AV-DiT: Efficient Audio-Visual Diffusion Transformer for Joint Audio and Video Generation

[1,♣]**Kai Wang,** [2,♠]**Shijian Deng,** [3,†]**Jing Shi,** [1,♡]**Dimitrios Hatzinakos**[*]**,** [2,♢]**Yapeng Tian**[*]
[1]University of Toronto, [2]University of Texas at Dallas, [3]Adobe Research
♣kaikai.wang@mail.utoronto.ca, ♠shijian.deng@utdallas.edu
†jingshi@adobe.com, ♡dimitris@comm.utoronto.ca, ♢yapeng.tian@utdallas.edu

Abstract

Recent Diffusion Transformers (DiTs) have demonstrated impressive capabilities in generating high-quality single-modality content, including images, videos, and audio. However, the potential of transformer-based diffusers to efficiently denoise the Gaussian noises towards superb multimodal content creation remains underexplored. To bridge this gap, we introduce AV-DiT, a novel and efficient audio-visual diffusion transformer designed to generate high-quality, realistic videos with synchronized visual and audio tracks. To minimize model complexity and computational costs, AV-DiT utilizes a modality-shared DiT backbone pre-trained on image-only data, with only lightweight, newly inserted adapters being trainable. This shared backbone facilitates the generation of both audio and video. Specifically, the video branch incorporates a trainable temporal attention layer into a frozen pre-trained DiT block for temporal consistency. Additionally, a small number of trainable parameters adapt the image-based DiT block for audio generation. An extra shared self-attention block from the DiT block, equipped with lightweight parameters, facilitates feature interaction between audio and visual modalities, ensuring alignment. Extensive experiments on the AIST++ and Landscape datasets demonstrate that AV-DiT achieves state-of-the-art performance in joint audio-visual generation with significantly fewer tunable parameters.

1 Introduction

In recent years, diffusion models [15] have emerged as powerful generative technologies, significantly advancing AI-generated content creation, including images [36, 35], videos [14, 39, 16], and audio [23, 27, 8]. However, most existing research focuses on generating single-modality content, overlooking the multimodality inherent in real-world perception. For example, many video diffusion models [3, 1] generate silent videos, lacking the crucial auditory component. Our work aims to bridge this gap by generating high-quality videos with accompanying audio.

For generating both audio and video content, a straightforward approach involves a two-stage process: generating silent videos using video diffusion models first, followed by generating corresponding audio based on these videos. However, this serial computation is time-consuming. Recent works like MM-Diffusion [37] and Seeing and Hearing [45] propose a more integrated approach, generating video and audio simultaneously using convolutional U-Net-based diffusion models. These frameworks require two separate large generation modules, increasing computational complexity. Given the inherent correlation between audio and visual modalities and the strength of existing image generation models, we question *whether a shared image generation module could be adapted for a joint audio-visual generation with minimal additional learnable parameters?*

Inspired by the success of diffusion transformers (DiTs) [33] in generating high-quality content, particularly with Sora's impressive performance in creating high-fidelity long videos [2], we propose AV-DiT, a novel and efficient Audio-Visual Diffusion Transformer. AV-DiT aims to simultaneously

[*]Equal advisement

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

generate high-quality, realistic audio and video with minimal computational cost by adapting pre-trained image DiTs, effectively addressing the aforementioned questions. Specifically, our AV-DiT leverages a shared DiT backbone pre-trained solely on ImageNet, along with lightweight trainable layers (e.g., LoRA and adapters) to extend image generation into joint audio and video generation. Compared with MM-Diffusion introducing an extra super-resolution module and involving full parameter updates, our end-to-end AV-DiT only trains the newly inserted layers while maintaining the DiT backbone frozen, greatly reducing the number of tunable parameters and computing memory.

To adapt the image diffusion backbones to audio and video domains, AV-DiT addresses three challenges: **1.** Image-to-video: Benefiting from the generalized capability in modeling spatial information, the pre-trained image backbone can spatially extract contextual information from each video frame. However, modeling temporal consistency across various video frames is crucial for video generation. Therefore, our AV-DiT equips the frozen image DiT with the spatial-temporal ability by inserting temporal adapters after spatial layers. **2.** Image-to-audio: Due to the large domain gap between image and audio, it is hard for the image backbone to directly learn the harmonic structure of the acoustic spectrogram. To solve this, AV-DiT enhances the image DiT to learn acoustic information by incorporating audio-specific LoRA and adapter; **3.** Multimodal alignment: During multimodal generation, the generation of each modality expects to benefit from the counterpart modality. Fortunately, the original self-attention block from image DiT can globally connect tokens from different modalities. To further facilitate the audio-video alignment, AV-DiT injects LoRA layers into the self-attention block of image DiT to shift the attention from close-domain token connection to cross-domain interaction. Extensive quantitative and qualitative experiments on the Landscape [24] and AIST++ [25] datasets demonstrate the effectiveness and efficiency of our proposed approach.

The main contributions of this work are as follows: **1)** We propose AV-DiT, the first multimodal diffusion transformer architecture for joint audio and video generation by leveraging an off-the-shelf frozen DiT pre-trained on image-only data and minimal trainable adapters; **2)** Our AV-DiT adapts the shared frozen pre-trained DiT for video generation by introducing temporal consistency, for audio generation by mitigating the domain gap, and for audio-visual alignment by multimodal interaction, demonstrating that the joint audio-video generation can benefit from a pre-trained image generator; **3)** Extensive experiments on the Landscape and AIST++ datasets show that our AV-DiT achieves competitive or even better performance than recent state-of-the-art methods in generating high-quality video and audio with fewer trainable parameters.

2 Method

In this section, we propose an audio-visual diffusion transformer (AV-DiT) by leveraging the off-the-shell DiT backbone pre-trained on image-only data, aiming to make full use of its generalization ability to reduce the training cost and model complexity as shown in Figure 1.

Video and Audio Latent Encoding: Our AV-DiT follows the LDMs to adopt the VAE encoder to project the input video and audio into the latent space before undergoing the AV-DiT denoising network. More specifically, the input videos $V \in \mathbb{R}^{B \times M \times H \times W \times 3}$ are first flattened along the batch B and temporal dimension M, and are then extracted by a learned VAE encoder, resulting in video latent features $z_v = \mathcal{E}_v(V)$, where $z_v \in \mathbb{R}^{(B*M) \times \frac{H}{r_v} \times \frac{W}{r_v} \times c_v}$, M means the number of video frames, c_v is the video channel and $r_v > 1$ denotes the video downsampling ratio. Meanwhile, the audio waveform is transformed into an image-like mel-spectrogram $A \in \mathbb{R}^{B \times T \times F}$ via the STFT operator which is then passed through a pre-trained VAE encoder to obtain the audio latent code $z_a = \mathcal{E}_a(A)$, where $z_a \in \mathbb{R}^{B \times \frac{T}{r_a} \times \frac{F}{r_a} \times c_a}$, T, F and c_a denote the temporal, frequency, and channel dimension, respectively, and $r_a > 1$ means the ratio of downsampling audio resolution. In our work, we employ the off-the-shell pre-trained VAEs from image LDM [36] and audio LDM [8] to extract the video and audio latent codes respectively and save them locally to reduce the training memory. Furthermore, similar to the ViTs, the z_v and z_a are transformed into the sequence of tokens by respective patch embedding layers, which are then followed by the positional encoding to yield the video $x_v \in \mathbb{R}^{(B*M) \times L_v \times D}$ and audio $x_a \in \mathbb{R}^{B \times L_a \times D}$ inputs of our proposed multimodal diffusion transformer. In addition, the time information is incorporated into the diffusion transformer blocks via a time embedder. Note that, in our proposed diffusion transformer, we adopt the frozen visual patch embedding and time embedding layers from the pre-trained image-based DiT while employing a trainable patch embedder for specifically addressing the audio modality.

AV-DiT Block Design: Once the audio and video inputs are obtained, a sequence of proposed AV-DiT blocks is adopted to perform the joint diffusion and denoising processes. In each AV-DiT block, the

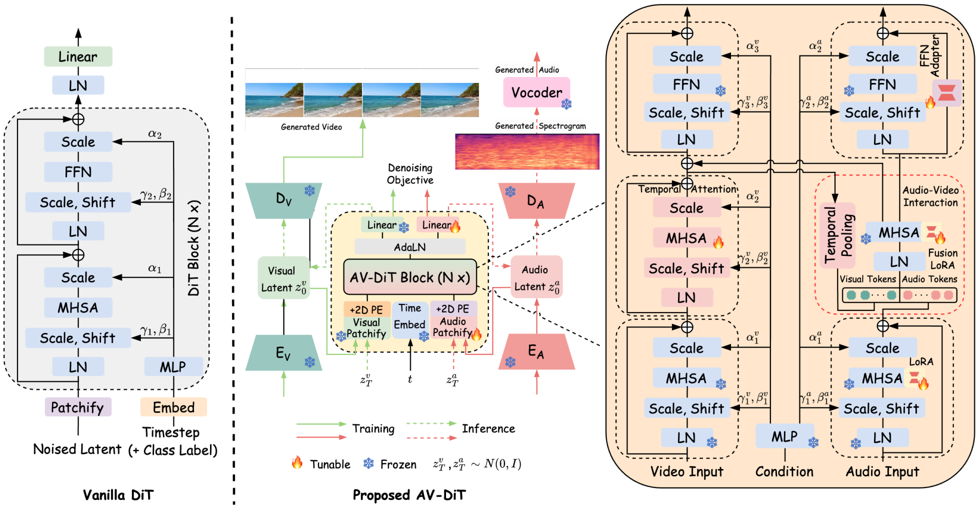

Figure 1: Illustration of vanilla DiT (left) and our proposed AV-DiT (right) for joint audio and video generation. Our AV-DiT leverages a shared frozen DiT backbone pre-trained on image-only data to simultaneously generate high-quality and realistic audio and video, where only inserted modality-specific adapters are trainable while the original pre-trained weights are frozen.

video x_v and audio x_a inputs are normalized by a shared adaptive layer normalization (AdaLN) [34] that regresses the scale $< \gamma_1^v, \gamma_1^a >$ and shift $< \beta_1^v, \beta_1^a >$ parameters from MLP conditioning block to introduce the time guidance. Then, the normalized video features undergo a frozen pre-trained multi-head self-attention (MHSA) to learn the spatial correlation within each video frame along L_v dimension'. Meanwhile, audio features share the same frozen pre-trained MHSA since the learnt attention weights have been explored to be an effective initialization for the audio domain [11, 26].

Bridging Audio-Visual Domain Gap. However, to further alleviate the domain gap between audio and image modality, we inject trainable LoRA layers into projection modules of frozen MHSA to transfer the knowledge from the image into the audio domain. Afterwards, a pair of scaling parameters $< \alpha_1^v, \alpha_1^a >$ from the conditioning block is used to control the information flow prior to the residual connection. It is worth noting that the weights of the conditioning block are inherited from the one of frozen DiT to provide consistent information guidance for both audio and video diffusion.

Temporal Adaptation. To enhance the frozen DiT's ability to model temporal dependencies, we insert a trainable temporal adapter after the frozen spatial MHSA block. This adapter, structurally identical to the frozen MHSA, ensures temporal consistency in video generation. Specifically, the output video features from the frozen MHSA are permuted into $x_{v,f} \in \mathbb{R}^{(B*L_v)\times M\times D}$ by swapping the M and L_v dimensions, which are then passed through the temporal adapter to learn the temporal dependency among different video frames. For efficiency, the temporal adapter enables the feature compression on the query and key projector of the MHSA block, whose compression ratio will be explored in the following section. Additionally, factors regressed from the frozen condition block incorporate time information into the temporal adapter via adaptive layer normalization.

Multimodal Alignment. To bridge the audio and video branches and learn multimodal alignment for improved joint generation, we first pool the video tokens temporally, then concatenate them with audio tokens, feeding the combined sequence into an MHSA block for mutual interaction. Instead of training MHSA as in [6], we reuse a frozen MHSA block from the pre-trained DiT and augment it with LoRA. This adapts the model's knowledge from the image domain to facilitate multimodal interaction, generating audio-steered video features and video-steered audio features to refine the respective tokens. The refined video and audio tokens are processed by a shared frozen feed-forward network (FFN) from the pre-trained DiT, enabling global feature extraction for each modality. Time conditions are incorporated into each branch via adaptive layer normalization (adaLN). Additionally, a simple learnable adapter (a bottleneck MLP) is connected in parallel with the audio branch's FFN, further adapting the learned image knowledge to benefit audio modeling. Finally, the output video and audio features from the AV-DiT block are passed to the next block for iterative processing.

Video and Audio Latent Decoding: After the final AV-DiT block, the sequences of video and audio tokens are required to be decoded into the predicted noises and diagonal covariance of corresponding modalities. To do so, a shared AdaLN followed by two separate liner decoding layers is adopted to

Table 1: Comparison between our AV-DiT and existing SOTA methods. Since Seeing and Hearing [45] only evaluates 200 samples, we also use the same number of samples for fair comparison. For metrics, we used the same scripts as MM-Diffusion [37]. Note that $*$ means the reproduced results using the released model weights from the authors.

Model	**AIST++**			**Landscape**			**Param.** ↓	**Inference Speed** ↑
	FVD ↓	KVD ↓	FAD ↓	FVD ↓	KVD ↓	FAD ↓		
GroundTruth	8.73	0.0036	8.46	17.83	-0.12	7.51	-	-
DIGAN (ICLR 2022) [48]	119.47	35.84	-	305.36	19.56	-	-	-
TATS-base (ECCV 2022) [7]	267.24	41.64	-	267.24	41.64	-	-	-
MM-Diffusion* (CVPR 2023) [37]	98.69	**18.90**	10.58	186.09	**9.21**	**10.61**	426.16M	0.009 sample/sec
AV-DiT (Ours)	**68.88**	21.01	**10.17**	**172.69**	15.41	11.17	**159.91M**	**0.032** sample/sec
Seeing and Hearing[1] (CVPR 2024) [45]	-	-	-	326.23	9.20	**12.76**	-	-
AV-DiT (Ours), 200 samples	-	-	-	**260.50**	**9.15**	14.15	**159.91M**	**0.032** sample/sec

decode video and audio tokens into corresponding predicted latent codes respectively, generating the original spatial feature layout via the feature rearrangement to obtain the predicted noise and covariance of video and audio for diffusion objective. Note that, the decoding layer from the video is directly from frozen DiT, while the other one is initialized and trained from scratch. Once diffusion forward is finished, two sampled Gaussian noises are separately fed into the trained AV-DiT to progressively perform noise removal, yielding the less noisy audio and video latent code at the last diffusion time step. Next, the decoders from pre-trained VAEs from image LDM and audio LDM are employed to simultaneously reconstruct the video and audio latent features back generated video frames and audio mel-spectrogram. Last but not least, a pre-trained HiFi-GAN [22] is used as a vocoder to transform the audio mel-spectrograms into the audio waveforms.

3 Experiments

Experimental Setups. Following previous work [37], we evaluate our proposed AV-DiT on two high-quality datasets including Landscape [24] and AIST++ [25] for joint audio and video generation. We adopt 16 video frames with the size of 256 × 256 as the visual input and Mel spectrogram converted from a 1.6-second waveform as the audio input. Our AV-DiT leverages the frozen pre-trained DiT XL/2 as the shared backbone. The AdamW [29] optimizer is used to train the model for 100k iterations by setting the learning rate as 5×10^{-4} and batch size as 16. Following common practice [37, 48, 7], we adopt the Frechet Video Distance (FVD) and Kernel Video Distance (KVD) to measure the visual quality of the generated videos and we calculate the Frechet audio distance (FAD) [37] between the pairs of ground-truth and generated audio in the space of latent features extracted by AudioCLIP [13]. Lower scores in these metrics indicate higher quality.

Comparison with SOTA methods. As shown in Table 1, we compare our proposed AV-DiT with existing state-of-the-art methods (i.e. DIGAN [48], TATS [7], MM-Diffusion [37], Seeing and Hearing [45]) on AIST++ and Landscape datasets. In general, our AV-DiT achieves competitive or even superior performance than existing methods on joint audio and video generation while involving significantly reduced trainable parameters. First, as for AIST++ dataset, compared with existing baselines, our AV-DiT achieves the best results in FVD (68.88), presenting impressive performance in generating high-quality videos as shown in Figure 2. Although recent MM-Diffusion obtained a lower KVD score than our AV-DiT, it adopts the full training of audio and video branches and involves much more trainable parameters than ours (159.91M vs 426.16M). In addition, MM-Diffusion uses a two-stage generation that first produces a small resolution and then upscales to a large resolution by an extra pre-trained super-resolution network, while our AV-DiT directly generates the target resolution in an end-to-end manner due to our efficient design. For audio evaluation, our AV-DiT yields a superior FAD score to existing methods without any audio-specific training, demonstrating that our AV-DiT efficiently adapts the pre-trained image diffusion to generate high-fidelity audio. Second, as for the Landscape dataset, our AV-DiT also achieves competitive performance in all evaluation metrics including FVD (172.69), KVD (15.41), and FAD (11.17) with reduced tunable parameters. It is worth mentioning that we generated the same number of videos (200 samples) as Seeing and Hearing for a fair comparison on Landscape. As shown in Table 1, we find that our AV-DiT attains better evaluation scores than Seeing and Hearing in all objective metrics, presenting excellent generative performance of sounding videos. Finally, we compare the performance of our methods and the MM-Diffusion baseline in inference efficiency. From Table 1, we observe that our AV-DiT possesses a three times faster inference speed than MM-Diffusion, showing the efficient generation capability of joint audio and video.

[1] We appreciate the authors of Seeing and Hearing [45] providing us their generated samples on Landscape.

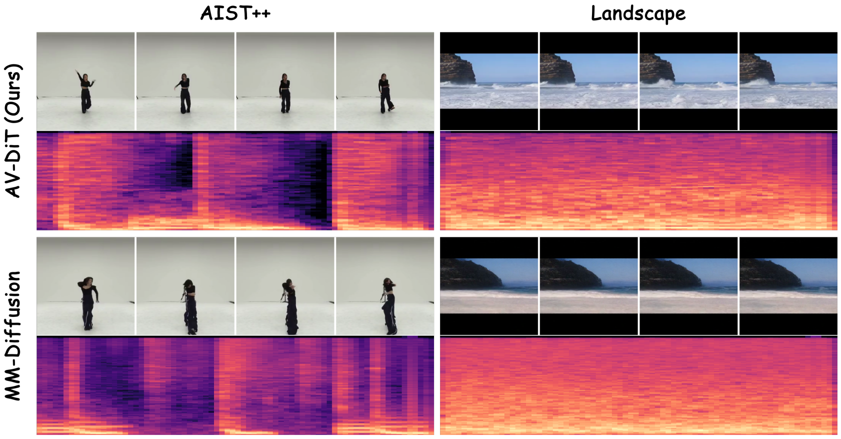

Figure 2: Qualitative examples of our AV-DiT and the MM-Diffusion. Ours generates higher quality and more realistic videos. Meanwhile, our generated audio spectrogram involves fewer artifacts and restores more approximate structures reflecting the visual scenes. For example, our audio sample of Landscape scenes possesses more details that demonstrate the sound of waves lapping on the shore.

Table 2: Influence of various adapter layers

Model	FVD ↓	KVD ↓	FAD ↓
AV-DiT	**68.88**	21.01	**10.17**
w/o Video temporal adapter	365.71	101.33	10.23
w/o Audio FFN adapter	72.81	22.03	10.28
w/o Audio LoRA	69.46	20.80	10.22
w/o Audio LoRA and adapter	74.01	21.63	10.21
w/o Fusion	72.11	**20.39**	11.87
w/o Fusion LoRA	73.62	21.95	10.19

Table 3: Different adapter ratios

Video Ratio	Audio Ratio	Fusion Ratio	FVD ↓	KVD ↓	FAD ↓
8	2	2	68.88	21.01	10.17
4	2	2	67.60	22.68	11.06
2	2	2	**67.13**	22.68	10.11
4	4	2	72.59	23.48	10.15
4	8	2	70.51	**20.70**	10.18
4	2	4	72.71	21.06	10.25
4	2	8	70.10	20.95	**10.03**

Ablation Study. We conducted the following ablation studies to investigate the proposed model.

Influence of Various Adapter Layers: To verify the effects of different adapter layers including LoRA, temporal adapter, and FFN adapter on generative performance, we establish various reference models by removing corresponding layers as shown in Table 2. First, omitting the temporal adapter leads to an obvious performance drop in terms of FVD (from 68.88 to 365.71) and KVD (from 21.01 to 101.33), showing that keeping the temporal consistency is very crucial for video generation. Second, the reference models without either an FFN adapter or LoRA in the audio branch attain a worse performance than our AV-DiT, presenting that audio-specific adapters are useful for the frozen pre-trained image generator to extend to joint audio and video generation. Finally, when removing the LoRA from the fusion module, the performance will be decreased, demonstrating that directly adapting the self-attention of the frozen DiT makes it hard to perform the multimodal interaction.

Different Adapter Ratio: To reduce the model complexity and computing burden, the downsampling operation is applied in each adapter layer. Therefore, we conduct the ablation study to analyze the influence of the choice of adapter ratios on generating performance as shown in Table 3. We find that increasing the ratio of the video temporal adapter will lead to a performance drop in terms of FVD, showing that properly increasing the trainable parameters of the temporal adapter is beneficial for learning the temporal dependency towards better generated videos. Moreover, it is worth mentioning that the best FAD score is achieved by setting a small ratio of audio adapters and a large ratio of fusion adapters. By considering the trade-off between performance and computing complexity, we determine $(8, 2, 2)$ as the ratio for video, audio and fusion part, respectively.

4 Conclusion

In this paper, we propose AV-DiT, the first multimodal diffusion transformer designed for joint audio and video generation. AV-DiT leverages a modality-shared, frozen DiT backbone pre-trained on image-only data along with lightweight trainable layers to generate audio and video jointly. Extensive experiments show that our AV-DiT achieves competitive or even better performance than recent methods while having significantly reduced trainable parameters and enhanced inference efficiency.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[3] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[5] Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[7] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022.

[8] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[10] Litong Gong, Yiran Zhu, Weijie Li, Xiaoyang Kang, Biao Wang, Tiezheng Ge, and Bo Zheng. Atomovideo: High fidelity image-to-video generation. *arXiv preprint arXiv:2403.01800*, 2024.

[11] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[12] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023.

[13] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

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

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[19] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

[20] HPC-AI Technology Inc. Open-sora: Democratizing efficient video production for all. https://github.com/hpcaitech/Open-Sora, 2024. Accessed: 2024-05-20.

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[22] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[23] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[24] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pages 34–50. Springer, 2022.

[25] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[26] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023.

[27] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[28] Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, and Zhou Zhao. Vit-tts: visual text-to-speech with scalable diffusion transformer. *arXiv preprint arXiv:2305.12708*, 2023.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: An empirical study on video diffusion with transformers. *arXiv preprint arXiv:2305.13311*, 2023.

[31] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

[32] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[34] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[37] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023.

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

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

[40] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, volume 1, page 6, 2019.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Kai Wang and Dimitrios Hatzinakos. Moma: Mixture-of-modality-adaptations for transferring knowledge from image models towards efficient audio-visual action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8055–8059. IEEE, 2024.

[43] Kai Wang, Yapeng Tian, and Dimitrios Hatzinakos. Towards efficient audio-visual learners via empowering pre-trained vision transformers with cross-modal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1837–1846, 2024.

[44] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[45] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. *arXiv preprint arXiv:2402.17723*, 2024.

[46] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023.

[47] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024.

[48] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022.

A Supplemental Material

This supplementary material provides some more related works including diffusion transformers, joint audio-video generation, and parameter-efficient generative models. In addition, we summarize the implementation details of the proposed method including architecture configuration and training process. Finally, we present more generated examples to show that our AV-DiT can produce high-quality, realistic videos with audio tracks.

A.1 Related Work

Revist Diffusion Transformers: Diffusion models [15] have presented impressive success in various generative tasks including image synthesis [36, 35, 38], video generation [14, 39, 3, 1, 44], audio generation [27, 8, 19], etc. In general, diffusion models include a forward diffusion process for gradually corrupting real samples to learn a noise predictor and a reverse process to progressively generate less noisy samples via a trained denoising network. To reduce the computational complexity, latent diffusion models (LDMs) [36] employ U-Net architectures to perform the diffusion on the latent space with lower feature dimensions. Recently, the diffusion transformer (DiT) [33] has emerged as an effective replacement for the U-Net backbone in various diffusers like SiT [31], SD3 [6], VDT [30], Latte [32], Sora [2], OpenSora [20], ViT-TTS [28], etc. However, existing transformer-based diffusion models concentrate on generating a single modality, constraining its potential application in multimodal generation. Hence, our AV-DiT is the first work to investigate how to utilize the DiT structure for joint audio and video generation.

A vanilla diffusion transformer (DiT) adopts transformer structure [41] to replace the commonly used U-Net denoiser, which follows the design of latent diffusion models (LDMs) [36] to operate on the latent space for reducing the computational complexity as shown in the left part of Figure 2. More specifically, the DiT first compresses the input image into the latent code z_0 with a reduced feature dimension by an encoder $\mathcal{E}$ from the pre-trained variational autoencoder (VAE). Then, the transformer blocks consisting of multi-head self-attention (MHSA) and feed-forward network (FFN) modules are operated on the latent space to perform the forward diffusion and backward denoising processes [15], where the time embedding and class embedding are incorporated for conditioning the generation by adaptive layer normalization. Note that, the patch embedding layer and position encoding are used to transform the latent code into the input tokens of the transformer.

Following denoising diffusion probabilistic models (DDPMs) [15], a forward noising process is defined to gradually corrupt the real latent z_0 with random Gaussian noise over a discrete time step t, which can be formulated as $q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I})$ where $(\beta_1, \beta_2, \dots, \beta_t, \dots, \beta_T)$ are noising schedule, and T is the pre-defined number of total diffusion steps. Therefore, the forward process generates a series of corrupted samples $(z_1, z_2, \dots, z_t, \dots, z_T)$ with ascending levels of noise via Markovian chain. By the re-parameterization method, the z_t can be directly obtained by $\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\alpha_t = 1 - \beta_t$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To recover the original latent, the reverse process of DiT leverages a learned transformer network θ to predict less noisy data z_{t-1} from the noisy input z_t at each timestep via inverting the forward process, which can be formulated as $p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \sigma_t^2 \mathbf{I})$. The training objective is to minimize a mean squire loss between the noise prediction $\epsilon_\theta(z_t, t)$ and ground truth, defined as follows: $\min_\theta \mathcal{L}_{simple} = \min_\theta \mathbb{E}_{z_t, t, \epsilon} ||\epsilon - \epsilon_\theta(z_t, t)||_2^2$. Afterwards, the data can be generated by progressively sampling z_{t-1} from $p_\theta(z_{t-1}|z_t)$ by a trained denoising transformer, yielding the generated latent code $\hat{z_0}$ which is recovered back to the generated image sample via learned decoder $\mathcal{D}$ from VAE. In this work, we employ the DiT as the diffusion backbone and adapt it to address joint audio-visual generation.

Problem Definition: Existing diffusion models typically perform forward and reverse processes to generate single-modality outputs like images, videos, or audio. However, our work tackles the more challenging problem of jointly generating high-quality audio and video. To simplify the problem, we directly start from the reverse process to introduce how to generate audio and video modalities from the Gaussian noises having the same feature sizes as those of audio and video latent code. Given a paired noise (z_T^a, z_T^v) randomly sampled from Gaussian distribution, the joint audio and video generation aims to train a joint denoising network θ_{av} to fit the reverse process by taking both modalities as the inputs and benefiting the generative quality of counterpart modality. In other words, the reverse process adopts the trained θ_{av} to predict less noisy video and less noisy audio by considering their own and counterpart modalities, which can be formulated as follows:

$$p_{\theta_{av}}(z^a_{t-1}|(z^a_t, z^v_t)) = \mathcal{N}(z^a_{t-1}; \mu_{\theta av}(z^a_t, z^v_t, t), \sigma_t^2\mathbf{I}) \tag{1}$$

$$p_{\theta_{av}}(z^v_{t-1}|(z^v_t, z^a_t)) = \mathcal{N}(z^v_{t-1}; \mu_{\theta av}(z^v_t, z^a_t, t), \sigma_t^2\mathbf{I}) \tag{2}$$

where $t = 1, 2, \ldots, T$ denotes the diffusion steps. The training objective of joint denoising network θ_{av} can be formulated as follows:

$$\mathcal{L}_{\theta_{av}} = \mathbb{E}_{z^v_t, z^a_t, t, \epsilon_v, \epsilon_a} ||\epsilon_v - \epsilon_{\theta_{av}}(z^v_t, z^a_t, t)||_2^2 + ||\epsilon_a - \epsilon_{\theta_{av}}(z^a_t, z^v_t, t)||_2^2 \tag{3}$$

where ϵ_v and ϵ_a mean the ground noises for corrupting video and audio modality in the forward process, respectively.

Our AV-DiT simultaneously generates audio and video by taming a modality-shared DiT pre-trained on image data. It will address three key challenges: 1) Enabling pre-trained Image DiT for video generation; 2) Adapting the pre-trained image DiT for audio generation by mitigating the domain gap; 3) Generalizing the pre-trained DiT for feature interaction or alignment between audio and video modalities[2].

Joint Audio-Video Generation: Unlike video generators that only create silent videos, joint audio-video generation aims to generate high-quality, realistic videos that can be watched and listened to simultaneously. MM-Diffusion [37] is the pioneering work that adopts diffusion models for generating audio-video pairs. It consists of separate video and audio branches to perform the joint multimodal denoising, where a random-shift based attention module is incorporated to learn the consistency between audio and video modalities. Furthermore, Seeing and Hearing [45] introduces a multimodal latent aligner based on ImageBind [9] to align the well-learnt latent features from pre-trained audio and video diffusion models, achieving the cross-modal generation without training the model from scratch. Compared with existing joint audio-video generation models, our AV-DiT leverages a modality-shared frozen diffusion model pre-trained on image data to generate audible videos via introducing lightweight trainable adapters.

Parameter-efficient Generative Models: Training or fine-tuning large diffusion models for downstream tasks is computationally expensive. Therefore, parameter-efficient generative models have been proposed to adopt the parameter-efficient fine-tuning (PEFT) strategy [17, 26, 5, 43, 42] to only tune partial layers while keeping the majority weights of pre-trained frozen. PEFT technologies primarily consist of adapter tunning for inserting bottleneck adapters [17], prompt tuning for injecting learnable prompt tokens at input space [21], and low-rank adaptation (LoRA) for approximating the model weights by a low-rank factorization [18]. To sufficiently leverage the pre-trained image generator to produce video, various lightweight adapters or temporal layers are inserted into frozen LDMs to capture the temporal consistency across video frames [46, 12, 10]. Besides, the trainable LoRA layers can be injected into frozen pre-trained LDM to introduce controllable conditions, guiding the generation of personalized images without relying on full fine-tuning [47]. Different from these works on closed-domain adaptation (i.e. image-to-image or image-to-video), our AV-DiT adapts the frozen image generator to address the joint audio and video generation via lightweight trainable layers (i.e. LoRA or adapters).

A.2 More Implementation Details

Datasets: Landscape is a high-fidelity audio-visual dataset and it features nine different nature scenes including raining, splashing water, thunder, etc. The dataset contains 928 videos crawled from YouTube as listed in [24] and creates 1,000 video clips of 10 seconds without overlap, spanning approximately 2.7 hours in total. **AIST++** is acquired from the AIST dataset [40], featuring dancing videos with 60 copyright-cleared songs. The dataset consists of 1,020 video clips having a total duration of 5.2 hours.

Implementation. We summarize the implementation details in table 4.

Data Pre-processing: Following the work [37], 16 video frames are sampled to construct a video clip and are then cropped into 256 × 256 resolution. Based on the duration of each video clip, we crop out the corresponding audio signal and then truncate or pad it into a 1.6-second waveform at the sampling rate of 16 KHz.

[2] Hereafter, the abbreviation DiT will refer to the diffusion transformer pre-trained on images unless otherwise specified.

Table 4: Model and Training Configurations

Model Configuration	Audio-Visual Diffusion Transformer (AV-DiT)
Architecture	
Feature Dimension	1152
Attention Head	16
Transformer Depth	28
Compression Ratio in Video Temporal Adapter	8
Compression Ratio in Audio LoRA and FFN Adapter	2
Compression Ratio in Fusion LoRA	2
Diffusion Process	
Diffision Steps	1000
Diffusion Noise Scheduler	Linear
Prediction Objective	Noise Prediction
Sampling Method	DDPM
Sampling Steps	250
Input Data	
Video Shape	$16 \times 256 \times 256 \times 3$
Video FPS	10
Audio Spectrogram Shape	160×64
Audio Sample Rate	16K
Training Setting	
Learning Rate	5e-4
Optimizer	AdamW
Batch Size	16
Loss Function	MSE
Training Iteations	100k
Training Hardware	NVIDIA RTX 6000 Ada Generation
Training OS	Linux-5.15.0-101-generic-x86_64-with-glibc2.35

Model Structure: We adopt the off-the-shelf pre-trained VAEs from Stable Diffusion to transform the input video into latent code with the size of $(32 \times 32 \times 4)$. Meanwhile, the STFT operator and pre-trained VAE from audio LDM (i.e. Tango [8]) are used to project the input audio into a latent feature with size of $(40 \times 16 \times 8)$. Our AV-DiT adopts the frozen pre-trained DiT XL/2 backbone as well as lightweight trainable layers to predict the noise during the forward process. Moreover, the used DiT is pre-trained on the ImageNet with the resolution of 256×256 by using 28 transformer layers with 16 attention heads and a feature dimension of 1152.

Diffusion Stage: We retain the diffusion configuration as ADM [4] which uses a linear noise schedule ranging from 1×10^{-4} to 2×10^{-2} via 1,000 time steps. In addition, we train our proposed AV-DiT for 100K iterations with a batch size of 16 and a constant learning rate of 5×10^{-4} via AdamW [29] optimizer. Note that, only newly inserted layers (e.g., adapters, LoRA, audio patch embedding and decoding layers), and bias terms are trainable, while the pre-trained DiT backbone is kept frozen during the training.

Efficient Training: We implement our AV-DiT by PyTorch framework and utilize one Nvidia RTX A6000 (48GB) GPU card for all experiments. To achieve efficient training and reduce the usage of computing memory, we first extract the latent features of audio and video input by using the off-the-shell STFT operator (from Tango [8]), audio VAE (from Tango [8]), and visual VAE (Stable Diffusion [36]). Then, we save the extracted audio-visual features locally and load them for latent denoising, saving remarkable computing memory so that we can run the experiments with only one GPU card. To further enhance the training speed, we adopt mixed precision training and gradient checkpointing during the training process. To present the training efficiency of our AV-DiT, we give the estimated usage of computing memory with the global batch size in Table 5. We can find that our AV-DiT can be trained in an affordable GPU machine.

Table 5: Estimated Costs of Computing Memory

Global Batch Size	Memory Usage per GPU Card	Number of GPU Cards
8	About 14-15GB	1
16	About 30GB	1
24	About 42GB	1

Evaluataion: In our experiments, we randomly generate 2,048 samples by using our trained AV-DiT to calculate objective evaluation scores, where the original generated and real videos are cropped out 64×64 resolution for computation efficiency. For a fair comparison, we evaluate all experiments by averaging 5 runs to reduce randomness.

A.3 More Ablation Studies

Scaling Backbone: When replacing AV-DiT's frozen DiT backbone with one pre-trained on larger image resolutions (i.e., 512x512), Table 6 shows improved performance. This validates that our parameter-efficient AV-DiT benefits from more powerful pre-trained image generators.

Cross-attention Fusion: Our AV-DiT adopts a joint self-attention for connecting the audio and video features for multimodal alignment. To explore the efficiency of our audio-video fusion mechanism, we also design the other reference model using cross-attention based fusion. More specifically, in each AV-DiT block, two cross-attention blocks are separately inserted before the FFN module of each branch for bi-directional conditioning audio and video. As shown in Table 7, we can find that our AV-DiT with self-attention fusion outperforms better than the one with cross-attention fusion while involving lower model complexity and trainable parameters.

Table 6: Different DiT backbones

Model	FVD $\downarrow$	KVD $\downarrow$	FAD $\downarrow$
AV-DiT (256*256 Backbone)	68.88	21.01	**10.17**
AV-DiT (ours) (512*512 Backbone)	**67.26**	**20.02**	10.23

Table 7: Self-attention fusion V.S. Cross-attention fusion

Model	FVD $\downarrow$	KVD $\downarrow$	FAD $\downarrow$	Param.
AV-DiT (ours) (Self-attention) Fusion	**68.88**	21.01	**10.17**	**159.91M**
AV-DiT (ours) (Cross-attention) Fusion	71.57	**20.19**	11.43	289.82M

A.4 Limitations:

Our experiments mainly focus on unconditional audio and video generation. It is meaningful to explore the effectiveness of the proposed AV-DiT in class-conditional and text-conditional generation of audio and video. Notably, our AV-DiT can be flexibly extended into the conditional joint generation of audio and video by injecting additional text prompts or audio-visual captions with minimal adjustment. Although our AV-DiT is more efficient than MM-Diffusion, further improvements are necessary for deployment in real-time applications. Lastly, due to the limited computational resources, our proposed AV-DiT was only implemented on the small datasets used in MM-Diffusion. In future work, we plan to evaluate our AV-DiT on large-scale datasets with real-world scenarios like VGGSound, and further improve the generative efficiency of joint audio and video by accelerating the inference speed.

A.5 Additional Qualitative Examples

For additional results on Landscapes, please check Fig. 3.

For more results on AIST++, please refer to Fig. 4.

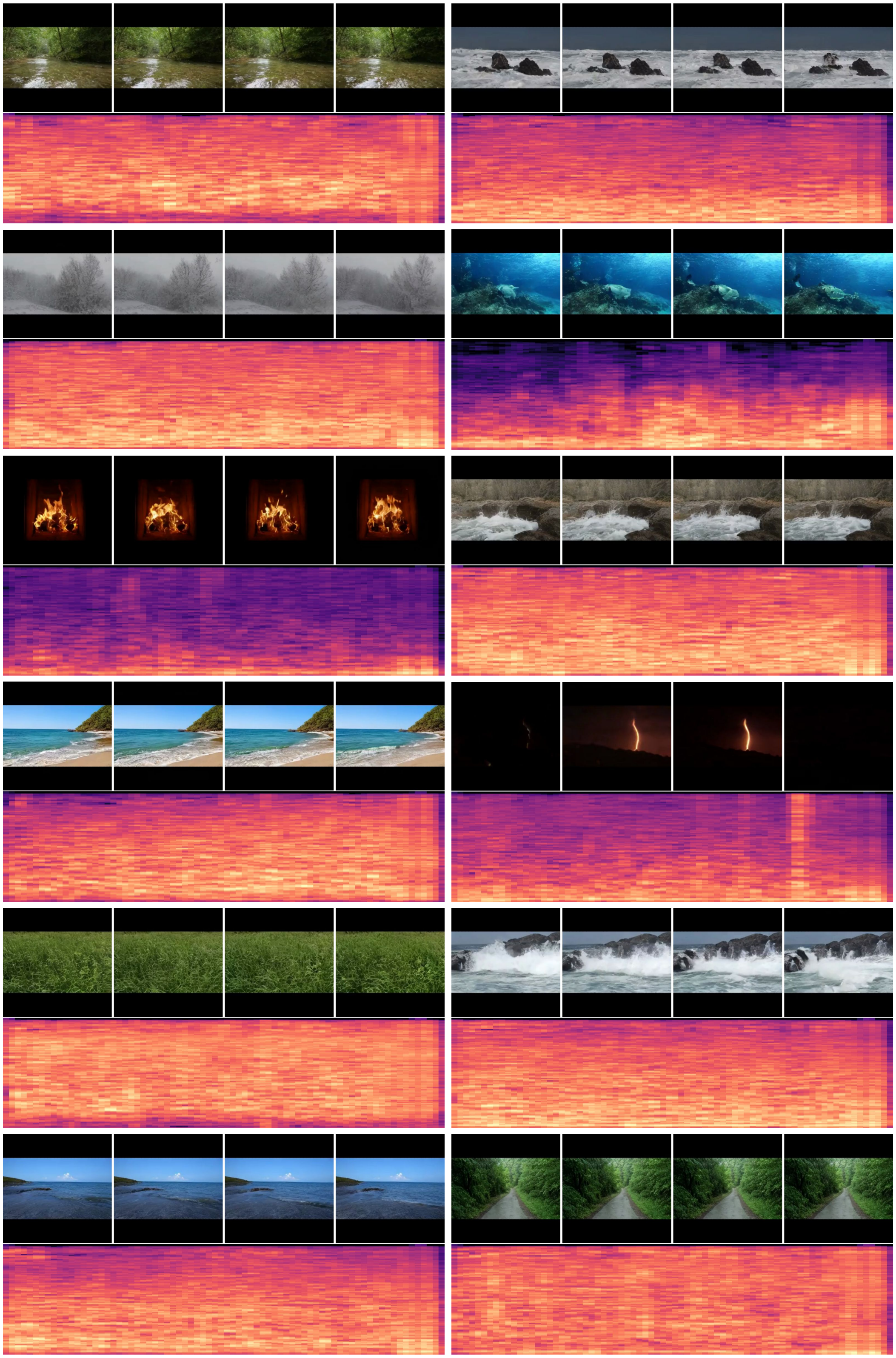

Figure 3: More results from our AV-DiT model on Landscape dataset. The generated audio and video are consistent with each other.

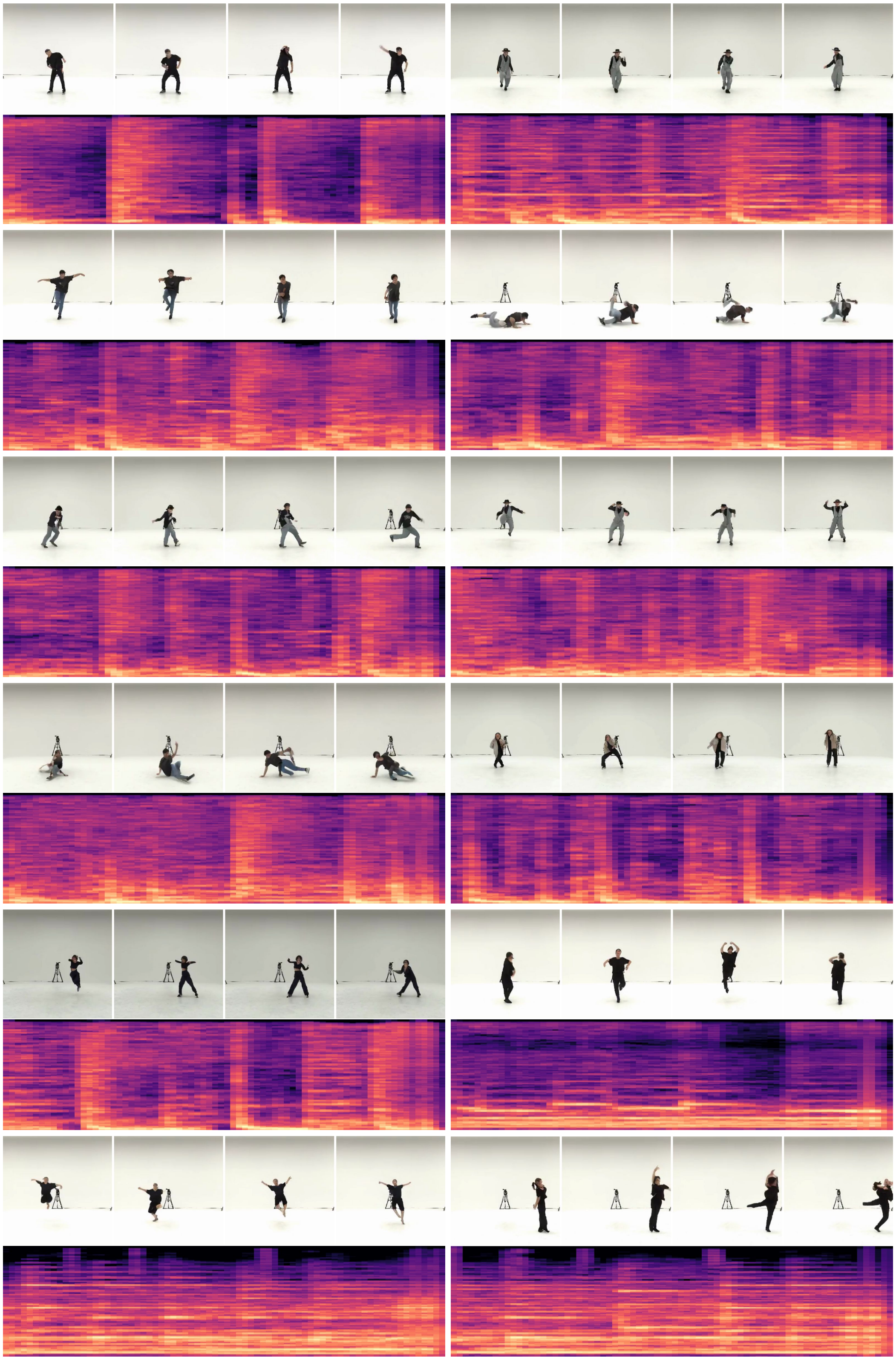

Figure 4: More results from our AV-DiT model on AIST++ dataset. The dance movements and music beats are well aligned.