

MM-LDM: Multi-Modal Latent Diffusion Model for Sounding Video Generation

Appendix

Anonymous Authors

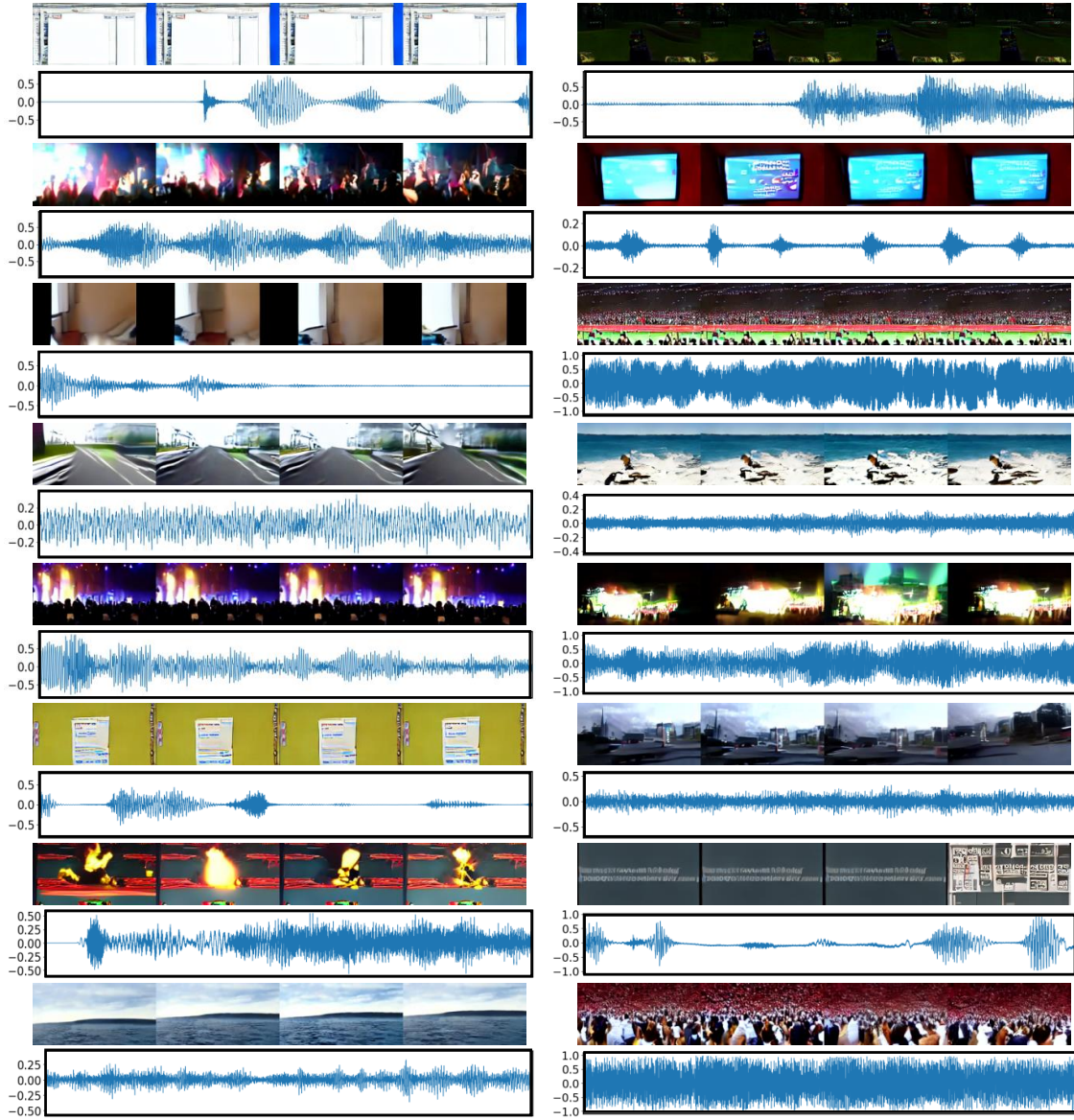


Figure 1: Samples of open-domain sounding video generation on the AudioSet dataset. All presented audios can be played in Adobe Acrobat by clicking corresponding wave figures.

A OPEN-DOMAIN VIDEO GENERATION

To evaluate the performance of our proposed method for open-domain generation, we train it on the AudioSet dataset with 2 A800 GPUs. Samples are presented in Fig. 1. It can be seen that our method can synthesize diverse video content with consistent audios, demonstrating the effectiveness of our method on open-domain sounding video generation.

B LIMITATIONS

Our research exhibits several limitations. First, when unifying the data representation of audio and video inputs, we resize the audio Mel Spectrogram image to the same spatial resolution with video frames, which may lead to information loss to some extent. Second, we utilize a shared KL-VAE and two modal-specific encoders to compress audio and video signals, which significantly reduces the computational complexity but can cause loss of signal information.

C SIGNAL DECODING

We design our signal decoder to utilize the generation power of image diffusion model for synthesizing more details when decoding both audio images and video frames, which is similar to [6]. When reconstructing video signals, the signal decoder takes multiple factors into account, including the video perceptual latent z_v , frame index i , audio semantic feature s_a , learnable modality embedding, and learnable class embedding. The video perceptual latent plays a vital role in guiding the video reconstruction in two ways. Firstly, it provides detailed spatial information for the i -th video frame by utilizing residual blocks. This spatial information is integrated into each residual block within the signal decoder following a zero convolution to provide spatial signal guidance, which is the most basic condition of our method. Secondly, we spatially average pool the video perceptual latent and add it with the timestep embedding to offer global content guidance. To incorporate cross-modal information, we rasterize the audio semantic feature and feed it to the cross attention layers. It is noteworthy that we share signal decoder parameters for both audio and video modalities to reduce computational complexity. To this end, we introduce learnable prompt embeddings specific to each modality for distinction. In addition, given that our classification head can predict class labels precisely, we further define a learnable embedding for each class to incorporate more prior knowledge. Both the modality embedding and the class embedding are concatenated with the audio semantic feature to feed cross-attention layers. When dealing with audio reconstruction, the signal decoder employs similar inputs, except for the frame index. To reduce training time and enhance the quality of reconstruction, we initialize our signal decoder with parameters of a pretrained image diffusion model [4] and open all parameters during training.

D DATASET DETAILS

We conduct experiments on two distinct and high-quality sounding video datasets downloaded from MM-Diffusion [5], namely Landscape [1] and AIST++ [2]. The Landscape dataset comprises 1,000 non-overlapping video clips recording nature scenes. Each video clip is 10 seconds, obtaining a total duration of about 2.7 hours and encompassing 300K frames. These videos are categorized into

nine classes, each corresponding to a different nature scene. The AIST++ dataset is a subset of the AIST dataset [8] and encompasses 1,020 video clips recording street dances. This dataset has a total duration of around 5.2 hours, encompasses street dance videos with 60 copyright-cleared dancing songs and contains a total of 560K frames. Each scene or dancing song is treated as a category in this paper.

E LONG SOUNDING VIDEO GENERATION

For long sounding video generation, we introduce a condition encoder as shown in Fig. 2 and perform experiments on the AIST++ dataset. We first generate an initial sounding video unconditionally. Then the condition encoder takes the generated sounding video as input and outputs a follow-up sounding video. Long sounding videos can be obtained by iteratively performing the conditional generation step. We also consider the audio or video continuation task by feeding the condition encoder with only video or audio inputs at the first iteration. During the training of the condition encoder, we randomly selected a generation task from 1) unconditional sounding video generation and conditional sounding video generation tasks based on 2) sounding videos, 3) videos, and 4) audios with uniform probability in each training iteration. Samples of long sounding video generation, audio continuation, and video continuation are partly presented in Fig. 3. More samples can be obtained in <https://anonymouss765.github.io/MM-LDM>. It can be seen that our method can synthesize consistent and realistic long sounding videos, demonstrating the effectiveness of our proposed method.

F MORE IMPLEMENTATION DETAILS

Structures of audio and video encoders. The audio encoder is constructed in a similar way to the encoder part of U-Net, which consists of residual blocks and spatial attention layers. Since video signals are temporally redundant [7], we uniformly select keyframes from the input video to feed our video encoder. The structure of the video encoder differs from the audio encoder in two key aspects. Firstly, it adds a temporal attention layer after each spatial attention layer to capture temporal relationships. Secondly, an additional temporal pooling layer is employed before the final layer to integrate temporal information.

Structures of audio and video projectors. Our projectors incorporate multiple residual blocks, spatial attention layers, and downsampling layers. Each residual block is succeeded by a spatial attention layer and a downsampling layer, except for the last block. By default, the audio and video projectors comprise three and four residual blocks, downsampling the audio and video perceptual latents by factors of 4 and 8, respectively.

Training settings. The multi-modal autoencoder is first trained without the adversarial loss for 30 epochs on the Landscape dataset and 10 epochs on the AIST++ dataset. Then, we introduce the adversarial loss and continue training for an additional 30 epochs on the Landscape dataset and 10 epochs on the AIST++ dataset. Notably, we stop the gradient of perceptual latents z_a and z_v when feed them to the projectors using *detach()*. For the training of the generator diffusion modal, i.e., DiT, we optimize it for 300 and 100

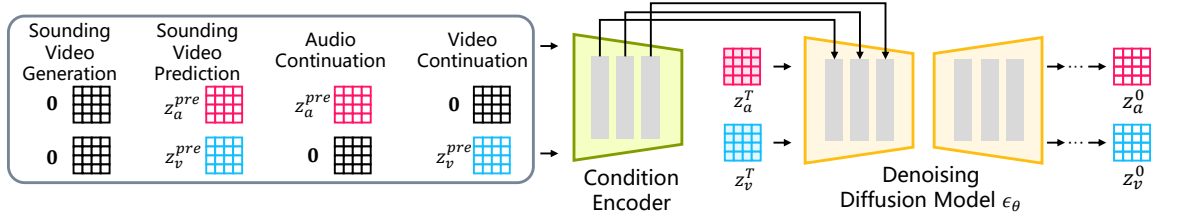


Figure 2: We extend our MM-LDM to long video generation by incorporating a condition encoder, thus enabling to model the sounding video generation, sounding video prediction, audio continuation, and video continuation simultaneously. z_a^{pre} and z_v^{pre} denotes the synthesized previous audio and video latents.

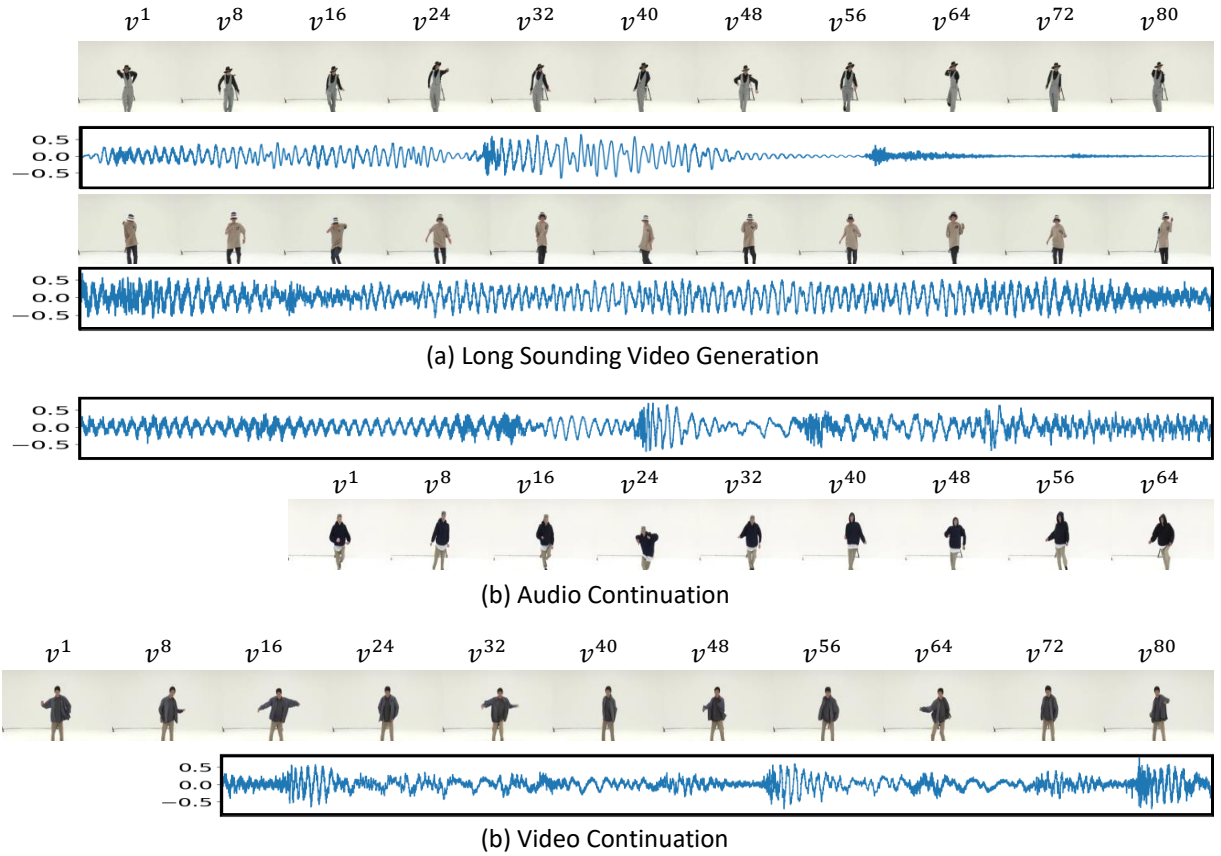


Figure 3: Samples of long video generation, audio continuation, and video continuation on the AIST++ dataset.

epochs on the Landscape and AIST++ datasets, respectively. Our methods are implemented using PyTorch [3], and all experiments are conducted on 8 NVIDIA A100 GPUs. The detailed settings of model hyper parameters are presented in Table. 1.

REFERENCES

- [1] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. 2022. Sound-guided semantic video generation. In *European Conference on Computer Vision*. 34–50.
- [2] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.
- [3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (2019).
- [4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Table 1: Hyper-parameters of the multi-modal auto-encoder and the DiT.

	Landscape	AIST++
	KL-VAE	
f	8	
	Video Encoder	
f_a	4	
f_o	2	
Input Shape	32	
Input Channels	4	
Output Channels	16	
Model Channels	320	
Num Res. Blocks	2	
Num Head Channels	64	
Attention Resolutions	[16, 8]	
Channel Multiplies	[1, 2]	
	Video Decoder (UNet)	
Input Shape	32	
Input Channels	4	
Output Channels	4	
Model Channels	320	
Num Res. Blocks	2	
Num Head	8	
Attention Resolutions	[32, 16, 8]	
Channel Multiplies	[1, 2, 4, 4]	
	Video Generator (DiT)	
Input Shape	16	
Input Channels	16	
Model Channels	1152	
Num Head	16	
Depth	28	
Mlp Ratio	4	

- [5] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. 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*. 10219–10228.
- [6] Mingzhen Sun, Weining Wang, Zihan Qin, Jiahui Sun, Sihan Chen, and Jing Liu. 2024. GLOBER: Coherent Non-autoregressive Video Generation via GLOBal Guided Video DecodER. *Advances in Neural Information Processing Systems* 36

- (2024).
- [7] Mingzhen Sun, Weining Wang, Xinxin Zhu, and Jing Liu. 2023. MOSO: Decomposing MOTion, Scene and Object for Video Prediction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 18727–18737.
- [8] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-Genre, Multi-Dancer, and Multi-Camera Database for Dance Information Processing.. In *ISMIR*, Vol. 1. 6.