Training-Free Multi-Character Audio-Driven Animation via Diffusion Transformer with Reward Feedback

Xingpei Ma*, Shenneng Huang*, Jiaran Cai*†, Yuansheng Guan*, Shen Zheng*, Hanfeng Zhao, Qiang Zhang, Shunsi Zhang

Guangzhou Quwan Network Technology

{maxingpei, huangshenneng, caijiaran, guanyuansheng, zhengshen, zhaohanfeng, zhangqiang, zhangshunsi}@52tt.com

Abstract

Recent advances in diffusion models have significantly improved audio-driven human video generation, surpassing traditional methods in both quality and controllability. However, existing approaches still face challenges in lip-sync accuracy, temporal coherence for long video generation, and multi-character animation. In this work, we propose a diffusion transformer (DiT)-based framework for generating lifelike talking videos of arbitrary length, and introduce a training-free method for multi-character audio-driven animation. First, we employ a LoRA-based training strategy combined with a position shift inference approach, which enables efficient long video generation while preserving the capabilities of the foundation model. Moreover, we combine partial parameter updates with reward feedback to enhance both lip synchronization and natural body motion. Finally, we propose a training-free approach, Mask Classifier-Free Guidance (Mask-CFG), for multi-character animation, which requires no specialized datasets or model modifications and supports audio-driven animation for three or more characters. Experimental results demonstrate that our method outperforms existing state-of-the-art approaches, achieving high-quality, temporally coherent, and multi-character audio-driven video generation in a simple, efficient, and cost-effective manner.

Project Page — https://playmate111.github.io/Playmate2/

1 Introduction

Benefiting from large-scale pre-training and advanced architectural designs, diffusion models (Rombach et al. 2022; Esser et al. 2024; Liu et al. 2024; Yang et al. 2024; Kong et al. 2024; Wan et al. 2025) have achieved significant advances in image and video synthesis, outperforming traditional generative adversarial networks (GANs) (Goodfellow et al. 2020) in both visual quality and temporal coherence. These advancements have greatly improved the generation of audio-driven human videos (Xue et al. 2024), making this capability a cornerstone of digital human research. Audio-driven human animation has broad applications in digital entertainment, film and gaming production, virtual reality, and digital storytelling.

Audio-driven human animation (Jiang et al. 2024) synthesizes realistic character videos with synchronized lip movements and natural body gestures from speech and auxiliary inputs. Recent advances in diffusion models have spurred their use in this domain, leading to two main categories: portrait animation and human animation. The first focuses on synthesizing facial expressions solely from audio signals, with little attention given to background dynamics (Tian et al. 2024; Xu et al. 2024a,b; Ji et al. 2024; Ma et al. 2023; Chen et al. 2024; Cui et al. 2025b). Such a restricted approach frequently compromises the realism of generated videos in complex scenes, leading to results that do not satisfy the demands of high-quality applications. The second employs video diffusion models to overcome the aforementioned spatial constraints, thereby achieving full-body animation generation (Lin et al. 2025; Fei et al. 2025; Wang et al. 2025a; Chen et al. 2025; Kong et al. 2025; Cui et al. 2025a). Despite progress, several challenges persist: 1) Existing methods often struggle to maintain accurate lip-sync while generating natural body movements; 2) In long video synthesis, current solutions often result in jittery motions and abrupt transitions, failing to preserve temporal coherence; 3) Most existing techniques are unable to animate scenes involving multiple characters using audio input; although some works achieve multi-character animation by constructing multi-speaker datasets and introducing significant modifications to the model architecture, such strategies are often resource-intensive and not scalable.

To address these challenges, leveraging the large-scale video diffusion model Wan2.1 (Wan et al. 2025), we propose a diffusion transformer (DiT)-based (Peebles and Xie 2023) framework for audio-driven facial and human video generation, aiming to enhance video quality and enable cost-effective multi-character animation. First, we adopt a LoRA-based (Hu et al. 2022) training strategy to preserve the capabilities of the foundation model while enabling long video generation. Next, we explore a training strategy that combines partial parameter updates with reward feedback, producing videos with accurate lip synchronization and natural body motions. Finally, inspired by Classifier-Free Guidance (CFG) (Ho and Salimans 2022), we introduce a training-

*These authors contributed equally.

†Project lead & Corresponding Author.

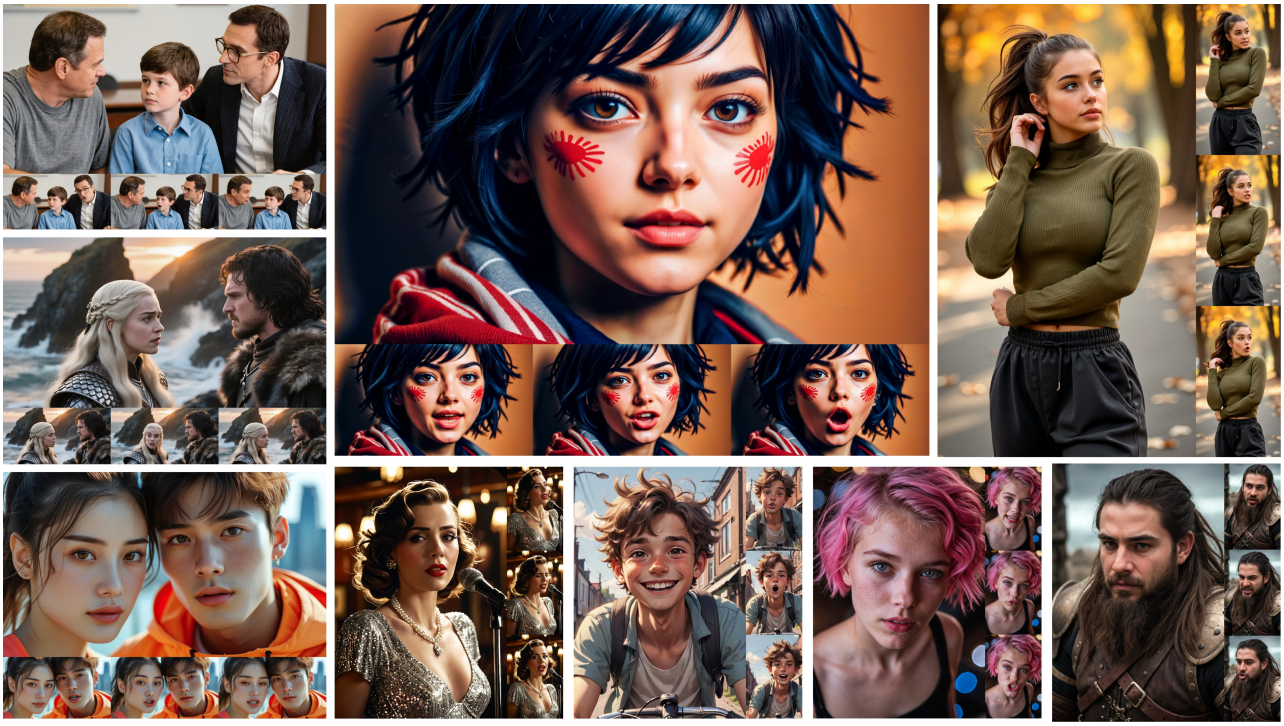

Figure 1: We present a novel DiT-based framework for generating high-quality, audio-driven human videos that effectively tackles key challenges related to temporal coherence in long sequences and multi-character animations. To the best of our knowledge, this is the first training-free approach capable of enabling audio-driven animation for three or more characters without requiring additional data or model modifications.

free approach, Mask Classifier-Free Guidance (Mask-CFG), to tackle the challenge of multi-character animation. This method does not require constructing specialized datasets or modifying the model architecture; instead, it achieves multi-character control through simple adjustments during inference, making it both efficient and cost-effective.

To the best of our knowledge, this is the first training-free approach capable of enabling audio-driven animation for three or more characters. In summary, our contributions are as follows:

- We propose a DiT-based framework for audio-driven human animation, combined with a LoRA-based training strategy for long video generation.
- We investigate a training strategy combining partial parameter updates with reward feedback to improve lip-sync accuracy while maintaining natural and adaptive body movements.
- We introduce a training-free method (Mask-CFG) to support multi-character animation, which is both efficient and cost-effective.

2 Related Work

2.1 Audio-Driven Portrait Animation

Prior work on audio-driven portrait animation has largely focused on lip-sync accuracy (Prajwal et al. 2020; Zhang et al. 2023b,a; Guo et al. 2021; Wang et al. 2024). Traditional approaches based on GANs, neural radiance fields (NeRF) (Mildenhall et al. 2021), and 3D Gaussian Splatting (Kerbl et al. 2023) have achieved strong results, yet often fail to model the subtle relationship between prosody and facial dynamics, leading to limited expressiveness and reduced visual realism. Recently, diffusion-based methods have enabled end-to-end talking video generation. EMO (Tian et al. 2024) improves inter-frame consistency for stable, natural synthesis. Hallo (Xu et al. 2024a) jointly addresses lip synchronization, expression, and pose. Sonic (Ji et al. 2024) emphasizes global perceptual coherence for diverse motions, while DICE-Talk (Tan et al. 2025) and Playmate (Ma et al. 2025) introduce emotional control for expressive portraits. Despite producing realistic outputs, these methods are primarily limited to facial animation and do not support full-body motion synthesis.

2.2 Audio-Driven Human Animation

To enable audio-driven human animation, recent methods leverage large-scale video diffusion models. CyberHost (Lin et al. 2024) proposes a one-stage framework with novel attention and human-prior-guided training for upper-body synthesis. Approaches like OmniHuman-1 (Lin et al. 2025), FantasyTalking (Wang et al. 2025a), SkyReels-Audio (Fei et al. 2025), and OmniAvatar (Gan et al. 2025) build on models such as Seaweed (Seawead et al. 2025), Hunyuan-

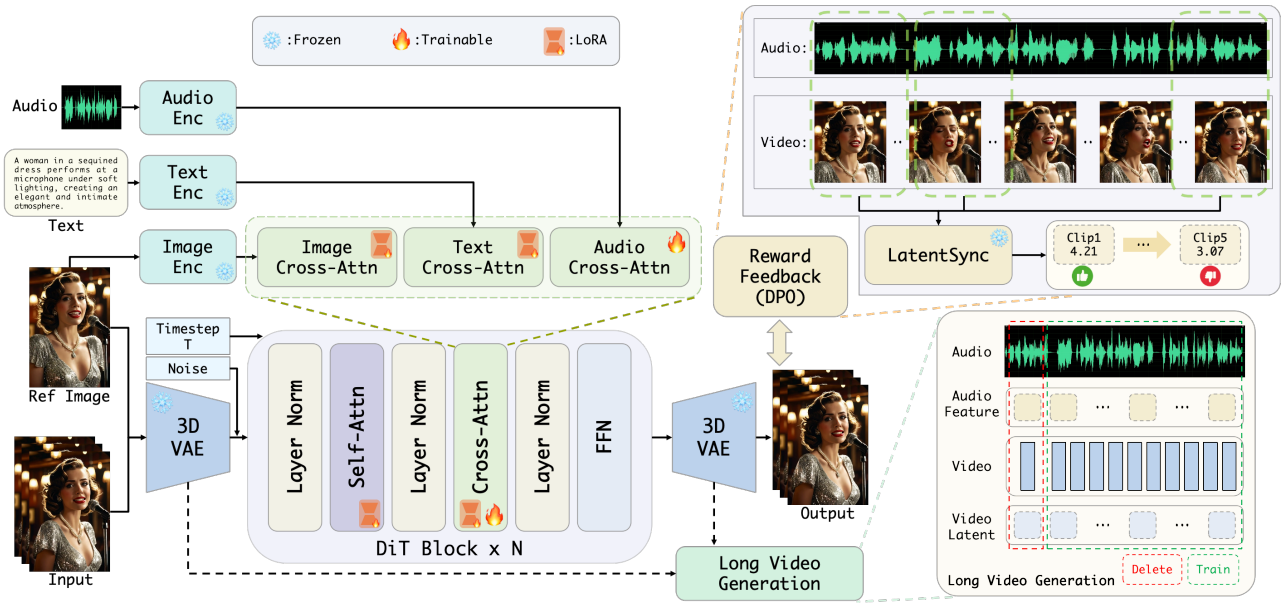

Figure 2: Overview of our method. Our framework leverages a LoRA-based training strategy and position shift inference to generate long, temporally coherent videos with consistent identity. A partial parameter update with reward feedback enhances lip synchronization and upper-body motion naturalness. Furthermore, we propose Mask-CFG, a training-free approach for multi-character animation that requires no additional data or model fine-tuning, yet supports audio-driven animation of three or more characters.

Video (Kong et al. 2024), and Wan2.1 for holistic motion generation. In multi-character scenarios, HunyuanVideo-Avatar (Chen et al. 2025) uses latent-space masking for localized, character-specific control, while MultiTalk (Kong et al. 2025) introduces Label Rotary Position Embedding with a multi-person dataset to resolve audio-person binding. Inspired by these advances, we base our approach on a large-scale video diffusion transformer for audio-driven human animation.

2.3 Direct Preference Optimization

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) is widely used to align large language models with human preferences. This paradigm has been extended to image and video generation via reward models or preference data (Zhang et al. 2024; Wallace et al. 2024; Liu et al. 2025). Recently, Direct Preference Optimization (DPO) (Rafailov et al. 2023) has gained traction in audio-driven animation. Hallo4 (Cui et al. 2025a) proposes a DPO framework for human-centric animation, leveraging a curated human preference dataset to align generated outputs with perceptual metrics related to motion-video alignment and facial expression naturalness. EchoMimicV3 (Meng et al. 2025) further adopts an alternating Supervised Fine-Tuning (SFT) and DPO training paradigm, enabling high-quality video generation with a 1.3B-parameter model. Building on these advances, we present a more efficient framework that integrates DPO to simultaneously enhance lip synchronization accuracy and facial expression naturalness in audio-driven video generation.

3 Methodology

Our method generates high-quality talking videos and enables efficient multi-character animation from a single image, text prompt, and audio clip. The overall framework is illustrated in Figure 2. Built upon the Wan2.1 video diffusion model, we propose a DiT-based architecture enhanced with a LoRA-based training strategy to support long-duration video generation (Section 3.1). We further introduce a partial-update training approach with reward feedback to improve visual fidelity (Section 3.2) and lip synchronization accuracy. Finally, we present a training-free solution, named Mask-CFG, for efficient and cost-effective multi-character audio-driven animation (Section 3.3).

3.1 LoRA-based Long Video Generation

The framework of our method is illustrated in Figure 2, where Wan2.1 serves as the foundational model. Specifically, we employ the causal 3D Variational Autoencoder (VAE) (Kingma, Welling et al. 2013) from Wan2.1 to compress both the reference image and the ground-truth video from pixel space to the latent space. Additionally, we use UMT5 (Chung et al. 2023) for text encoding and CLIP (Radford et al. 2021) for image encoding. For audio input, we utilize Wav2Vec (Baevski et al. 2020) to extract audio tokens containing rich multi-scale acoustic features, which are then injected into the DiT through cross-attention mechanisms.

HunyuanVideo-Avatar (Chen et al. 2025) uses the Time-Aware Position Shift Fusion method from Sonic (Ji et al. 2024) to enable long video generation. OmniAvatar (Gan et al. 2025) reuses the final latent of the current segment

as the initial latent for the next, and applies reference image embedding to preserve identity and maintain frame overlap for temporal consistency. Our experiments show that these two methods fail to achieve satisfactory performance in long video generation. This issue stems from the special architecture of video diffusion models, such as Wan2.1, which are designed to support joint training on both video and image data. In particular, given an input video $V \in \mathbb{R}^{(1+T)\times H\times W\times 3}$, where the frames of V follow the $1+T$ input format, Wan2.1 divides the video into $1+T/4$ chunks. Then, Wan-VAE compresses the spatio-temporal dimensions of these chunks to $[1+T/4, H/8, W/8]$, while the first frame is only spatially compressed to better handle image data. This independent processing of the first frame tends to cause forgetting and drifting issues.

To address this issue, we divide the video into $T/4$ chunks and encode each chunk into a single latent representation. Subsequently, we employ the LoRA training approach, which enables the model to efficiently adapt to long video generation while maintaining high-quality output and low computational cost during training. Notably, we do not add audio cross-attention layers at this stage; instead, we apply LoRA training only to the self-attention and cross-attention modules within the Wan2.1 DiT blocks.

3.2 Partial-update Training and DPO

Audio Cross-Attention. After completing the first LoRA-based training stage, we obtain a diffusion transformer capable of seamless long video generation. Next, we introduce the Audio Cross-Attention module and adopt the Flow Matching (Lipman et al. 2022) approach used in Wan2.1 to update its parameters. Specifically, we aggregate every four consecutive audio frames into a single representation to ensure temporal alignment between the audio features and the compressed video latent representation. The Audio Cross-Attention mechanism is defined as:

$$z' = \text{CrossAttn}(z_v, z_a) = \text{Attn}(Q_v, K_a, V_a), \tag{1}$$

where $z_v \in \mathbb{R}^{b\times f\times(w\times h)\times c}$ and $z_a \in \mathbb{R}^{b\times f\times l\times c}$ denote the video and audio tokens, respectively. Here, f, h and w represent the number of frames, height, and width of the latent video representation, while l denotes the sequence length of the audio tokens. Q_v, K_a, and V_a are the video query, audio key, and audio value matrices, respectively.

Finally, we use the following Flow Matching objective to update the parameters of the Audio Cross-Attention module:

$$\mathcal{L} = \mathbb{E}_{z_0,z_1,z_a,t} \left\| v_{\theta_a}(z_t, z_a, t; \theta_a) - v_t \right\|^2, \tag{2}$$

where z_1 denotes the latent embedding of the training sample, and z_0 denotes the initial noise sampled from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The latent variable z_t is linearly interpolated between z_0 and z_1, and its time derivative $v_t = \frac{dz_t}{dt} = z_1 - z_0$ serves as the regression target. The model predicts this velocity as $v_{\theta_a}(z_t, z_a, t; \theta_a)$, where θ_a represents the parameters of the Audio Cross-Attention module, and z_a denotes the audio features used for conditioning.

Reward Feedback. To further improve lip-sync accuracy and align the model with human preferences, we introduce DPO for optimization after completing the aforementioned stages. Hallo4 (Cui et al. 2025a) presents the first audio-driven portrait DPO dataset that captures human preferences in lip-sync and facial naturalness via annotator rankings. Unlike Hallo4, which relies on human annotators to construct the dataset, we introduce DPO in a more efficient and cost-effective manner.

Direct Preference Optimization formulates the alignment of models with human preferences as a policy optimization task, based on pairwise preference data $\mathcal{D} = \{(x, y^w, y^l)\}$, where y^w is preferred over y^l. As shown in Figure 2, for each training sample, we randomly select five segments and employ LatentSync (Li et al. 2024) to compute the Sync-C score for each; the highest-scoring segment is selected as y^w, and the lowest as y^l. Finally, we use the Flow-DPO loss proposed by VideoReward (Liu et al. 2025) to train the model:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} = -\,\mathbb{E}_{y^w,y^l,t}\Bigg[\log\sigma\Bigg(-\frac{\beta_t}{2}\Big(\|v^w - v_{\theta_a}(y_t^w,t)\|^2 \\ - \|v^w - v_{\text{ref}}(y_t^w,t)\|^2 - \|v^l - v_{\theta_a}(y_t^l,t)\|^2 \\ + \|v^l - v_{\text{ref}}(y_t^l,t)\|^2\Big)\Bigg)\Bigg],\end{aligned} \tag{3}$$

where v_{ref} denotes the reference model, initialized from the previously fine-tuned diffusion model; v^w and v^l denote the velocity fields derived from the preferred sample y^w and the dispreferred sample y^l, respectively. Here, $\beta_t = \beta(1-t)^2$, and the expectation is taken over $(y^w, y^l) \sim \mathcal{D}$ and $t \sim [0,1]$. The overall training loss during this stage is:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{diff}} + \lambda\mathcal{L}_{\text{DPO}}, \tag{4}$$

where $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{DPO}}$ denote the losses in Equation (2) and Equation (3), respectively, and λ is set to 0.1.

3.3 Mask-CFG for Multi-Character Audio-driven Animation

After the training stage described above, we obtain a diffusion transformer that achieves accurate lip-sync and strong alignment with human preferences. We now introduce a training-free method to enable multi-character audio-driven video generation. Methods such as MultiTalk (Kong et al. 2025) and HunyuanVideo-Avatar achieve this capability by constructing multi-speaker datasets and modifying the cross-attention mechanism. In contrast, we enable multi-character animation by improving the classifier-free guidance (CFG) mechanism during inference—without any training or model modification—resulting in a simple, efficient, and framework-agnostic approach. Specifically, we propose Mask-CFG, which leverages spatial masks to route audio conditions to specific characters. Given an audio condition set $A = \{a_1, a_2, \ldots, a_n\}$, the corresponding binary mask set is defined as $M = \{m_1, m_2, \ldots, m_n\}$. Here, a_1 is considered to be silent audio, and m_1 serves as the background mask. Each $m_i \in \{0,1\}^{H\times W}$ denotes a binary segmentation of the input image, and the masks are exhaustive

and mutually exclusive, satisfying $\bigvee_{i=1}^{n} m_i = \mathbf{1}$, meaning their union covers the entire image region. Under classifier-free guidance, the conditional distribution $p(a_i \mid x_t)$ leads to the following formulation:

$$
\begin{aligned}
p(a_i \mid x_t) &= p\left(a_i \,\middle|\, \sum_{j=1}^{n} m_j \odot x_t\right) \\
&= \frac{p(\sum_{j=1}^{n} m_j \odot x_t \mid a_i) p(a_i)}{p(\sum_{j=1}^{n} m_j \odot x_t)} \\
&= \frac{\prod_{j=1}^{n} p(m_j \odot x_t \mid a_i) p(a_i)}{\prod_{j=1}^{n} p(m_j \odot x_t)} \\
&= \frac{p(m_i \odot x_t \mid a_i) p(a_i) \prod_{j=1, j\neq i}^{n} p(m_j \odot x_t)}{p(m_i \odot x_t) \prod_{j=1, j\neq i}^{n} p(m_j \odot x_t)} \\
&= \frac{p(m_i \odot x_t, a_i)}{p(m_i \odot x_t)} \\
&= p(a_i \mid m_i \odot x_t).
\end{aligned}
\tag{5}
$$

Substituting $p(a_i \mid x_t) = p(a_i \mid m_i \odot x_t)$ into the CFG score term $\nabla_{x_t} \log p(a_i \mid x_t)$ and combining it with the standard CFG formula, we obtain:

$$
\begin{aligned}
\hat{v}_\theta(x_t, a, t) &= \nabla_{x_t} \log p(x_t) + \lambda \nabla_{x_t} \log p(a \mid x_t) \\
&= \nabla_{x_t} \log p(x_t) + \lambda \nabla_{x_t} \log \prod_{i=1}^{n} p(a_i \mid x_t) \\
&= \nabla_{x_t} \log p(x_t) + \lambda \sum_{i=1}^{n} \nabla_{m_i \odot x_t} \log p(a_i \mid m_i \odot x_t) \\
&= \nabla_{x_t} \log p(x_t) + \lambda \sum_{i=1}^{n} m_i \odot \nabla_{x_t} \log p(a_i \mid x_t) \\
&= v_\theta(x_t, t) + \sum_{i=1}^{n} \lambda_i m_i \odot [v_\theta(x_t, a_i, t) - v_\theta(x_t, t)].
\end{aligned}
\tag{6}
$$

Through the above Mask-CFG approach, we achieve multi-character audio-driven video generation in a training-free manner, with the visualization workflow shown in Figure 3.

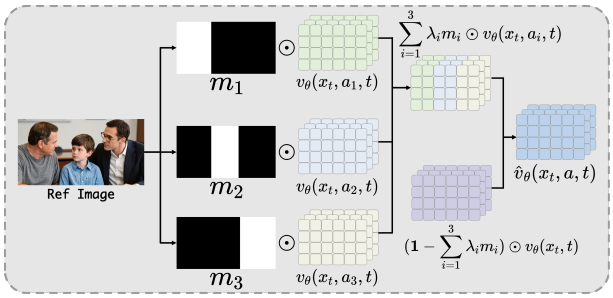

Figure 3: Workflow of Mask-CFG.

4 Experiment

4.1 Experimental Setups

Datasets. We collect our training data from public datasets (including AVSpeech (Ephrat et al. 2018) and OpenHuman-Vid (Li et al. 2025)) and sources we collect ourselves. To ensure high data quality, we employ tools such as Koala-36M (Wang et al. 2025b) to filter out videos with low brightness or poor aesthetic quality. Through this standardized selection process, we obtain over 300,000 training samples, with a total duration exceeding 800 hours. To demonstrate the effectiveness of our multi-character audio-driven approach in a training-free manner, all training samples are single-person talking videos. For evaluation, we use two public datasets: CelebV-HQ (Zhu et al. 2022), which features diverse scenes, and HDTF (Zhang et al. 2021), which provides high-resolution videos and a larger number of subjects, to assess the animation capabilities of our method.

Implementation Details. The training weights of our first LoRA stage are initialized from the pretrained Wan2.1-I2V-14B-720P model, and the training process is conducted using 16 NVIDIA A100 GPUs for a total of 5000 steps. Subsequently, we use 32 NVIDIA A100 GPUs and conduct additional training for 100,000 steps to obtain the first v_{ref}. Next, we introduce DPO-based refinement training for another 100,000 steps, during which v_{ref} is updated every 10,000 steps. The model operates at a resolution of 720×1280. We employ AdamW as the optimizer and set the learning rate to 1×10^{-5}. After completing the above training pipeline, we introduce Mask-CFG during the inference stage to enable multi-person audio-driven animation, with λ set to 5.0.

Evaluation Metrics. We evaluate the superiority of our method using several widely adopted metrics from prior work. Specifically, we employ the Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD) to assess the visual quality and diversity of the generated content. Audio-visual synchronization is measured using Sync-C and Sync-D. Furthermore, we conduct an analysis of both perceptual quality, using Image Quality Assessment (IQA), and aesthetic appeal with the Aesthetic Score Estimator (ASE).

4.2 Results and Analysis

We conduct both qualitative and quantitative evaluations of our method by comparing it with SOTA audio-driven animation approaches, including Sonic, Hallo3, FantasyTalking, HunyuanVideo-Avatar, MultiTalk, and OmniAvatar. Since the work Hallo4 has not yet released its code and models, a direct comparison is not feasible.

Quantitative Results. As shown in Table 1, our method significantly outperforms existing approaches in FID and FVD across both test datasets. On the HDTF benchmark, we achieve the best results in all image and video quality metrics (FID, FVD, IQA, ASE) and performs competitively in lip synchronization. On the CelebV-HQ test set, our method achieves the best scores in FID, FVD, and Sync-D, and ranks second in the remaining metrics(IQA, ASE, and Sync-C), with only a marginal gap to the best result. Overall, our method delivers superior quantitative performance compared to current SOTA methods.

Method	HDTF/CelebV-HQ					
	FID ↓	FVD ↓	IQA ↑	ASE ↑	Sync-C ↑	Sync-D ↓
Sonic	46.47/87.61	213.15/232.65	7.53/6.37	4.58/3.11	6.91/5.28	8.57/8.15
Hallo3	33.16/80.17	185.40/159.04	7.96/7.15	4.81/3.76	6.55/4.64	9.01/9.17
FantasyTalking	38.17/78.72	86.89/138.22	7.62/7.16	4.83/3.82	3.56/3.22	11.16/10.14
HunyuanVideo-Avatar	34.80/78.85	175.00/230.41	7.95/7.29	5.13/4.06	7.43/4.81	8.12/8.11
MultiTalk	38.51/77.92	172.02/206.46	8.35/7.24	5.71/3.95	**8.57/5.64**	**6.97**/7.67
OmniAvatar	36.19/82.40	137.19/169.66	8.14/**7.35**	5.35/**4.14**	7.72/5.36	7.66/7.76
Ours (w/o DPO)	29.05/76.25	86.10/152.33	7.94/7.27	5.66/3.99	7.89/5.28	7.53/7.84
Ours (w/ DPO)	**27.63/66.11**	**81.86/133.78**	**8.38**/7.33	**5.96**/4.13	8.15/5.49	7.32/**7.66**

Table 1: Quantitative comparisons of video quality and lip synchronization with other competing methods on two test datasets. The best results are in bold, and the second-best are underlined.

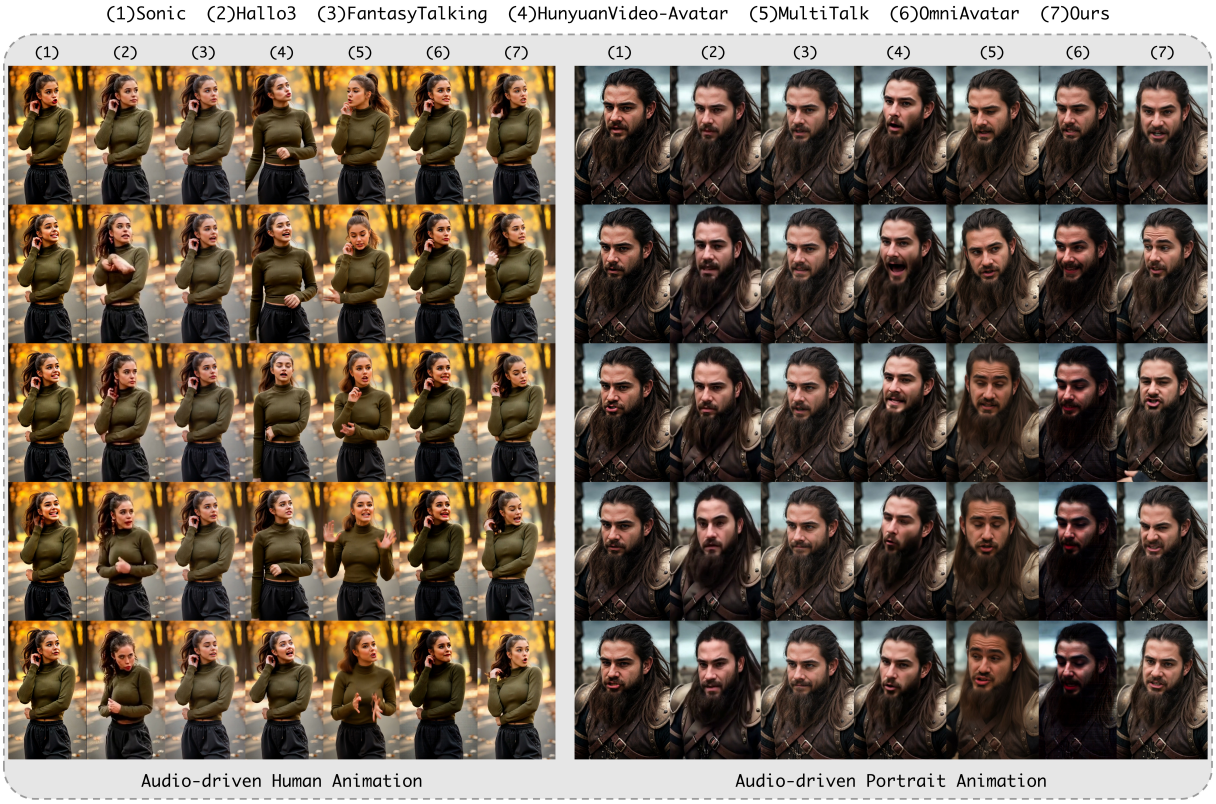

Figure 4: Qualitative comparison with other competing methods.

Qualitative Results. We conducted qualitative comparisons with existing methods. As shown in Figure 4, for human animation, our method generates videos with more natural variations in the foreground, background, and character movements, as well as higher overall quality. In contrast, Sonic, HunyuanVideo-Avatar, and OmniAvatar produce unnatural facial expressions and inaccurate lip synchronization, while FantasyTalking exhibits motion only in the mouth region, with minimal changes elsewhere. Hallo3 and MultiTalk show noticeable artifacts in the face and hands. For portrait animation, Hallo3, HunyuanVideo-Avatar, and MultiTalk fail to maintain character consistency, whereas FantasyTalking animates only the mouth with limited motion in other areas. Sonic demonstrates limited facial expressiveness, and OmniAvatar suffers from severe color distortion. In comparison, our method generates more natural and vivid facial expressions and more aesthetically pleasing visual effects, resulting in superior video quality.

Figure 5: Qualitative comparison of long video generation results.

Methods	LS ↑	VD ↑	N ↑	VA ↑
Sonic	3.50	3.14	3.21	3.21
Hallo3	2.86	2.79	2.79	2.86
FantasyTalking	1.93	2.64	2.57	2.71
HunyuanVideo-Avatar	3.54	3.14	3.05	2.86
MultiTalk	<u>3.93</u>	<u>3.79</u>	**3.93**	<u>3.79</u>
OmniAvatar	3.71	3.77	3.21	3.29
Ours	**4.02**	**3.98**	<u>3.90</u>	**4.11**

Table 2: User Study results. The best results are in bold, and the second-best are in underlined.

User Study. To further validate the effectiveness of our proposed method, we conducted a user study with 50 participants, who rated the videos using a 5-point Mean Opinion Score (MOS) scale across four critical dimensions: Lip Synchronization (LS), Video Definition (VD), Naturalness (N), and Visual Appeal (VA). As shown in Table 2, our method achieves higher scores in LS, VD, and VA. Although the naturalness score is slightly lower than that of MultiTalk, it still significantly outperforms all other methods, demonstrating competitive performance. This comprehensive evaluation highlights the superiority of our approach in generating realistic and diverse talking animations while maintaining consistent identity representation and high visual fidelity.

4.3 Ablation Studies

Ablation on Long Video Generation. We conducted ablation experiments on the LoRA-based long video generation method described in Section 3.1. Specifically, we trained a model without incorporating the improvements outlined in that section, and then generated long videos using the approaches from OmniAvatar and HunyuanVideo-Avatar. As illustrated in Figure 5, the final latent extension strategy used in OmniAvatar suffers from error accumulation over time, leading to significant degradation in the quality of the generated long video. The Time-aware Position Shift Fusion method employed in HunyuanVideo-Avatar produces visible artifacts in the transition regions due to the special input format of the DiT backbone. In contrast, our method generates temporally coherent and identity-consistent long videos, effectively preserving both visual fidelity and temporal smoothness.

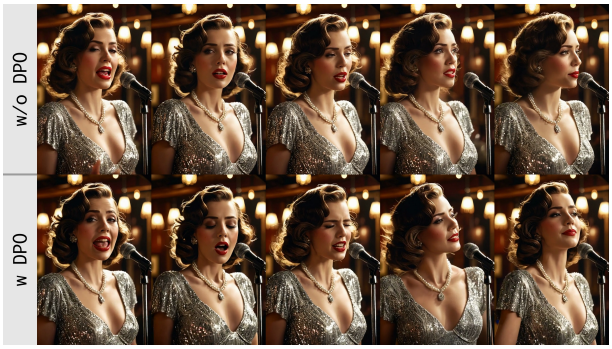

Figure 6: Visual comparison of DPO ablation study.

Ablation on the Reward Feedback. We train models with and without DPO to evaluate the Reward Feedback method (Section 3.2) both quantitatively and qualitatively. As shown in Table 1, incorporating DPO leads to consistent improvements across all metrics, indicating enhanced video quality and lip synchronization accuracy. Qualitatively (Figure 6), the model with DPO generates rich, context-appropriate facial expressions for singing audio, while the ablated version produces flat and under-expressive results. These results demonstrate that our DPO-based approach improves not only fidelity and synchronization but also expressiveness, yielding outputs better aligned with human preferences.

5 Conclusion

We present a novel DiT-based framework for high-quality, audio-driven human video generation, addressing key challenges in long-sequence temporal coherence and multi-character animation. Our method enables arbitrarily long video generation via LoRA-based training and a position shift inference technique, preserving temporal coherence, identity consistency, and the integrity of the pre-trained model. To further enhance synchronization and motion naturalness, we introduce a partial parameter update scheme combined with reward feedback, which improves both lip synchronization accuracy and upper-body dynamics. Furthermore, we propose Mask-CFG, a training-free approach for multi-character animation that requires no additional data or model fine-tuning, yet supports audio-driven animation of three or more characters. To the best of our knowledge, this is the first training-free method to enable audio-driven animation for three or more characters. Extensive experiments show that our method surpasses existing SOTA approaches in terms of visual quality, temporal consistency, and scalability.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Chen, Y.; Liang, S.; Zhou, Z.; Huang, Z.; Ma, Y.; Tang, J.; Lin, Q.; Zhou, Y.; and Lu, Q. 2025. HunyuanVideo-Avatar: High-Fidelity Audio-Driven Human Animation for Multiple Characters. *arXiv preprint arXiv:2505.20156*.

Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; and Ma, C. 2024. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*.

Chung, H. W.; Constant, N.; Garcia, X.; Roberts, A.; Tay, Y.; Narang, S.; and Firat, O. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*.

Cui, J.; Chen, Y.; Xu, M.; Shang, H.; Chen, Y.; Zhan, Y.; Dong, Z.; Yao, Y.; Wang, J.; and Zhu, S. 2025a. Hallo4: High-Fidelity Dynamic Portrait Animation via Direct Preference Optimization and Temporal Motion Modulation. *arXiv preprint arXiv:2505.23525*.

Cui, J.; Li, H.; Zhan, Y.; Shang, H.; Cheng, K.; Ma, Y.; Mu, S.; Zhou, H.; Wang, J.; and Zhu, S. 2025b. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21086–21095.

Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W. T.; and Rubinstein, M. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Fei, Z.; Jiang, H.; Qiu, D.; Gu, B.; Zhang, Y.; Wang, J.; Bai, J.; Li, D.; Fan, M.; Chen, G.; et al. 2025. SkyReels-Audio: Omni Audio-Conditioned Talking Portraits in Video Diffusion Transformers. *arXiv preprint arXiv:2506.00830*.

Gan, Q.; Yang, R.; Zhu, J.; Xue, S.; and Hoi, S. 2025. OmniAvatar: Efficient Audio-Driven Avatar Video Generation with Adaptive Body Animation. *arXiv preprint arXiv:2506.18866*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5784–5794.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Ji, X.; Hu, X.; Xu, Z.; Zhu, J.; Lin, C.; He, Q.; Zhang, J.; Luo, D.; Chen, Y.; Lin, Q.; et al. 2024. Sonic: Shifting Focus to Global Audio Perception in Portrait Animation. *arXiv preprint arXiv:2411.16331*.

Jiang, D.; Chang, J.; You, L.; Bian, S.; Kosk, R.; and Maguire, G. 2024. Audio-Driven Facial Animation with Deep Learning: A Survey. *Information*, 15(11): 675.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*.

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

Li, C.; Zhang, C.; Xu, W.; Lin, J.; Xie, J.; Feng, W.; Peng, B.; Chen, C.; and Xing, W. 2024. LatentSync: Taming Audio-Conditioned Latent Diffusion Models for Lip Sync with SyncNet Supervision. *arXiv preprint arXiv:2412.09262*.

Li, H.; Xu, M.; Zhan, Y.; Mu, S.; Li, J.; Cheng, K.; Chen, Y.; Chen, T.; Ye, M.; Wang, J.; et al. 2025. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7752–7762.

Lin, G.; Jiang, J.; Liang, C.; Zhong, T.; Yang, J.; and Zheng, Y. 2024. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*.

Lin, G.; Jiang, J.; Yang, J.; Zheng, Z.; and Liang, C. 2025. OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human Animation Models. *arXiv preprint arXiv:2502.01061*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, J.; Liu, G.; Liang, J.; Yuan, Z.; Liu, X.; Zheng, M.; Wu, X.; Wang, Q.; Qin, W.; Xia, M.; et al. 2025. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.

Ma, X.; Cai, J.; Guan, Y.; Huang, S.; Zhang, Q.; and Zhang, S. 2025. Playmate: Flexible Control of Portrait Animation via 3D-Implicit Space Guided Diffusion. In *Forty-second International Conference on Machine Learning*.

Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; and Deng, Z. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv e-prints*, arXiv–2312.

Meng, R.; Wang, Y.; Wu, W.; Zheng, R.; Li, Y.; and Ma, C. 2025. EchoMimicV3: 1.3 B Parameters are All You Need for Unified Multi-Modal and Multi-Task Human Animation. *arXiv preprint arXiv:2507.03905*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, 484–492.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Seawead, T.; Yang, C.; Lin, Z.; Zhao, Y.; Lin, S.; Ma, Z.; Guo, H.; Chen, H.; Qi, L.; Wang, S.; et al. 2025. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*.

Tan, W.; Lin, C.; Xu, C.; Xu, F.; Hu, X.; Ji, X.; Zhu, J.; Wang, C.; and Fu, Y. 2025. Disentangle Identity, Cooperate Emotion: Correlation-Aware Emotional Talking Portrait Generation. *arXiv preprint arXiv:2504.18087*.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, 244–260. Springer.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, M.; Wang, Q.; Jiang, F.; Fan, Y.; Zhang, Y.; Qi, Y.; Zhao, K.; and Xu, M. 2025a. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*.

Wang, Q.; Shi, Y.; Ou, J.; Chen, R.; Lin, K.; Wang, J.; Jiang, B.; Yang, H.; Zheng, M.; Tao, X.; et al. 2025b. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8428–8437.

Wang, X.; Ruan, T.; Xu, J.; Guo, X.; Li, J.; Yan, F.; Zhao, G.; and Wang, C. 2024. Expression-aware neural radiance fields for high-fidelity talking portrait synthesis. *Image and Vision Computing*, 147: 105075.

Xu, M.; Li, H.; Su, Q.; Shang, H.; Zhang, L.; Liu, C.; Wang, J.; Yao, Y.; and Zhu, S. 2024a. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*.

Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024b. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*.

Xue, H.; Luo, X.; Hu, Z.; Zhang, X.; Xiang, X.; Dai, Y.; Liu, J.; Zhang, Z.; Li, M.; Yang, J.; et al. 2024. Human motion video generation: A survey. *Authorea Preprints*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Zhang, R.; Gui, L.; Sun, Z.; Feng, Y.; Xu, K.; Zhang, Y.; Fu, D.; Li, C.; Hauptmann, A.; Bisk, Y.; et al. 2024. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023a. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8652–8661.

Zhang, Z.; Hu, Z.; Deng, W.; Fan, C.; Lv, T.; and Ding, Y. 2023b. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3543–3551.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*, 650–667. Springer.