
Supplementary Material: Audio-Sync Video Generation with Multi-Stream Temporal Control

Shuchen Weng^{1†} Haojie Zheng^{1,2†} Zheng Chang³ Si Li³ Boxin Shi^{4,5‡} Xinlong Wang^{1‡}

¹Beijing Academy of Artificial Intelligence

²School of Software and Microelectronics, Peking University

³School of Artificial Intelligence, Beijing University of Posts and Telecommunications

⁴State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University

⁵Nat'l Eng. Research Ctr. of Visual Tech., School of Computer Science, Peking University

{scweng, wangxinlong}@baai.ac.cn, suimu@stu.pku.edu.cn

{zhengchang98, lisi}@bupt.edu.cn, shiboxin@pku.edu.cn

A. Appendix

A.1 Task Differences

To further present the advantages of our MTV framework, we clarify the distinctions between our approach and other audio-relevant tasks. We discuss relevant methods published before the date of this paper submission (May 15th, 2025).

A.1.1 Audible Video Generation

Audio-sync video generation. Our method belongs to the topic of audio-sync video generation, which receives user-provided audios as inputs, offering the potential for free scene creation with optional text descriptions. With the recent advancements in video models, our comparisons focus on very recent methods (*e.g.*, TempoTokens [1] and Xing *et al.* [2]). Since TATS [3] does not provide the custom audio processing, and other methods [4, 5] are tailored for specific visual categories (*e.g.*, landscapes), we select the one with the higher citations [4] for finetuning to general scenarios. Among these, our method is the first to leverage demixed audio tracks for multi-stream control, achieving state-of-the-art performance across six metrics.

Video-audio joint generation. Different from our audio-sync video generation, video-audio joint generation task aims to generate videos with accompanying audios based on user-provided instructions, where most methods in this area are proposed recently [2, 4, 6–8]. Since the audio is co-generated with the video from a shared input (*e.g.*, text descriptions) that typically lacks explicit temporal control signals, *users typically have limited direct control over the precise event timing within the generated video*. While MM-Diffusion [4] discusses training-free strategies to adapt such joint generation methods for audio-sync video generation, our comparison results in Sec. 5.1 indicate that this adaptation approach still has room for improvement.

Summary. Both video-audio joint generation and audio-sync video generation belong to the audible video generation task. Although video-audio joint generation offers advantages in directly producing audible videos, *audio itself is inherently temporal and closely synchronized with the visual world, making it an ideal control signal for precise temporal guidance*. This makes it highly suitable for controllable video generation, unlocking potential applications (*e.g.*, bringing historical recordings to visual life and creating rich visual narratives for podcasts).

[†] Equal contributions. [‡] Corresponding authors.

A.1.2 Audible Image Animation

Audio-driven image animation. Audio-driven image animation aims to generate dynamic visuals from a static image, synchronized with user-provided audios. Most of these methods [9–13] handle general objects and scenarios but still struggle with specific feature synchronization (*e.g.*, for speech and events). Animating talking humans is another sub-topic, which requires a human image to be driven mainly by speech. While most methods in this area [14–17] only focus on the head and facial expressions, a few recent methods [18, 19] extend to half-body or full-body generation. Compared to audio-sync video generation, while the reference image required by these methods allows for animating pre-defined subjects, this reliance may also limit the creation freedom for diverse and dynamic video generation.

Discussion with talking human methods. Since code for both CyberHost [18] and OminiHuman-1 [19] is unavailable, we additionally compare our method with SadTalker [20] and Hallo3 [21]. Since both SadTalker [20] and Hallo3 [21] can only animate the frontal face of a single person, it is infeasible to make a comprehensive evaluation even on our single character subset (as videos for single character also contain many frames without a clear frontal face). Consequently, we provide qualitative comparisons in Fig. S1. These results show that our method effectively demonstrates realistic human gestures and reasonable camera movement. In contrast, Hallo3 [21] presents a more static video (*e.g.*, less gesture and stable background), while SadTalker [20] only modifies the face and pastes the remaining regions directly from the source image. Notably, since both SadTalker [20] and Hallo3 [21] require an additional reference image, we take the reference image as the first frame to leverage our model’s keyframe guidance capability for a fair comparison.



Figure S1: Visual comparison results with state-of-the-art methods for talking human.

A.2 Analysis of MST-ControlNet Depth

As presented in Sec. 4.2, we feed features into N interval interaction blocks within the MST-ControlNet for the interval-wise control. To investigate the impact of this hyperparameter N , we evaluate variants with different depths. The quantitative results presented in Tab. S1 show that increasing N consistently improves the overall visual quality (FVD) and temporal consistency (Temp-C). However, lip motion synchronization metrics demonstrate that they are improved until $N = 4$ before declining. Text-video (Text-C) and audio-video (Audio-C) consistency remain largely stable across different values of N . This suggests a potential trade-off between general video quality and specific lip motion synchronization when varying the depth of interval interaction blocks. Considering this trade-off, we choose $N = 4$ as the setting for our main reported results.

A.3 Metrics Details

As described in Sec. 5.1, we adopt six metrics to quantitatively evaluate performance. We present their details below: (i) Frechét Video Distance (FVD) [22] is used to assess the video quality by computing the distance between feature distributions from real videos and generated videos. (ii) The Temporal consistency (Temp-C) is measured by calculating the cosine similarity between consecutive frame embeddings from the CLIP image encoder [23]. (iii) Text consistency (Text-C) is evaluated

Table S1: Quantitative experiment results of comparison and ablation. \uparrow (\downarrow) means higher (lower) is better. Throughout the paper, best performances are highlighted in **bold**.

Method	FVD \downarrow	Temp-C (%) \uparrow	Text-C (%) \uparrow	Audio-C (%) \uparrow	Sync-C \uparrow	Sync-D \downarrow
$N = 1$	677.51	94.94	26.49	26.32	2.85	9.47
$N = 4$	626.06	95.40	26.55	26.22	3.17	9.43
$N = 8$	570.62	96.09	26.46	26.26	2.74	9.55
$N = 16$	485.84	97.02	26.44	26.25	2.42	9.45

by cosine similarity between text descriptions and generated videos using VideoCLIP-XL [24]. (iv) Audio consistency (Audio-C) is evaluated by cosine similarity between input audios and generated videos using ImageBind [25]. (v) Sync-C and Sync-D [26] are common metrics used to evaluate lip motion synchronization.

Notably, AV-Align [1] is another potential metric for evaluating audio-video alignment. This metric detects energy peaks in audio [27] and motion peaks in video [28], respectively. It then validates whether a peak detected in one modality is also detected in the other within a three-frame temporal window, and vice versa. Although this metric is intuitive and reasonable, it seems unsuitable for evaluating the cinematic videos that our MTV framework focuses on. As shown in Tab. S2, real videos unexpectedly achieve the lowest score with this metric. As a result, we only report this metric in the supplementary materials.

Table S2: AV-Align scores for comparison methods. The higher scores are considered better in theory.

Method	MM-Diffusion [4]	TempoTokens [1]	Xing <i>et al.</i> [2]	Ours (MTV)	Real videos
AV-Align (%)	33.60	33.66	32.49	25.21	23.19

A.4 Dataset Details

Our dataset processing pipeline is illustrated in Fig. S2, with full processing details provided in Sec. 3 of the main paper. Additional dataset samples from our five subsets (*i.e.*, basic face, single character, multiple characters, sound event, and visual mood) are provided in an *anonymous* GitHub link ¹. Each sample includes a video with its corresponding demixed audio tracks, serving to clearly illustrate the concept of ‘audio demixing’.

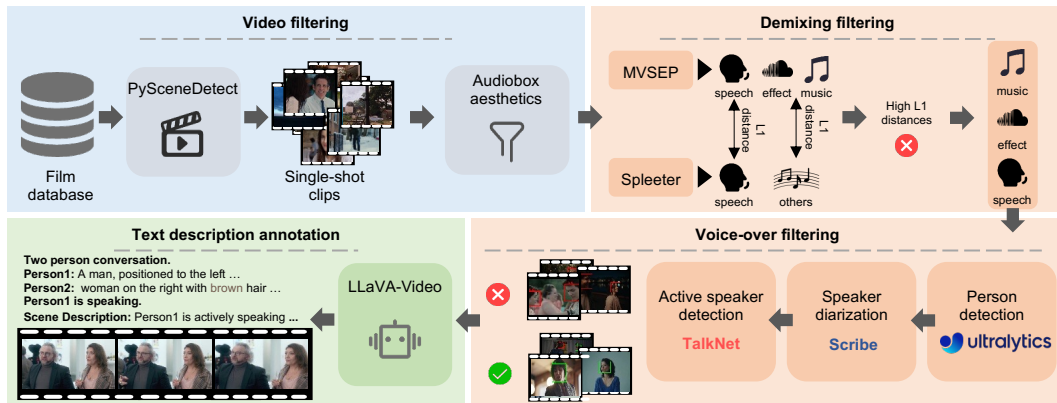


Figure S2: Dataset processing pipeline for our DEMIX dataset.

¹<https://anonymous.4open.science/w/MTV-F4C4/>

A.5 Evaluation on Additional Datasets

To demonstrate the generalization capabilities of our MTV framework, we follow TempoTokens [1] to conduct additional experiments on both the Landscape [29] and AudioSet-Drum [30] datasets. To make a fair comparison, we fine-tune all baseline methods (MM-Diffusion [4], TempoTokens [1], and Xing *et al.* [2]) on our DEMIX dataset using their official training schedules, and evaluate them on these separate datasets. As shown in Tab. S3, our MTV framework still achieves significantly better performance. Since neither dataset includes human talking, the lip synchronization metrics (*i.e.*, Sync-C and Sync-D) are not applicable for this evaluation.

Table S3: Quantitative comparison results in Landscape and AudioSet-Drum datasets.

Method	Landscape				Audio-Drum			
	FVD ↓	Temp-C ↑	Text-C ↑	Audio-C ↑	FVD ↓	Temp-C ↑	Text-C ↑	Audio-C ↑
MM-Diffusion [4]	807.65	94.74	14.66	16.59	1520.09	94.59	14.90	14.11
TempoTokens [1]	797.33	94.67	21.73	18.86	1512.97	94.28	23.18	15.59
Xing <i>et al.</i> [2]	838.03	94.71	21.04	18.70	1589.46	94.49	23.73	17.84
Ours (MTV)	697.51	96.98	25.35	23.37	1511.53	97.50	25.62	39.61

A.6 Organization of Supplementary Video

We provide a supplementary video to dynamically showcase our audio-sync video generation results. The video is structured as follows: (i) **Versatile capabilities across five scenarios.** We demonstrate five generation scenarios to show our capabilities in character-centric narrative, multi-character interaction, sound-triggered events, music-shaped ambiance, and camera movement. (ii) **Application across four typical scenarios.** We present four application scenarios for character creation, keyframe guidance, long video generation, and scene transitions. (iii) **Controllability across four key aspects.** We showcase four aspects to control the generated results, including appearance, lip motion, event timing, and visual mood. (iv) **Comparison with state-of-the-art methods.** We compare with relevant audio-sync video generation methods [1, 2, 4] to demonstrate our superior performance. (v) **Ablation study.** We present the ablation study results to demonstrate the effectiveness of our proposed modules. (vi) **Discussion with talking human methods.** We illustrate the task difference with talking human methods [20, 21], where our method animates humans with more realistic human gestures and reasonable camera movement. For a fair comparison with these reference-based methods, we take the reference image as the first keyframe to leverage our model’s keyframe guidance capability.

References

- [1] G. Yariv, I. Gat, S. Benaim, L. Wolf, I. Schwartz, and Y. Adi, “Diverse and aligned audio-to-video generation via text-to-video model adaptation,” in *AAAI*, 2024.
- [2] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen, “Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners,” in *CVPR*, 2024.
- [3] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh, “Long video generation with time-agnostic vqgan and time-sensitive transformer,” in *ECCV*, 2022.
- [4] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, “MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation,” in *CVPR*, 2023.
- [5] M. Zhao, W. Wang, T. Chen, R. Zhang, and R. Li, “TA2V: Text-audio guided video generation,” *TMM*, 2024.
- [6] Y. Mao, X. Shen, J. Zhang, Z. Qin, J. Zhou, M. Xiang, Y. Zhong, and Y. Dai, “TAVGBench: Benchmarking text to audible-video generation,” in *ACM Multimedia*, 2024.
- [7] A. Hayakawa, M. Ishii, T. Shibuya, and Y. Mitsufuji, “MMDisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation,” in *ICLR*, 2025.

- [8] K. Wang, S. Deng, J. Shi, D. Hatzinakos, and Y. Tian, “AV-DiT: Efficient audio-visual diffusion transformer for joint audio and video generation,” *arXiv preprint arXiv:2406.07686*, 2024.
- [9] S. H. Lee, G. Oh, W. Byeon, J. Bae, C. Kim, W. J. Ryoo, S. H. Yoon, J. Kim, and S. Kim, “Sound-guided semantic video generation,” in *ECCV*, 2022.
- [10] M. Chatterjee and A. Cherian, “Sound2Sight: Generating visual dynamics from sound and context,” in *ECCV*, 2020.
- [11] G. Le Moing, J. Ponce, and C. Schmid, “CCVS: Context-aware controllable video synthesis,” in *NeurIPS*, 2021.
- [12] Y. Jeong, W. Ryoo, S. Lee, D. Seo, W. Byeon, S. Kim, and J. Kim, “The power of sound (TPoS): Audio reactive video generation with stable diffusion,” in *ICCV*, 2023.
- [13] L. Zhang, S. Mo, Y. Zhang, and P. Morgado, “Audio-synchronized visual animation,” in *ECCV*, 2024.
- [14] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang, “Hallo2: Long-duration and high-resolution audio-driven portrait image animation,” in *ICLR*, 2025.
- [15] J. Jiang, C. Liang, J. Yang, G. Lin, T. Zhong, and Y. Zheng, “Loopy: Taming audio-driven portrait avatar with long-term motion dependency,” in *ICLR*, 2025.
- [16] H. Wei, Z. Yang, and Z. Wang, “AniPortrait: Audio-driven synthesis of photorealistic portrait animation,” *arXiv preprint arXiv:2403.17694*, 2024.
- [17] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, “SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation,” in *CVPR*, 2023.
- [18] G. Lin, J. Jiang, C. Liang, T. Zhong, J. Yang, and Y. Zheng, “CyberHost: A one-stage diffusion framework for audio-driven talking body generation,” in *ICLR*, 2025.
- [19] G. Lin, J. Jiang, J. Yang, Z. Zheng, and C. Liang, “OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models,” *arXiv preprint arXiv:2502.01061*, 2025.
- [20] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, “SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation,” in *CVPR*, 2023.
- [21] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu, “Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer,” in *CVPR*, 2025.
- [22] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” *arXiv preprint arXiv:1812.01717*, 2018.
- [23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [24] J. Wang, C. Wang, K. Huang, J. Huang, and L. Jin, “VideoCLIP-XL: Advancing long description understanding for video clip models,” *arXiv preprint arXiv:2410.00741*, 2024.
- [25] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “ImageBind: One embedding space to bind them all,” in *CVPR*, 2023.
- [26] J. S. Chung and A. Zisserman, “Out of time: automated lip sync in the wild,” in *ACCV Workshops*, 2017.
- [27] S. Böck and G. Widmer, “Maximum filter vibrato suppression for onset detection,” Citeseer.
- [28] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Artificial intelligence*, 1981.

- [29] S. H. Lee, G. Oh, W. Byeon, C. Kim, W. J. Ryoo, S. H. Yoon, H. Cho, J. Bae, J. Kim, and S. Kim, "Sound-guided semantic video generation," in *ECCV*, 2022.
- [30] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.