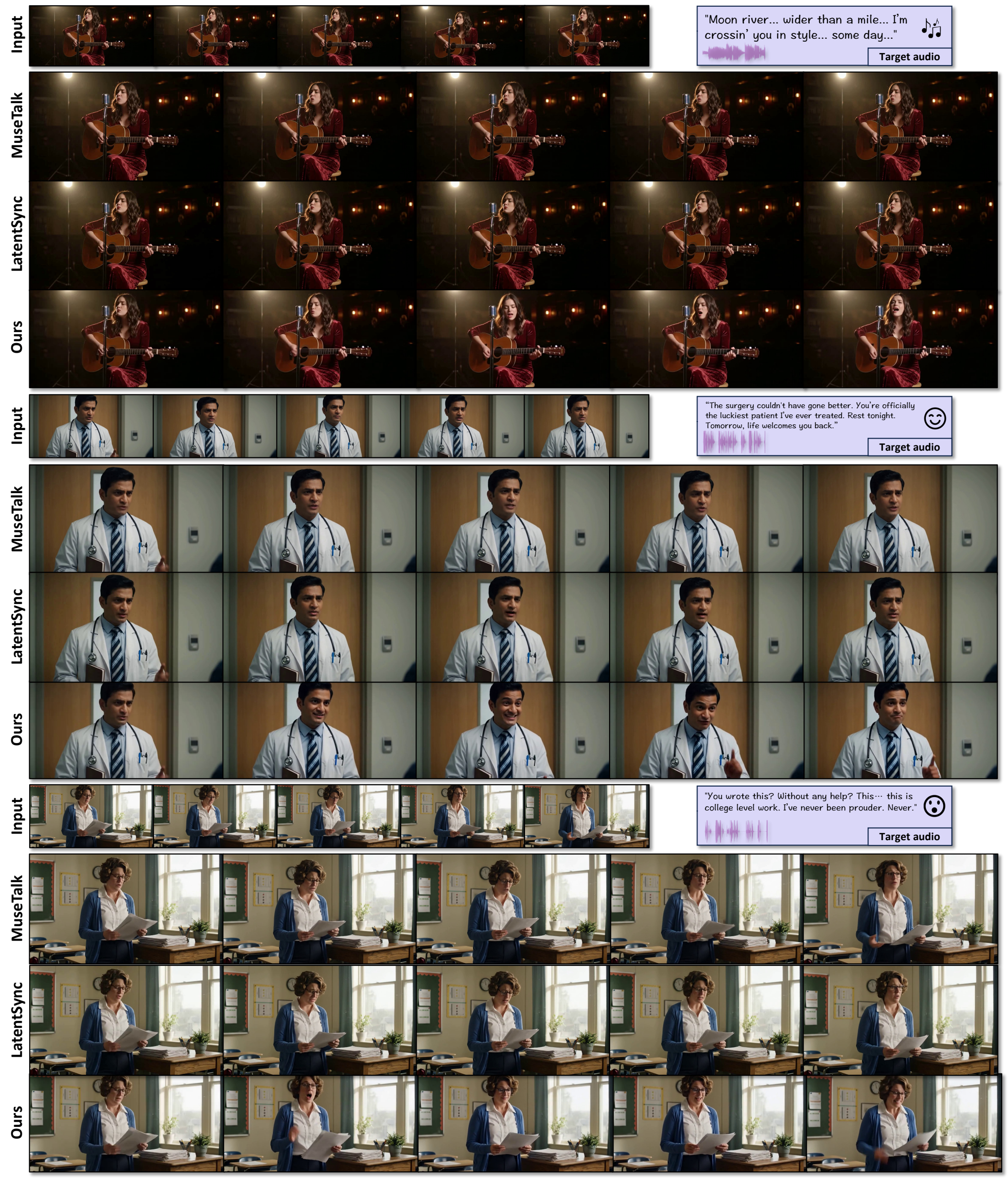

Figure 6: A visual comparison between the video dubbing methods.

identity preserving. As seen in Table 1 when comparing with traditional video dubbing methods. Since LatentSync Li et al. (2024) and MuseTalk Zhang et al. (2025) is limited to edit the oral region, the rest of the video is kept unchanged, making their FID and FVD extremely good. Given that, the metrics are not showing the true visual quality difference.

Currently, there is not automatic full body motion-audio alignment metric available. We find the music-motion alignment metric like beat consistency score fail to differentiate the performance of the methods. Meanwhile, Sync-C and Sync-D cannot accurately depict the lip synchronization when there is large head motion in the video. For a comprehensive evaluation of full body motion alignment, we conduct a user study. The result is shown in Table 4. The participant is asked to rank the results (placing the best to the 1-st place, the worst to the 3-rd place) by the video dubbing methods including MuseTalk Zhang et al. (2025), LatentSync Li et al. (2024), and our method. The number in the table shows the averaged ranking of the corresponding method. Benefited by the strong audio-driven

Dataset	Model	Metrics				
		FID↓	FVD ↓	Sync-C↑	Sync-D↓	CSIM ↑
HDTF	FantacyTalking	32.06	**110.36**	3.78	10.80	0.684
	Hallo3	36.48	144.65	7.20	8.61	0.674
	OmniAvatar	26.63	112.49	7.06	8.63	0.752
	MultiTalk	27.61	133.58	9.02	6.96	0.754
	Ours	**27.14**	132.54	**9.18**	**6.84**	0.751
CelebV-HQ	FantacyTalking	37.53	237.58	2.93	10.79	0.654
	Hallo3	42.36	258.65	5.63	9.12	0.591
	OmniAvatar	37.41	250.67	5.88	8.68	0.703
	MultiTalk	34.79	230.41	7.25	7.70	0.711
	Ours	**33.96**	**230.12**	**7.41**	**7.59**	0.713
EMTD	FantacyTalking	36.66	**298.24**	3.60	11.31	0.626
	Hallo3	44.71	326.94	5.68	9.56	0.512
	OmniAvatar	**29.47**	308.14	6.93	8.55	0.694
	MultiTalk	33.80	315.33	8.13	7.50	0.702
	Ours	33.27	314.68	**8.34**	**7.36**	0.709

Table 2: Quantitative comparisons between our method and audio-driven image-to-video models.

Model	FID↓	FVD ↓	Sync-C↑	Sync-D↓
Ours (M0)	32.69	322.04	8.51	7.31
Ours (M1)	**32.21**	**307.21**	7.96	8.11
Ours (M2)	42.17	376.53	8.23	7.44
Ours (M3)	32.55	312.17	**8.60**	**7.16**

Table 3: Ablation experiment results on EMTD.

human animation architecture and the sparse-frame video dubbing paradigm, our method achieves the best results in both lip and body motion synchronization. It demonstrates the limitation of traditional video dubbing methods that the editing region is restricted to mouth, resulting to misaligned body motion.

4.2 Qualitative experiments

Comparison with counterparts We conduct a visual comparison between our method and traditional video dubbing methods in Fig. 6. The first input example is a static video. It showcases when only editing mouth regions, traditional video dubbing methods cannot drive the head and body by the audio track. The following two inputs are dynamic videos. Compared to the counterparts, InfiniteTalk is not only able to generate plausible audio-aligned lip movements, but also synchronized face, head, and body movements with matched emotional expressions. As a new paradigm, sparse-frame video dubbing also demonstrates its necessity for modern audio-driven human animation video-to-video applications.

Camera control A visual comparison on the camera control methods are shown in Fig. 7. Using InfiniteTalk alone will not replicate the subtle camera movement of source video. With SDEdit Meng et al. (2022) or Uni3C Cao et al. (2025), we can achieve fine-grained camera control. Comparing SDEdit with Uni3C. We find that with Uni3C, the model fails to preserve the video background.

4.3 Ablation study

To systematically evaluate which training strategy performs the best in sparse-frame video dubbing, we conduct ablation studies comparing the four different reference frame positioning methods introduced in Section 3.4. All ablated models are rigorously benchmarked using the 40-video test set sampled from EMTD under identical conditions. As shown in Table 3, our fine-grained reference

Model	Lip Sync.↓	Body Sync. ↓
MuseTalk	2.57	-
LatentSync	2.32	1.92
Ours	**1.11**	**1.09**

Table 4: Human evaluation between video dubbing methods on motion synchronization. We do not compare our method with MuseTalk in body synchronization because MuseTalk has exactly the same body movement to LatentSync.

Figure 7: A visual comparison on the camera control.

frame positioning during training is the key to achieve reliable visual quality and audio-motion synchronization.

5 Conclusion

We introduce sparse-frame video dubbing, a novel paradigm for audio-driven video-to-video generation that employs reference keyframes to maintain emotional cadence and camera trajectories while liberating facial, head, and body dynamics to synchronize organically with dubbed audio. We propose InfiniteTalk, an audio-driven generator that overcomes critical limitations in long-form synthesis. By incorporating transient frame conditioning for seamless transitions, motion-provoking sampling to activate natural gestures, and adaptive camera control, InfiniteTalk achieves state-of-the-art lip, head, and body synchronization while eliminating identity drift and motion artifacts across extended sequences. Extensive validation confirms its superiority in producing natural, audio-aligned dynamics essential for immersive dubbed content.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`.

Antoni Bigata, Rodrigo Mira, Stella Bounareli, Michał Stypułkowski, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Keysync: A robust approach for leakage-free lip synchronization in high resolution, 2025. URL `https://arxiv.org/abs/2505.00497`.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Chenjie Cao, Jingkai Zhou, shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 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, 2024.

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*, 2025a.

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, pp. 2403–2410, 2025b.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

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.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Zhengcong Fei, Hao Jiang, Di Qiu, Baoxuan Gu, Youqiang Zhang, Jiahua Wang, Jialin Bai, Debang Li, Mingyuan Fan, Guibin Chen, and Yahui Zhou. Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers, 2025. URL `https://arxiv.org/abs/2506.00830`.

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.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *arXiv preprint arXiv:2305.10474*, 2023.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022b.

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.

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.

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.

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. *arXiv preprint arXiv:2505.22647*, 2025a.

Zhe Kong, Le Li, Yong Zhang, Feng Gao, Shaoshu Yang, Tao Wang, Kaihao Zhang, Zhuoliang Kang, Xiaoming Wei, Guanying Chen, et al. Dam-vsr: Disentanglement of appearance and motion for video super-resolution. *arXiv preprint arXiv:2507.01012*, 2025b.

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.

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*.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2023. URL `https://arxiv.org/abs/2209.03003`.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Yuming Li Chenguang Ma Rang Meng, Xingyu Zhang. Echomimicv2: Towards striking, simplified, and semi-body human animation, 2025.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019. URL `https://arxiv.org/abs/1904.05862`.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024a.

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*, pp. 244–260. Springer, 2024b.

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.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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. *arXiv preprint arXiv:2504.04842*, 2025.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023a.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023b.

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

Zijian Zhang Rox Min Zuozhuo Dai Jin Zhou Jiangfeng Xiong Xin Li Bo Wu Jianwei Zhang Kathrina Wu Qin Lin Aladdin Wang Andong Wang Changlin Li Duojun Huang Fang Yang Hao Tan Hongmei Wang Jacob Song Jiawang Bai Jianbing Wu Jinbao Xue Joey Wang Junkun Yuan Kai Wang Mengyang Liu Pengyu Li Shuai Li Weiyan Wang Wenqing Yu Xinchi Deng Yang Li Yanxin Long Yi Chen Yutao Cui Yuanbo Peng Zhentao Yu Zhiyu He Zhiyong Xu Zixiang Zhou Zunnan Xu Yangyu Tao Qinglin Lu Songtao Liu Daquan Zhou Hongfa Wang Yong Yang Di Wang Yuhong Liu Weijie Kong, Qi Tian and along with Caesar Zhong Jie Jiang. Hunyuanvideo: A systematic framework for large video generative models, 2024. URL `https://arxiv.org/abs/2412.03603`.

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023.

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.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

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.

Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high-fidelity video dubbing via spatio-temporal sampling, 2025. URL `https://arxiv.org/abs/2410.10122`.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

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.