

Bind-Your-Avatar: Multi-Character-Talking Video Generation with Dynamic 3D-mask-based Embedding Router

Supplementary Material

1. Overview of Supplementary Material

This supplementary document provides comprehensive details, additional experiments, and implementation specifics to support the findings in the main paper. The content is organized as follows:

- **Section 2: Additional Experimental Results.** We present a comparative analysis against the concurrent work MultiTalk, a component analysis of our router design choices (L-RoPE, layer-shared, sequential training, training-free alternatives), a user study, mask visualizations, and an evaluation of generalization beyond two characters.
- **Section 3: Additional Model Details.** We provide an in-depth description of the Intra-Denoise Router architecture and detail the loss functions employed, including the geometric priors and weak supervision losses used to enforce spatiotemporal consistency.
- **Section 4: Additional Training Details.** We elaborate on our three-stage training curriculum and the specific teacher-forcing strategy implemented to stabilize the joint training of the router and the diffusion model.
- **Section 5: Additional Data Details.** We outline the rigorous data cleaning criteria and the comprehensive pre-processing pipeline constructed for the MCTV dataset, including segmentation, speech separation, and captioning.
- **Section 6 and 7: Code and Ethical Considerations.** Finally, we provide details on code and resource availability, and discuss the broader implications of our method, including ethical considerations for responsible usage.

2. Additional Experimental Results.

2.1. Comparison with Concurrent Works.

To evaluate the performance of our approach in both single- and multi-character scenarios, we compare it with the recent concurrent work MultiTalk, which also targets multi-character talking video generation and supports single-speaker settings. We report results on AV-Speech (single-speaker), MCTV-Test, and MCTV-I (multi-character).

As shown in Table 1, MultiTalk achieves stronger re-

sults in the single-character AV-Speech setting, which aligns with its substantially larger training corpus. On the multi-character benchmarks (MCTV-Test and MCTV-I), our method attains comparable performance to MultiTalk across most metrics, demonstrating that the intra-denoise routing mechanism is effective for handling multi-person scenarios without requiring character-specific spatial priors.

Overall, while our approach is not superior in the single-speaker case, it remains competitive in multi-character settings and exhibits stable performance across different benchmarks.

2.2. Component Analysis

To further validate the design choices of our embedding router, we conduct a controlled study by integrating alternative designs within our unified framework while keeping all other training hyperparameters consistent. As shown in Table 2, we compare the following variants: (1) **L-RoPE**: replacing our router with the L-RoPE mechanism from MultiTalk [9]; (2) **Layer-shared Router**: sharing a single router across all DiT layers instead of using layer-specific routers; (3) **Sequential Training**: training single-character and multi-character stages sequentially rather than jointly; (4) **Training-Free**: applying SAM2 on the estimated clean frame at an intermediate denoising step to obtain masks without training a router module. Our method consistently outperforms all variants across both MCTV-Test and MCTV-I benchmarks with negligible inference overhead, demonstrating clear architectural advantages for multi-character coordination.

2.3. User Study

To complement the automatic metrics, we conduct a user study with 30 participants on MCTV-Test. Videos from all methods are shuffled and rated on three dimensions: lip-sync accuracy, naturalness, and expressiveness. Scores are normalized to 100. As shown in Table 3, our method achieves the best lip-sync and expressiveness scores, and is competitive with MultiTalk on naturalness, further validating the effectiveness of our approach for multi-character talking video generation.

Table 1. Quantitative comparisons of our method with the concurrent work MultiTalk across single-character (AV-Speech) and multi-character (MCTV-Test, MCTV-I) benchmarks.

Dataset	Model	Metrics				
		Face Sim \uparrow	Sync-C \uparrow	Sync-D \downarrow	FID \downarrow	FVD \downarrow
AV-Speech	MultiTalk[9]	72.66	6.89	7.44	48.6	211.48
	Ours	72.25	6.76	8.15	44.5	233.55
MCTV-Test	MultiTalk[9]	68.27	5.84	8.05	45.55	388.71
	Ours	72.65	6.12	7.45	26.88	298.20
MCTV-I	MultiTalk[9]	72.05	5.68	8.02	/	/
	Ours	71.28	5.78	7.89	/	/

Table 2. Component analysis on MCTV benchmarks.

Model	MCTV-Test					MCTV-I			DiT Time
	Face-Sim \uparrow	Sync-C \uparrow	Sync-D \downarrow	FID \downarrow	FVD \downarrow	Face-Sim \uparrow	Sync-C \uparrow	Sync-D \downarrow	(min)
L-RoPE	72.14	5.98	7.77	26.62	301.56	71.09	5.68	8.02	6.75
Layer-shared Router	71.09	5.68	8.02	27.56	315.78	70.64	5.71	7.95	6.88
Sequential Training	72.45	5.78	7.99	26.67	307.81	70.78	5.44	8.41	6.88
Training-Free	71.10	5.55	8.22	26.21	299.09	70.73	5.31	8.78	7.24
Ours	72.65	6.12	7.45	26.88	298.20	71.28	5.78	7.89	6.88

Table 3. User study on MCTV-Test (scores normalized to 100).

Model	Lip-Sync \uparrow	Naturalness \uparrow	Expressiveness \uparrow
Sonic	46.1	65.4	53.7
Sonic (Concat)	64.8	61.2	55.1
Hallo3	40.5	73.0	70.8
Hallo3 (Concat)	72.3	63.6	71.2
MultiTalk	78.6	83.9	74.5
Ours	79.2	81.4	78.6

2.4. Mask Visualization

To demonstrate the temporal consistency and spatial accuracy of the masks predicted by our Intra-Denoise Router, we visualize the predicted masks across multiple frames in Figure 1. The results show that our router produces smooth, coherent masks that accurately track character regions throughout the video, even under significant head movements and pose changes.

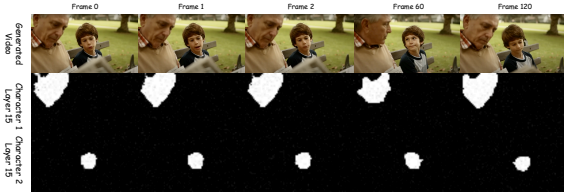


Figure 1. Visualization of the predicted masks across frames. The router produces temporally consistent and spatially accurate masks for each character.

2.5. Generalization Beyond Two Characters

To rigorously assess robustness for complex multi-character scenarios, we curate another challenging test suite of 30 samples with more than 2 characters, including 10 samples each with 3, 4, and more than 4 characters respectively. Quantitative and qualitative results are provided in Table 4 and Figure 2. A key confound in such scenarios is that increasing character count naturally reduces the face ratio and amplifies background complexity, both of which may affect performance independently. To isolate the impact of character count, we design a controlled experiment, referred to “More than 4 Characters (w/ edited)” in Table 4. Specifically, starting from 3-character, we gradually insert irrelevant characters while strictly preserving the appearance and spatial arrangement of target characters.

Empirically, our method sustains strong performance with only mild decreases as the number of character increases. Surprisingly, the controlled “edited” setup shows performance nearly identical to the 3- and 4-character subsets. This finding highlights our approach is robust to larger character counts and the performance drop in unconstrained cases is mainly attributed to lower face ratios and richer scene complexity rather than the character count itself.

3. Additional Model Details.

3.1. Intra-Denoise Router Architecture

Our router module integrates the rich cross-modal representations captured by pretrained facial cross-attention. Specifi-

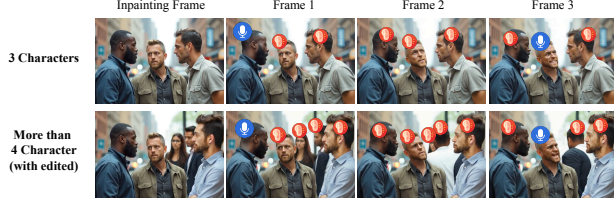


Figure 2. Visualization of inference results on unedited (3 Characters) and edited (More than 4 Characters) inpainting frames. Red microphone and blue ear icons indicate speaking and listening characters respectively. Our method maintains consistent performance as character count increases when confounding factors are controlled.

Test Set	Face Sim \uparrow (%)	Sync-C \uparrow	Sync-D \downarrow	FID \downarrow	FVD \downarrow
3 Characters	67.26	6.29	7.99	29.87	322.56
4 Characters	66.56	6.22	8.25	29.45	329.18
> 4 Characters	64.22	5.60	9.12	30.07	333.59
> 4 Characters (w/ edited)	66.78	6.12	7.95	29.58	327.18

Table 4. Quantitative comparisons in multi-character-talking video generation task across subsets with different number of characters. The best results are in **Bold**.

cally, we utilize the query and key outputs from a pretrained multi-head face cross-attention module as inputs to the router module. Both query and key undergo layer-wise linear transformations followed by reshaping operations that preserve the head dimension. The attention weights, computed via matrix multiplication between transformed queries and keys, encode the correspondence between facial features and visual tokens. These weights are enhanced with 3D RoPE positional encoding [12] before processing through multiple spatio-temporal attention layers that model interdependencies among visual tokens. A final linear layer with softmax activation generates the predicted mask output. This design holistically captures both isolated feature-token correspondences and spatio-temporal token dependencies, which are essential for producing precise, smooth masks.

3.2. Loss Function for Intra-Denoise Router

The Intra-Denoise Router predicts masks by modeling correlations between facial embeddings and visual tokens, and is trained with cross-entropy and weakly supervised losses. Unlike [5], we treat the background as an additional class ($n + 1$) in a multi-class classification framework, reducing the influence of noisy background outputs during inference. Given visual tokens \mathbf{v} from transformer layer l and multiple character reference images \mathbf{I}_r , the router predicts a 3D mask \mathbf{M} , and is mainly optimized by cross-entropy loss:

$$\mathcal{L}_r = \sum_{l,i,t,h,w}^{L,n,T',H'/\tau,W'/\tau} -y_{l,i,t,h,w} \log(\hat{y}_{l,i,t,h,w}), \quad (1)$$

where $y_{l,i,t,h,w}$ and $\hat{y}_{l,i,t,h,w}$ are the ground-truth extracted from SAM2 [10] and predicted mask, respectively, with l, i, t, h , and w indexing the layer, character, time, and spatial dimensions.

Compared to [5], which treats each visual token as an isolated entity and struggle to produce a smooth mask, we incorporate **geometric priors** which naturally constrain the shape of 3D mask to enhance mask smoothness and consistency, followed by three constraints: (1) Spatiotemporal Consistency, (2) Layer-wise Consistency, (3) Identity Exclusivity. In light of the geometric priors, we further design some weak supervision losses to regularize the training of the router module:

$$\mathcal{L}_{st} = \sum_{l=1, i=1}^{L, n} \|\nabla_{\text{Spatiotemporal}} M_{l,i,t}\|_1, \quad (2)$$

$$\mathcal{L}_{\text{layer}} = \sum_{i=1, t=1, h=1, w=1}^{n, T', H'/\tau, W'/\tau} \text{Var}(\{M_{l,i,t,h,w}\}_{l=1}^L), \quad (3)$$

where $\nabla_{\text{Spatiotemporal}} M_{l,i,t}$ denotes the discrete spatiotemporal gradient of the mask \mathbf{M} . We observed that the loss term \mathcal{L}_r inherently enforces Identity Exclusivity, ensuring that only one character is predicted at any given temporal-spatial position. The overall loss to optimize the router module is defined as $\mathcal{L}_{\text{router}}$:

$$\mathcal{L}_{\text{router}} = \mathcal{L}_r + \lambda_{st} \mathcal{L}_{st} + \lambda_l \mathcal{L}_{\text{layer}}, \quad (4)$$

where λ_{st} , and λ_l are the loss weights.

4. Additional Training Details.

4.1. Multi-Stage Training

We divide the training process into three consecutive stages to progressively enhance the model’s capabilities, from identity preservation to audio-driven motion control, and finally to multi-character audio-driven animation.

Stage 1: Excluding speech audio conditions, we focus on identity preservation and the ability to generate from an inpainting frame. The full parameters of the transformer and face embedding extractor are unfrozen. We introduce a 50% probability of randomly dropping the inpainting frame, forcing the model to generate robust outputs in both conditional and unconditional scenarios. Following [13], a Dynamic Mask Loss with a 50% application rate is adopted.

Stage 2: Audio conditions are added, and Low-Rank Adaptation (LORA) [7] is incorporated into the video DiT architecture as an additional trainable module. All parameters are fixed except for the audio encoder, audio cross-attention, face encoder, face cross-attention, and the DiT’s LORA. Random drops of different condition combinations

are applied to strengthen classifier-free guidance generation [6].

Stage 3: The embedding router module is introduced and trained jointly with the entire denoising process. All parameters are fixed except for the embedding router module, audio cross-attention, face cross-attention, and the DiT’s LORA. We adopt a two-task hybrid training paradigm by dividing the training process into single-character and multi-character animation tasks. These tasks use distinct training data while sharing the same network parameters. For single-character scenarios, we replicate the input conditions n times to match the model’s expected multi-character input format. We also employ a teacher-forcing training strategy to stabilize the training process.

4.1.1. Teacher-Forcing Training Strategy

We observed that training the router module jointly with the entire denoising process often leads to a trivial solution, where visual tokens become easy to classify but lack meaningful visual representation. Meanwhile, the router’s training depends on the model being sufficiently adapted to the data distribution to perform high-quality denoising. Without accurate router predictions, the diffusion model progressively loses its ability to effectively integrate and utilize the provided conditions.

Thus, we employ a teacher-forcing training strategy, where we force the input of mask-guided cross attention in the transformer diffusion model to be the ground truth mask value, and detach the computational graph of the router module from the denoising process. To enhance noise robustness, we apply a dropout operation and add Gaussian noise to the forced mask values, effectively serving as a form of data augmentation.

4.2. More Implementation Details of Training

Across all training phases, the three stages consist of 10,000, 40,000, and 10,000 steps, respectively, taking a total of 9 days to complete. To enhance classifier-free guidance, the reference image, inpainting frame, and audio are each randomly dropped with a probability of 0.05 during training. Our training sets include: (1) single-character datasets: **AVSpeech** [4] and (2) multi-character datasets: our **MCTV** dataset.

5. Additional Data Details.

5.1. Data Cleaning

We apply a series of automated data cleaning steps to ensure dataset quality—a stage we deem crucial for training. First, we filter raw videos by resolution, duration, and FPS. Then we employ a face detection model [8] to extract bounding boxes and isolate segments containing exactly two characters. Next, we calculate the bounding-box-to-inter-character-distance ratio to exclude segments where the characters’

bodies are too distant. Finally, lip-sync quality filtering is performed. Directly applying Sync-C [3] to segments generates noisy outputs due to overlapping speech from both characters. Consequently, segments with simultaneous dialogue are mistakenly filtered out. To overcome this, we utilize the AV_MossFormer2_TSE_16K model [11] to separate speech tracks and remove irrelevant noise. Sync-C is computed individually for each character’s isolated audio and cropped facial video, then the average of these two scores serves as our filtering criterion.

5.2. Processing Pipeline

To enable multi-character video training, we construct a comprehensive preprocessing pipeline following data cleaning. This pipeline includes segment extraction, speech separation, audio embedding extraction, and caption generation. Specifically, we use SAM2 [10] to obtain per-character segmentation masks, which are downsampled to the latent space and used as ground truth for the routing masks. AV_MossFormer2_TSE_16K is applied to separate speech tracks from overlapping videos, establishing the Audio-Character Matrix A_{ac} that encodes the correspondence between audio and characters. Audio embeddings are extracted from each isolated speech segment using Wav2Vec [1]. Finally, QWEN2-VL [2] is employed to generate descriptive captions for all segments. Figure 3 illustrates the overall automated pipeline.

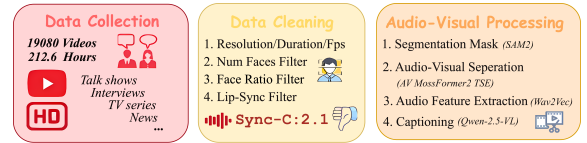


Figure 3. Visualization of the automated data processing pipeline for the MCTV dataset.

6. Code and Resources

To ensure reproducibility, we release our implementation in the supplementary material.

7. Ethical Considerations

Beyond technical contributions, it is crucial to consider the broader implications of our method. Like many generative technologies, it holds the potential to synthesize misleading or deceptive content. We acknowledge this risk and stress the importance of responsible use. We strongly encourage downstream users to adhere to ethical standards, including transparency in synthetic media usage and the development of detection tools to identify generated content.

References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020.
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [3] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Asian Conference on Computer Vision Workshops*, 2016.
- [4] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.
- [5] Zhengcong Fei, Debang Li, Di Qiu, Changqian Yu, and Mingyuan Fan. Ingredients: Blending custom photos with video diffusion transformers, 2025.
- [6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2022.
- [8] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
- [9] Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025.
- [10] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [11] Alibaba SGLab. Clearvoice: An open-source audio-visual speech processing toolkit. <https://github.com/modelscope/ClearerVoiceStudio>, 2025.
- [12] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.
- [13] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Computer Vision and Pattern Recognition Conference*, 2025.