Let Them Talk: Audio-Driven Multi-Person Conversational Video Generation

Zhe Kong[1,2,3]*, **Feng Gao**[2]*, **Yong Zhang**[2]†, **Zhuoliang Kang**[2], **Xiaoming Wei**[2], **Xunliang Cai**[2], **Guanying Chen**[1], **Wenhan Luo**[3]†

[1]Shenzhen Campus of Sun Yat-sen University [2]Meituan
[3]Division of AMC and Department of ECE, HKUST

https://meigen-ai.github.io/multi-talk/

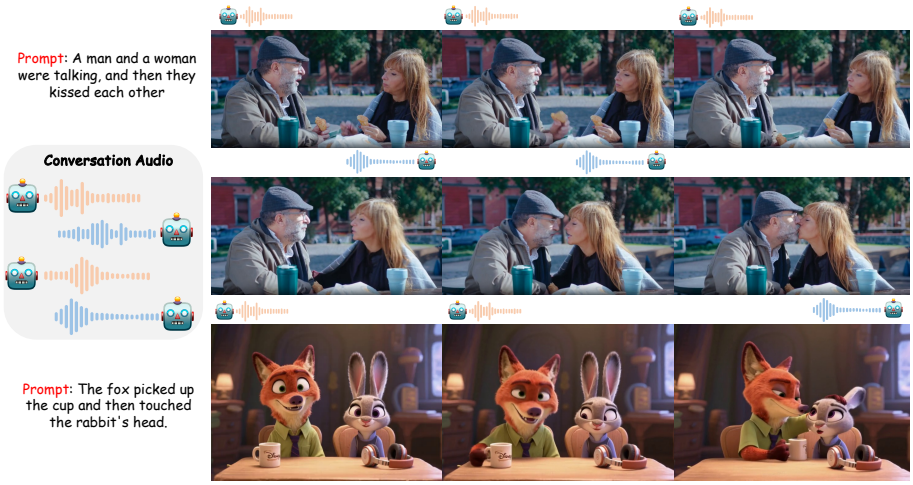

Figure 1: We propose MultiTalk, a novel framework for audio-driven multi-person conversational video generation. Given a multi-stream audio input and a prompt, MultiTalk generates a video containing interactions following the prompt, with consistent lip motions aligned with the audio.

Abstract

Audio-driven human animation methods, such as talking head and talking body generation, have made remarkable progress in generating synchronized facial movements and appealing visual quality videos. However, existing methods primarily focus on single human animation and struggle with multi-stream audio inputs, facing incorrect binding problems between audio and persons. Additionally, they exhibit limitations in instruction-following capabilities. To solve this problem, in this paper, we propose a novel task: Multi-Person Conversational Video Generation, and introduce a new framework, MultiTalk, to address the challenges during multi-person generation. Specifically, for audio injection, we investigate several schemes and propose the Label Rotary Position Embedding (L-RoPE) method to resolve the audio and person binding problem. Furthermore, during training, we observe that partial parameter training and multi-task training are crucial for preserving the instruction-following ability of the base model. MultiTalk achieves superior performance compared to other methods on several datasets, including talking head, talking body, and multi-person datasets, demonstrating the powerful generation capabilities of our approach.

*Equal Contribution.
†Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Audio-driven human animation aims to generate natural and vivid human-centric videos with synchronized facial expressions and body movements from audio control signals. This field has made significant progress recently, and existing methods can be roughly divided into two categories: talking head generation and talking body generation.

Most human animation methods [1, 2, 3, 4, 5, 6] focus on talking head generation. These methods utilize diffusion models to match audio features to visual frames, enabling the synthesis of vivid talking head videos with enhanced video quality and realistic facial expressions. However, they are constrained to achieve precise audio-aligned facial movements and often neglect other related motions, such as hand and body. Recently, several methods [7, 8, 9, 10, 11] have utilized video diffusion models [12, 13, 14] and successfully achieved talking body generation. By leveraging mixed data training strategies or using additional hand pose data, they can synchronize body movements with the audio. Despite these advancements, several constraints remain. Existing methods primarily target single-person animation and cannot handle multi-person scenarios, such as conversational video generation. They lack the capability for dual-stream audio injection. Additionally, they exhibit limitations in instruction-following capabilities. For instance, generated videos may fail to precisely follow instructions when a text prompt describes a large range of body movement.

In this paper, we propose a new task: audio-driven multi-person conversational video generation. This task has diverse applications, including multi-character movie scenes making and e-retailers' livestreaming. Compared to audio-driven single-human animation, this task presents three main challenges: 1) As conversations involve audio from multiple persons, the model should accommodate multi-stream audio inputs; 2) Each person within the conversation should be driven by only one audio stream to prevent incorrect face and audio binding; 3) Each person in the generated video is dynamic, requiring an adaptive method for person localization. Despite the success of existing methods in achieving subtle expressions and realistic motions for a single person, challenges remain in creating multiple-person videos. Specifically, existing methods cannot handle multi-stream input audio and are limited to a single audio stream. Additionally, when reference images contain multiple people, the audio tends to drive all individuals to speak simultaneously, resulting in consistent lip motions across all persons. This complicates the achievement of alternating speech in conversational video.

To complete this new task, we propose a novel framework, MultiTalk, for audio-driven multi-person conversational video generation. Multi-stream audio injection often encounters incorrect binding between the audio and the person. We investigate several schemes for audio injection and introduce the Label Rotary Position Embedding (L-RoPE) method. By assigning identical labels to audio embeddings and video latents, it effectively activates specific regions within the audio cross-attention map, thereby resolving incorrect binding issues. Furthermore, we explore a set of training strategies, including multi-stage training, partial parameter training, and multi-task training. Our observations highlight the importance of the latter two strategies. After incorporating a multi-event dataset for image-to-video, the instruction-following ability of the base model is preserved.

Our main contributions are summarized as follows: (1) We propose a novel task, *i.e.,* audio-driven multi-person conversational video generation, and introduce a novel framework to address the challenges. (2) We investigate several schemes for multi-stream audio injection and propose the Label Rotary Position Embedding method to resolve the inaccurate audio binding problem in multi-person video generation. (3) We explore a set of training strategies, including multi-stage training, partial parameter training, and multi-task training. We observe that the latter two are crucial for preserving the instruction-following ability of the base model, especially with limited compute resources and data. The multi-event dataset for the image-to-video is quite crucial. (4) We conduct evaluations on various datasets, such as talking face, talking body, and multi-person conversation. The results demonstrate the effectiveness of the proposed method.

2 Related Work

2.1 Audio-driven Human Animation

Pioneering audio-driven human animation works [15, 16, 17, 18, 19, 20, 21] typically consist of two components. They first employ an audio-to-motion model to transform motion signals into intermediate representations such as 3DMM [22] and FLAME [23]. Subsequently, motion-to-video

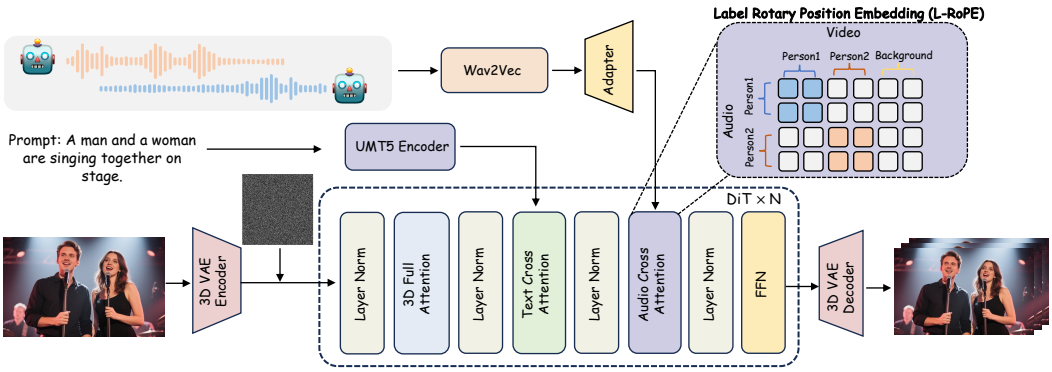

Figure 2: The overall pipeline of the proposed MultiTalk framework. Our framework incorporates an additional audio cross-attention layer to support audio conditions. To achieve multi-person conversational video generation, we propose a Label Rotary Position Embedding (L-RoPE) for multi-stream audio injection.

rendering techniques, such as GANs, are employed to project these intermediate representations into dynamic portrait animations. Despite notable successes, limitations in audio-to-motion models' ability to capture intricate facial expressions and head movements significantly constrain the authenticity and naturalness of synthesized videos.

Recently, end-to-end audio-to-video synthesis methods [1, 24, 2, 4, 3, 25, 5, 6] omit intermediate representation and directly utilize a single diffusion model to integrate audio cues with facial dynamics. These methods demonstrate enhanced potential, exhibiting superior naturalness and consistent portrait animation capability. However, they are constrained to support only head movement. To achieve audio-driven body animation, CyberHost [7] proposes a one-stage audio-driven talking body generation framework equipped with a Region Attention Module and Human-Prior-Guided Conditions to address common synthesis degradations in half-body animation. EMO2 [9] introduces a two-stage framework, first generating hand movements and subsequently using them as control signals in the second stage to enable holistic facial expressions and upper body motions. OmniHuman [8] employs a mixed data training strategy with multimodal motion conditioning to overcome the scarcity of high-quality data. EchomimicV2 [10] proposes an Audio-Pose Dynamic Harmonization strategy, requiring an additional hand pose sequence as input alongside audio. However, these audio-driven human animations can only animate a single person and cannot achieve multi-stream audio-driven image animation.

2.2 Video Diffusion Model

The success of text-to-image diffusion models and their downstream applications [26, 27, 28, 29] has sparked considerable interest in exploring their potential for video generation. Video diffusion models can be roughly divided into two categories: text-to-video models and image-to-video models. Early video diffusion models [30, 31, 12] typically leverage the U-Net architecture for video generation, attempting to extend the 2D U-Net pretrained on text-to-image tasks into 3D to generate continuous video frames. Recent works [32, 33, 14] have adopted a DiT (Diffusion-in-Transformer) architecture [34], significantly advancing video generation technology. These DiT-based methods replace the U-Net with a Transformer, incorporating a 3D VAE as the encoder and decoder. By expanding the training dataset, DiT networks learn motion priors for various objects and scenes. Video diffusion models demonstrate substantial potential in tackling intricate video generation tasks and provide a strong visual backbone for various downstream tasks [35, 36, 37, 38, 39]. Due to its excellent performance in human generation, a DiT-based image-to-video diffusion model is adopted as the backbone of our method to fully leverage its human generative prior.

3 Method

The overall architecture of the proposed method is illustrated in Fig. 2, showcasing an audio-driven multi-person conversational video generation framework. In Section 3.1, we first briefly describe the network architecture of the video foundational model. Then, in Section 3.2, we introduce the

integration of audio conditions via an audio cross-attention mechanism for single-person animation. Subsequently, in Section 3.3, we present our investigation into multi-stream audio injection and introduce the proposed L-RoPE method for audio and person binding. In Section 3.4, we explain our training strategy. Finally, we describe our method for long video generation in Section 3.5.

3.1 Preliminaries

In this study, we adopt a DiT-based video diffusion model as our foundational model, which is built upon the DiT architecture and incorporates a 3D Variational Autoencoder (VAE). This design achieves compression in both spatial and temporal dimensions. A textual encoder is utilized to generate the text-conditioned input, denoted as c_{text}. Additionally, the extracted global context from the CLIP image encoder [40] is injected into the DiT model along with c_{text} via decoupled cross-attention.

3.2 Audio-Driven Single Person Animation

Our foundational model is an image-to-video diffusion model capable of animating a reference image to generate a video. However, it does not natively support audio as an input. To incorporate an additional audio condition, we add layers consisting of layer normalization and an audio cross-attention mechanism after the text cross-attention in each DiT block.

Audio Embedding Extraction To extract acoustic audio embeddings, we employ Wav2Vec [41], a widely utilized audio feature extractor. In audio-driven human animation, since current motion is influenced by both preceding and succeeding audio frames, we follow [1] and concatenate audio embeddings proximal to the current frames, described as follows:

$$a_i = Concat(a_{i-\lfloor \frac{k}{2} \rfloor}, \cdots, a_i, \cdots, a_{i+\lfloor \frac{k}{2} \rfloor}) \tag{1}$$

where k denotes the context length.

In the audio cross-attention layer, queries are derived from video latents, while keys and values originate from audio embeddings. These elements execute frame-by-frame attention calculations. Due to the temporal compression of the 3D VAE, the frame length of video latents is shorter than that of audio embeddings, complicating direct calculations between them. To address this, we propose an audio adapter for audio compression. Specifically, suppose the input audio contains l frames. We first divide the audio embedding into the initial frame a_1 and the subsequent frames $a_{[2:l]}$ along the temporal dimension. Next, we downsample $a_{[2:l]}$ get $Down(a_{[2:l]})$, and then encode a_1 with $Down(a_{[2:l]})$ separately through several MLP layers. After concatenating, we encode the concatenated features to obtain the compressed audio condition c_a. This process is represented as:

$$c_a = MLP(Concat(MLP(a_1), MLP(Down(a_{[2:l]})))). \tag{2}$$

3.3 Audio-Driven Multi-Person Animation

Existing methods fail to address the problem of multi-human generation driven by multi-audio streams. In this paper, we introduce a novel task: audio-driven multi-person conversational video generation. To tackle this challenge, we propose a new framework, MultiTalk, specifically designed to handle multi-stream audio injection and rectify incorrect audio and person binding. The overall architecture of MultiTalk is depicted in Fig.2. We first investigate several schemes for multi-stream audio injection. Then, to accurately identify each person's motion region in generated videos, we propose an adaptive person localization method. Finally, we introduce the proposed L-RoPE method to effectively bind audio and persons.

Multi-stream Audio Injection Schemes. Multi-person conversational video generation, unlike single audio-driven video generation, requires the model to accommodate multi-stream audio inputs. To find an effective method for audio injection, we explore four distinct injection schemes, as illustrated in Fig. 3.

Our first attempt involved directly concatenating the multi-stream audio embeddings z_{a1} and z_{a2}, then calculating the audio cross-attention results with video latent z_t, as shown in Fig. 3 a). Another strategy is to calculate the multi-stream audio embeddings z_{a1} and z_{a2} separately with z_t, and then followed by an adding operation to calculate these two components, as seen in Fig.3 b). However,

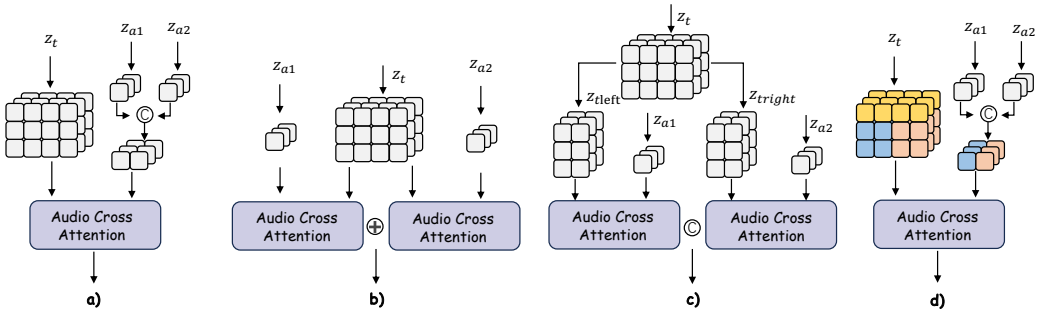

Figure 3: Investigation on different injection strategies for multi-stream audio condition.

these two attempts failed to bind the multi-stream audio with its corresponding video latent region. The network cannot learn to bind audio to different persons through training directly. Given that the individuals in the generated video are typically positioned on the left and right sides, we attempted to simplify binding by splitting the video latents into left and right segments, as demonstrated in Fig. 3c). Each video latent segment computes attention results with the corresponding audio embedding separately, and the two attention results are concatenated as the final output. Although this attempt successfully binds multi-stream audio to different persons, its generalization capacity is limited. Specifically, it is only effective for videos with minimal movement range. When a person exhibits extensive motion, directly applying this simple operation results in audio binding failures. To address these shortcomings, we propose an adaptive method for multi-stream audio injection, named L-RoPE, as illustrated in Fig. 3d).

Adaptive Person Localization. Before utilizing L-RoPE, the model must adaptively track the localization of each individual. Given a reference image I contains two persons, we first find the subject localization within I, resulting in the set $M = \{M_{p1}, M_{p2}, M_b\}$. Here, M_{p1} and M_{p2} represent the mask regions for each person, and M_b denotes the mask covering the background in the reference image. Collectively, they satisfy the relation $I = M_{h1} \cup M_{h2} \cup M_b$. The self-attention map reflects the similarity of generated video latents across different frames. In the I2V model, the first frame of the video also serves as the reference image, enabling the creation of a reference-image-to-video attention map $A_{r2v} \in R^{fhw \times 1hw}$, as depicted in Fig. 4 a). Here, f denotes the frame length in latent space, while h and w represent the height and width, respectively. Since the reference image contains multiple subjects within M, we calculate the average similarity of each latent in z_t with the subjects in the reference image, yielding $S \in R^{fhw \times 3}$. In this matrix, $S(i, j)$ represents the similarity between the i-th token in the video latents and the j-th subject in M. By leveraging the similarity captured in the self-attention map, we can adaptively locate each person in the video.

L-RoPE for Audio and Person Binding. Rotary Position Embedding (RoPE) [42] is a relative positional encoding technique that effectively captures inter-token relationships in large language models (LLMs). Known for its proficiency in modeling long sequences, RoPE has also been employed in video diffusion models, such as CogVideoX [32], Hunyuan Video [33], and Wan [14], among others, to facilitate multi-resolution, multi-aspect ratio, and variable duration video generation. It is utilized to generate position-aware query and key embeddings for time, height, and width within the video latents during the self-attention layer of the DiT block. In this paper, we introduce the Label Rotary Position Embedding (L-RoPE) method, aimed at binding multi-stream audio to multiple persons within the audio cross-attention layers of the DiT block.

Specifically, take the query q as an example. q is a sequence of N vectors $\{q_i\}_{i=1}^{N}$. We compute an angle θ_i for each vector q_i using its label $l_i \in \mathbb{R}$, and rotate q_i with θ_i to obtain $\hat{q}_i$:

$$\theta_i = l_i * \theta_{base} \tag{3}$$

$$\hat{q}_i = LRoPE(q_i, l_i) = q_i e^{l_i \theta_i} \tag{4}$$

where θ_{base} is a pre-defined base angle.

In the audio cross-attention mechanism, queries are derived from the video latent z_t, whereas keys and values originate from the multi-stream audio embeddings z_{a1} and z_{a2}. Appropriately assigning

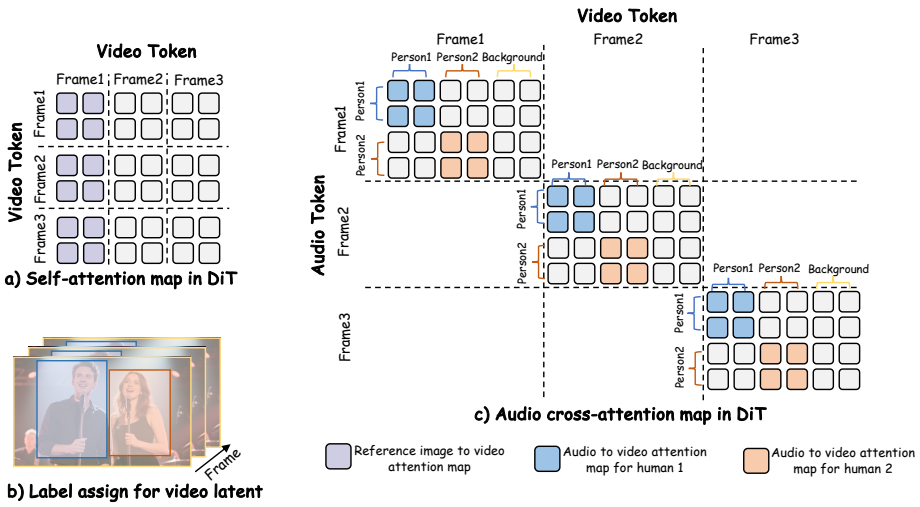

Figure 4: Analysis for different components in the DiT. a) We utilize the reference-image-to-video self-attention map in DiT for person localization. b) We assign different labels to the multiple subjects in the video. c) Assigning a close label for video and audio can activate a specific region in the audio cross-attention map.

Figure 5: Instruction-following capability comparison between different training strategies.

labels l to video and multi-stream audio is crucial. As depicted in Fig. 4b, video latents encompass regions corresponding to multiple persons and the background. We adopt a specific strategy for label assignment. For person regions, due to varying sensitivity driven by audio in different parts of the body, we first assign a numerical range for each person, (a, b). Then, we determine the category $C \in \mathbb{R}^{fhw}$ of each vector in q through $argmax_j(S\left[i,j\right])_{i=1}^{fhw}$. Finally, taking the first person as an example, the label for person1 can be calculated through the normalization function, $Norm(S\left[i,j\right]_{j=C[person1]}, a, b) = \frac{s_{i,j}-min(S_{,j})}{max(S_{,j})-min(S_{,j})} * (b-a) + a$. This method is applied for each person in the same manner, but using different label ranges. Specifically, we define the visual label range as $\{0-4\}$ for the first person and $\{20-24\}$ for the second person. Conversely, for the background and dual audio, they directly utilize a static value as their label. The background should not be associated with audio, hence we assign it the label 12. For multi-audio embedding, as shown in Fig. 3d, we first concatenate the multi-stream audio embeddings and subsequently assign different labels c_{a1} and c_{a2} to them. To bind the multi-stream audio with the two persons respectively, we set c_{a1} as 2 and c_{a2} as 22.

3.4 Training Strategy

Two-stage training. The training stages and associated data sources are essential for achieving effective multi-person animation. We divide the training process into two stages, progressively enhancing the model's capabilities in audio and lip synchronization. The first stage primarily focuses on developing the model's ability to animate a single person. Subsequently, in the second stage, we employ training data that contains dual-stream audio to facilitate multi-human animation.

Partial Parameter Training. In our method, only the network parameters in the audio cross-attention and audio adapter are updated, while all other network parameters are frozen during training. We also compare this strategy with full parameter training. Our findings indicate that network training parameters are crucial; when the compute resources and data are limited, fully parameterized training can lead to not only the degradation in the model's instruction-following ability, especially for motion and interaction, but also cause hand and object distortion. Conversely, training only the audio cross-attention does not result in this issue and the instruction-following ability of the base model can be well preserved.

Multi-task training. During training, we adopt a multi-task hybrid paradigm, dividing model training into multiple tasks, including audio + image to video (AI2V) training and image to video (I2V) training. Different tasks utilize distinct training data while sharing the same network parameters. For AI2V tasks, both the reference image and audio are used as conditions. In the I2V task, the audio condition is removed by zeroing the audio embedding. Additionally, the training data used for the I2V task is unique, comprising mainly of multi-event videos with interactions among human, object, and scene, which is crucial for the alignment between the motion description in the prompt and the generated video.

Multi-task training substantially impacts the results, as shown in Fig.5. Utilizing only talking head and talking body data for AI2V training diminishes the network's instruction-following capability. Conversely, incorporating I2V training allows the model to retain its instruction-following ability.

3.5 Long Video Generation

Although the model can generate video lengths of up to a few frames, this is still insufficient for real-world applications. To address this issue, we introduce an autoregressive-based method to facilitate long video inference. Specifically, within the I2V model, the first frame of the video is typically used as the condition for inference. In contrast, we incorporate the last 5 frames of the previously generated video as additional conditions for inference. Following 3D VAE compression, these conditional frames are reduced to 2 frames of latent noise. We pad zeros to the subsequent frames and concatenate them with latent noise and a video mask. These are then input into DiT for inference, enabling longer video generation.

4 Experiments

4.1 Settings

Datasets. We collect a video dataset of about 2K hours for the first stage training, which covers the face or body of a single talking person. We also collect about 200K video clips that contain multiple events and human-object/environment interactions. The average clip duration is about 10 seconds. For the second stage training, we collect 100 hours of videos consisting of conversations between two persons. For evaluation, we employ three distinct types of testing datasets: the talking head dataset, the talking body dataset, and the dual-human talking body dataset with interactive scenarios. For the talking head dataset, we employ two publicly available datasets, HDTF [43], and CelebV-HQ [44] for evaluation purposes. For the talking body dataset, we utilize the EMTD [10] dataset. Since we are the first to propose a dual-human talking body task, no public dataset is available. We collect a dataset containing 40 videos (referred to as MTHM) sourced from the internet.

Evaluation Metrics. We utilize the commonly used metrics to evaluate the methods. Frechet Inception Distance (FID) [45] and Fréchet Video Distance (FVD) [46] are used to assess the quality of the generated data. Expression-FID (E-FID) is used to evaluate the expressiveness of the facial in the generated video. Sync-C [47] and Sync-D [47] are utilized to measure the synchronization between audio and lip movements.

Implementation Details. We adopted Wan2.1-I2V-14B as the foundational video diffusion model for our experiments. The model is trained using a constant learning rate of $2e-5$, incorporating a warm-up strategy, and optimized using the AdamW optimizer. During training, we only fine-tuned the audio cross-attention layer and adapter while keeping other layers frozen. The proposed method was trained using 64 NVIDIA H800-80G GPUs. In stage 1 of the training process, the batch size was set to 64, whereas in stage 2, the batch size was adjusted to 32.

Table 1: Quantitative comparison with other competing methods on talking head generation, including HDTF and CelebV-HQ datasets.

Methods	HDTF					CelebV-HQ				
	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓
AniPortrait [24]	3.09	10.94	1.32	32.83	112.21	2.09	11.29	1.66	37.17	250.24
VExpress [21]	5.79	8.37	8.92	60.49	200.60	4.30	8.98	10.01	67.34	345.87
Echomimic [4]	5.36	8.99	1.27	60.82	240.07	4.16	9.55	2.87	63.72	318.08
Hallo3 [3]	6.55	8.49	1.12	33.98	153.31	5.57	8.58	1.51	40.81	212.91
Sonic [25]	8.35	**6.43**	1.22	29.53	**89.34**	6.68	7.31	1.85	39.89	224.48
Fantasy Talking [11]	3.61	10.78	1.36	32.64	103.01	3.14	10.43	1.77	37.54	218.43
MultiTalk-single (Ours)	**8.54**	6.69	**1.00**	**24.01**	95.99	7.07	**7.13**	**1.41**	**32.31**	219.19
MultiTalk-multiple (Ours)	8.53	6.81	1.24	27.27	124.06	**7.33**	7.18	1.48	34.08	**184.86**

Table 2: Quantitative comparison with other competing methods on talking body generation, including EMTD dataset.

Methods	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓
Echomimic v2 [10]	6.31	8.41	1.91	35.99	**163.60**
Fantasy Talking [11]	3.32	11.41	1.98	37.68	284.29
MultiTalk-single (Ours)	8.18	**7.28**	1.67	32.05	221.86
MultiTalk-multiple (Ours)	**8.34**	7.30	**1.51**	**31.93**	238.77

4.2 Comparisons with Competing Methods

Quantitative Evaluation. To verify the effectiveness of our method, we compare it with several state-of-the-art human animation methods. For talking head generation, we compare with AniPortrait [24], VExpress [21], EchoMimic [4], Hallo3 [3] Sonic [25] and Fantasy Talking [11]. For talking body comparison, we compare with EchoMimicV2 [10] and Fantasy Talking [11].

Quantitative comparisons, including both talking head and talking body analyses, are presented in Table 1 and Table 2, respectively. Our method surpasses most other approaches across a majority of metrics, exhibiting superior performance in lip synchronization and video quality, which underscores the effectiveness of our approach.

Qualitative Evaluation. To demonstrate the visual effectiveness of the proposed method, we compare and visualize the results alongside some competitive methods, as shown in Fig. 6. Upon providing instructions via a text prompt, only our method successfully responded to the instructions, highlighting its robust instruction-following capability. Additionally, our method generates fewer artifacts in the produced video, attesting to the quality of our approach.

Figure 6: Qualitative comparison with other competing methods.

Table 3: Ablation study about the label range selection in L-RoPE on MTHM dataset.

Variant	Label for video		Label for audio		Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓
	person1	person2	person1	person2					
a)	0–2	2–4	1	3	7.47	7.22	3.22	52.87	506.49
b)	0–4	20–24	2	22	7.56	7.13	3.16	54.20	508.01

As the first method for multi-person generation, there is no directly comparable approach available. We compare our method with the video concatenation technique, which involves generating the left and right video patches separately and subsequently concatenating them. The comparison results are presented in Fig. 7. Our method effectively handles interactive scenarios, avoiding inconsistencies between the left and right segments of the video. Besides, we also visualize the self-attention map for the specific person, highlighted in the red box. Our method can adaptively identify the localization of the person, thereby benefiting the audio binding.

4.3 Analyses

Multi-stream vs Single-stream. Our initial model for multi-stream audio training is derived from a single human animation model. To investigate whether multi-stream audio training would lead to performance degradation, we compared the performance of the single human animation model with multiple human animation models on both the talking head and talking body datasets. The results, presented in Table 1 and Table 2, show that our multiple human animation models achieve performance comparable to that of the single human animation models, indicating that multi-stream audio training does not result in model degradation.

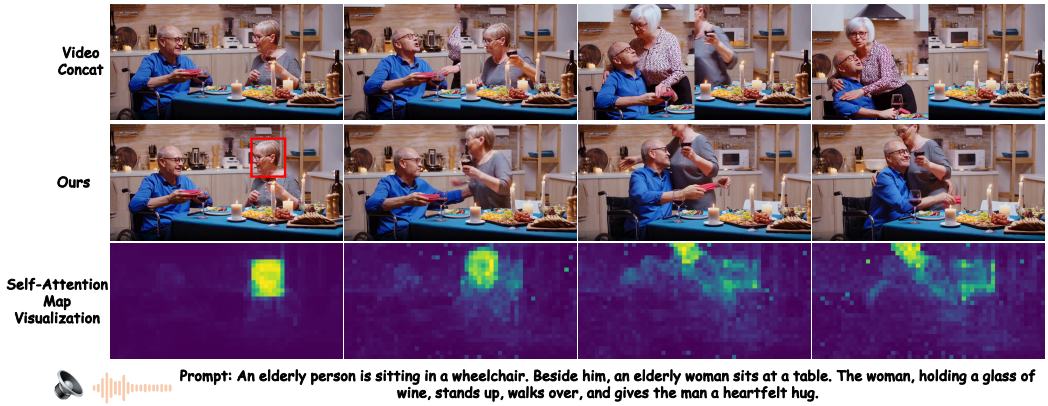

Figure 7: Qualitative comparison with video concat method in multi-human animation.

Label Selection for L-RoPE To validate the effectiveness of L-RoPE within MultiTalk, we conduct an ablation study focusing on label range selection. The evaluation dataset is the collected conversation data, MTHM. The experimental results are presented in Table 3. These results demonstrate that different label choices for various persons yield comparable metrics, indicating that L-RoPE is not sensitive to label range variations.

Table 4: Ablation study for different audio inject strategies (Corresponding to Fig. 3).

	Sync-C↑	Sync-D↓
a	3.49	10.73
b	3.07	11.26
c	7.09	8.00
d (Ours)	**7.56**	**7.13**

Different Audio Injection Strategies We conducted an additional ablation study to investigate the impact of different audio injection strategies. The results are summarized in Table 4, with each row corresponding to an audio injection strategy as illustrated in Fig. 3. Strategies (a) and (b) fail to bind multi-stream audio to the corresponding video latent regions. Strategy (c) employs a hard mask-based audio binding approach, which is capable of associating multi-stream audio with different persons; however, its effectiveness is limited to videos with minimal motion. When a person exhibits extensive

movement, this strategy also results in failure cases. In contrast, our proposed L-RoPE method (d) achieves the best results across all tested scenarios, demonstrating the superiority of our approach.

Table 5: Ablation study for different training strategies.

	Cross-attention Training	Full Parameter Training
MPS↑	**59.5**	40.5

Different Training Strategies To quantitatively evaluate the impact of different training strategies on the model's instruction-following ability and hand/object distortion, we conducted an additional ablation study. Specifically, we utilized a reward model [48] to directly compare the cross-attention training strategy and full parameter training, using the Multi-dimensional Preference Score (MPS) as an evaluation metric. The results, shown in Table 5, demonstrate that training only the cross-attention layers leads to a higher MPS score. This provides clear quantitative evidence that optimizing only the cross-attention layers leads to better performance compared to full-parameter training, particularly when computational resources and data are limited.

5 Conclusion

This paper introduces a novel task: audio-driven multi-person conversational video generation, and presents a new framework, MultiTalk, to accomplish this task. Multi-stream audio conditions are effectively injected using the proposed L-PoRE method, ensuring accurate audio and person binding. Furthermore, our findings demonstrate that partial parameter training and multi-task training are essential for maintaining the instruction-following ability of the base model, equipping our model with powerful instruction-following capability.

Limitation. We observe that our method performs better using real audio than using synthesized audio in terms of facial expression. The reason might be that our model is trained exclusively on real audio, which typically contains rich emotional cues and natural prosody. As a result, the generated videos exhibit more expressive and realistic facial behaviors when driven by real audio. In contrast, most current TTS-generated audio lacks emotional variation and nuanced expressiveness, leading to video outputs that appear less vivid and natural We will explore ways to mitigate the gap between real and synthesized audio for animation in future work.

Societal Impacts This paper introduces an effective approach for audio-driven multi-person conversational video generation to the community. However, this technology also raises ethical concerns. Beyond the risk of generating fake videos of celebrities, there are broader implications, including the potential for misuse in creating deepfakes for misinformation, defamation, fraud, or harassment. Such synthetic videos could be used to impersonate individuals, manipulate public opinion, or violate privacy. These risks are not unique to our approach, but are common across the broader field of human animation and generative models.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372480), in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012839), in part by 2025 Tencent AI Lab Rhino-Bird Focused Research Program, and in part by HKUST-MetaX Joint Lab Fund (No. METAX24EG01-D).

References

[1] 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.

[2] 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.

[3] 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.

[4] 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, 2025.

[5] 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.

[6] 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.

[7] 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*.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.

[14] 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*, 2025.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018.

[23] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Change Loy, and Ran He. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis. *IEEE TCSVT*, 33(3):1247–1261, 2022.

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

[25] 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.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[29] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*, pages 253–270. Springer, 2024.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[35] Sijie Zhao, Wenbo Hu, Xiaodong Cun, Yong Zhang, Xiaoyu Li, Zhe Kong, Xiangjun Gao, Muyao Niu, and Ying Shan. Stereocrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos. *arXiv preprint arXiv:2409.07447*, 2024.

[36] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. *arXiv preprint arXiv:2412.07744*, 2024.

[37] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Towards multiple character image animation through enhancing implicit decoupling. In *ICLR*.

[38] Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. *arXiv preprint arXiv:2408.15239*, 2024.

[39] 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. In *SIGGRAPH*, pages 1–11, 2025.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.

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

[42] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[43] 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.

[44] 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.

[45] 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. 30, 2017.

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

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

[48] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *CVPR*, pages 8018–8027, 2024.

[49] 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, 2025.

[50] 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.

[51] 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.

[52] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, pages 251–263. Springer, 2016.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our our contributions:(1) propose a novel task (2) propose a framework as solution (3)training strategies, are accurately reported in the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We present the implementation details in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: [No]

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are provided in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We use the commonly used metrics that do not involve error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resource information are presented in the experiment part.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the rules.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in the supplementary.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Those works are properly cited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Task Definition

Audio-driven multi-person conversational video generation is defined as follows: Given a reference image containing multiple persons and corresponding audio streams (with a one-to-one correspondence between each person and their audio), the goal is to synthesize a video sequence in which all persons appear together in the same frame, and each person's lip movements are temporally synchronized with their respective audio input. Unlike previous single-person talking face generation tasks, this new task requires the joint modeling of multi-person interactions, spatial consistency, and audio-visual synchronization within a unified, end-to-end generative framework, requiring only a single diffusion process.

In contrast to approaches that generate videos for each person independently and subsequently composite them, this new task demands integrated modeling of multiple individuals, offering several key advantages:

- Higher computational efficiency: Only a single inference process is required, substantially reducing computational costs.
- Global consistency: The unified framework enables better control over the overall coherence of the generated content, such as coordinated camera movements, lighting, and scene dynamics.
- Enhanced interaction modeling: This approach is inherently more suitable for capturing interactions among individuals, enabling natural and contextually appropriate reactions (e.g., when one person is speaking, others can display attentive or responsive behaviors).

B Dataset and Implementation Details

B.1 Dataset Details

In this paper, we utilize three distinct testing datasets: the talking head dataset, the talking body dataset, and the dual-human talking body dataset with interactive scenarios. For the talking head and talking body datasets, we employ conventional evaluation techniques for comparison with other methods. However, for the dual-human talking body dataset, where each reference image contains two persons, we evaluate Sync-C, Sync-D, and E-FID by splitting the video into two segments: the left part and the right part. Each segment contains only one person and their corresponding audio. We then average the scores of these two segments to derive the final result for this dataset. Fig.8 showcases some examples of our dual-human dataset.

Figure 8: Some examples of our MTHM dataset.

All data used in our experiments were collected from publicly available sources on the internet. Our data collection process follows the best practices established by previous works [49, 50, 51], ensuring that our methods are consistent with the standards in the community. All data sources are under the CC BY 4.0 International license. Our dataset comprises approximately 2,700 unique subjects, with approximately 71% male and 29% female. The distributions of age and race are presented in Table 6 and 7, respectively.

Table 6: The distributions of the age of the training dataset.

Age	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70+
Percentage	0%	0.02%	20.77%	57.11%	19.03%	2.95%	0.12%	0%

Table 7: The distributions of race of the training dataset.

Race	white	black	middle eastern	asian	latino hispanic	indian
Percentage	61.55%	6.22%	4.89%	21.76%	4.77%	0.81%

For data preprocessing, we closely follow the procedures described in [49] and further filter out samples exhibiting large facial movements or unsynchronized speech and mouth motion [52]. This ensures the high quality and reliability of our dataset.

B.2 Sample Details

In all the experiments and evaluations conducted within this paper, we utilize 40 sampling steps. To filter out undesired variations in diffusion models, we employ the following negative prompt during sampling: "bright tones, overexposed, static, blurred details, subtitles, style, works, paintings, images, static, overall gray, worst quality, low quality, JPEG compression residue, ugly, incomplete, extra fingers, poorly drawn hands, poorly drawn faces, deformed, disfigured, misshapen limbs, fused fingers, still picture, messy background, three legs, many people in the background, walking backwards." Additionally, we employ Qwen-VL for reference image captioning.

B.3 Inference Time

Although our method introduces an additional audio condition, the computational time required to pass through the DiT backbone remains the same as in Wan2.1, and it requires 40 steps for inference. Furthermore, all acceleration strategies available for Wan2.1—such as TeaCache and model distillation—are also applicable to our approach. For example, when employing a distilled model (such as lightx2v), the total number of inference steps is reduced to 4 steps per video.

C Analyses

C.1 Full Parameter Training vs Cross-attention Training

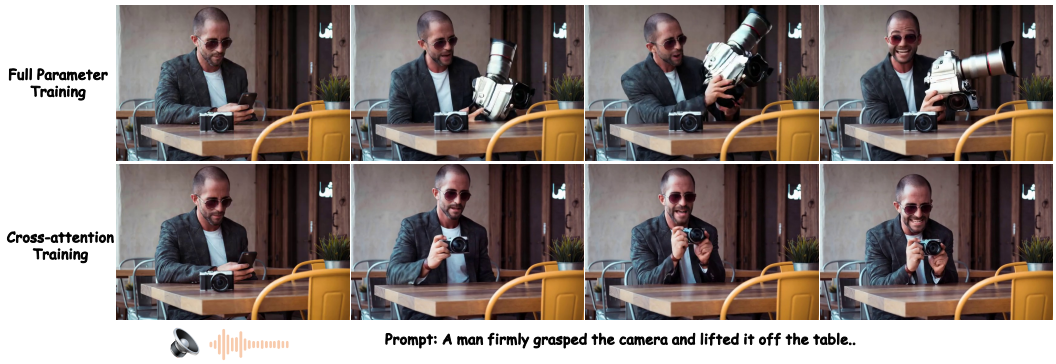

Figure 9: Comparison between full parameter training and cross-attention training.

We compare full parameter training with fine-tuning only the audio cross-attention layer. Our findings indicate that network training parameters are crucial. When compute resources and data are limited, fully parameterized training can lead not only to degradation in the model's instruction-following ability, especially for motion and interaction, but also to hand and object distortion. Conversely, training only the audio cross-attention does not result in these issues, and the instruction-following ability of the base model is well preserved. The comparison results between full parameter training

and cross-attention training are shown in Fig. 9. It can be seen that full parameter training degrades the model's instruction-following ability and causes hand distortion.

C.2 Long Video Generation

Utilizing the autoregressive-based method facilitates the long video generation of our method. The experimental results for long video generation are shown in Fig.10. This example shows a generated result containing 305 frames.

Figure 10: The generation result of long videos.

C.3 Emotional Expressions

We use the same reference image and specify emotional (e.g., angry, sad, happy) via the text prompt. Our model successfully generates videos with the corresponding emotional expressions, as shown in Fig. 11.

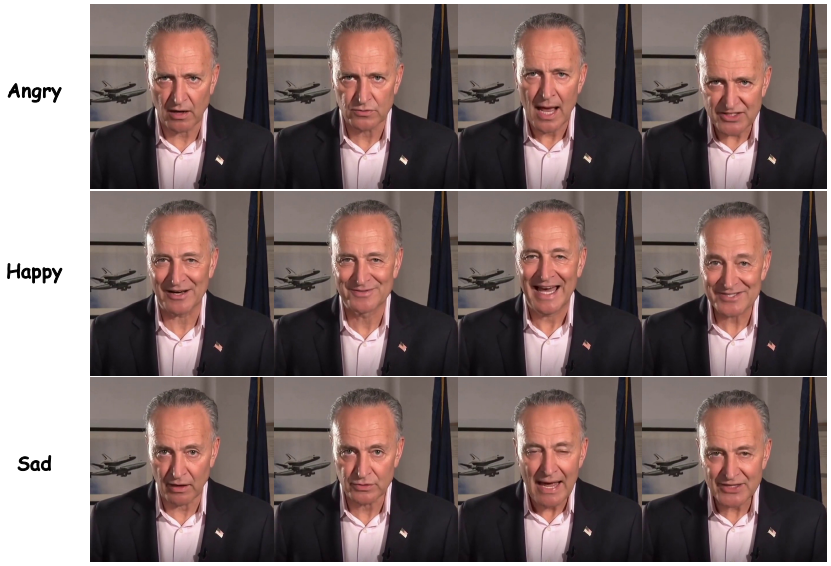

Figure 11: Generation results for videos with different emotions using the same reference image.

C.4 Generalization to More Speakers

In the two-speaker setting, we assign distinct, non-overlapping ranges of video and audio labels to each individual (e.g., video labels 0–4 for person 1 and 20–24 for person 2; audio labels 2 for

person 1 and 22 for person 2). For scenarios involving more individuals, we conducted additional experiments to verify that this labeling scheme can be extended by assigning new, non-overlapping ranges (e.g., video labels 40–44 and audio label 42 for a third person). This flexible approach enables the model to accommodate a greater number of speakers simply by expanding the label ranges for both video and audio streams, which demonstrates the generalizability of our L-RoPE framework. The visualization results for scenarios involving three speakers are shown in Fig. 12. Importantly, the L-RoPE extension strategy remains effective even when the number of persons during inference differs from that during training. By assigning dedicated label ranges, each person's lip movements are accurately aligned with their respective audio streams.

Figure 12: The generation results generalize to scenarios with three persons.