MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation Supplementary Material

Ludan Ruan[1]*, Yiyang Ma[2], Huan Yang[3]†, Huiguo He[3],
Bei Liu[3], Jianlong Fu[3], Nicholas Jing Yuan[3], Qin Jin[1], Baining Guo[3]
[1]Renmin University of China, [2]Peking University, [3]Microsoft Research
[1]{ruanld,qinj}@ruc.edu.cn, [2]myy12769@pku.edu.cn,
[3]{huayan,v-huiguohe,bei.liu,nicholas.yuan,jianf,bainguo}@microsoft.com

In this supplementary material, we introduce the algorithm details in Sec. 1. Next, we propose more details of human study in Sec. 2. Finally, we show more visualization results in Sec. 3.

1. Algorithm Details

In this section, we introduce all the implementation details to ensure the reproducibility of our results. We formally list the details of **Architecture**, **Diffusion Process**, **Training Settings** of the Coupled U-Net and super-resolution network in Table. 1.

2. Details of Human Study

To subjectively evaluate the generative quality of our MM-diffusion, we conduct 2 kinds of human study as written in the main paper: MOS and Turing test. For MOS, we asked testers to rate the video quality, audio quality and video-audio alignment based on the standards in Table 2. For Turing test, we asked the common users to vote the given video: 1). It is generated by machine; 2). It cannot be determined if the video is machine-generated or real; 3). It is real. We regard the latter two votes as passing the Turing test.

3. Additional Samples

In this section, we show more unconditional generation results of video-audio pairs from Landscape, AIST++ and AudioSet [1] in Sec. 3.1. Next, we visualize more zero-shot conditional generation results in Sec. 3.2. All results are sampled with 1,000 steps for the best quality.

3.1. Unconditional Generation Results

Firstly, we show more unconditional generation results from Landscape and AIST++ in Figure 1 and Figure 2 respectively. To verify the generative capability of our MM-Diffusion on the open domain, we further train our Coupled U-Net on the largest audio event dataset AudioSet [1] with paired videos on the open domain. It contains 2.1 million video clips of 10 seconds, which covers 632 event classes in total. The audio in AudioSet is complete and the amount of data is sufficient, but the video quality is not high. Therefore, we filter 20k videos of high quality according to the video frame rate and video size. We scale up our Coupled U-Net by enlarging the base channel from 128 to 256, other settings remain unchanged. The visualization results are shown in Figure 3. All corresponding videos in MP4 format are packed in the supplementary video.

3.2. Zero-shot Conditional Generation Results

In this section, we propose video-based audio generation on AIST++. As is shown in Figure 4 (a), taking the same video as input, our model can generate different audios corresponding to the dancing beats. Symmetrically, we next propose audio-based video generation on Landscape. With the results in Figure 4 (b), we find that our model can generate diverse video scenes of the sea to the given wave sound.

References

[1] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 1

*This work was performed when Ludan Ruan was visiting Microsoft Research Asia as research interns

†Corresponding author.

	Coupled U-Net	Super-Resolution
Architecture		
Base channel	128	192
Channel scale multiply	1,2,3,4	1,1,2,2,4,4
Blocks per resolution	2 + 1down/up sample	2 + 1down/up sample
Video downsample scale	H/2, W/2	H/2, W/2
Audio downsample scale	T/4	N/A
Video attention scale	2,3,4	4,5,6
Audio conv dilations	1,2,4,...2^{10}	N/A
Cross-modal attention scale	2,3,4	N/A
Cross-modal attention window size	1,4,8	N/A
Attention head dimension	64	48
Step embedding dimensions	128	192
Step embedding MLP layers	2	2
Diffusion Process		
Diffusion noise schedule	linear	linear
Diffusion steps	1000	1000
Prediction target	ϵ	ϵ
Learn sigma	False	True
Sample method	DPM solver	DDIM
Sample step	N/A	25
Training Settings		
Video shape	$16 \times 64 \times 64$	LR: 64×64, HR: 256×256
Video fps	10	N/A
Audio shape	1×25600	N/A
Audio sample rate	16,000 Hz	N/A
Augmentation	N/A	Gaussian noise, $\sigma \in [0, 20], p = 0.5$ JPEG compression, $q \in [20, 80], p = 0.5$ Random flip, $p = 0.5$
Weight decay	0.0	0.0
Dropout	0.1	0.1
Learning rate	1e-4	1e-4
Batch size	128	48
Training steps	100,000	270,000
Training hardware	$32 \times V100$	$8 \times V100$
EMA	0.9999	0.9999

Table 1. The implementation details of our Coupled U-Net and super-resolution network.

Score	Video/Audio Quality	Video-Audio Alignment
1	Pure noise, completely unrecognizable content.	The video and audio are total noise and they are completely irrelevant.
2	The video/audio has development, but the video/audio type can not be recognized.	The type of video/audio can not be recognised and they are irrelevant.
3	Video/audio can be recognized as specific type, but very unnatural.	The type of video/audio can be recognized but they are misaligned.
4	The video/audio is natural, but can be recognized as generated content.	Video and audio are basically matched, but the detail in correspondence is lacking.
5	The video/audio is so natural that can not be recognized if is from generation or real-world.	The video and audio are consistent in detail and very natural.

Table 2. The score description of MOS for video/audio quality and video-audio alignment.

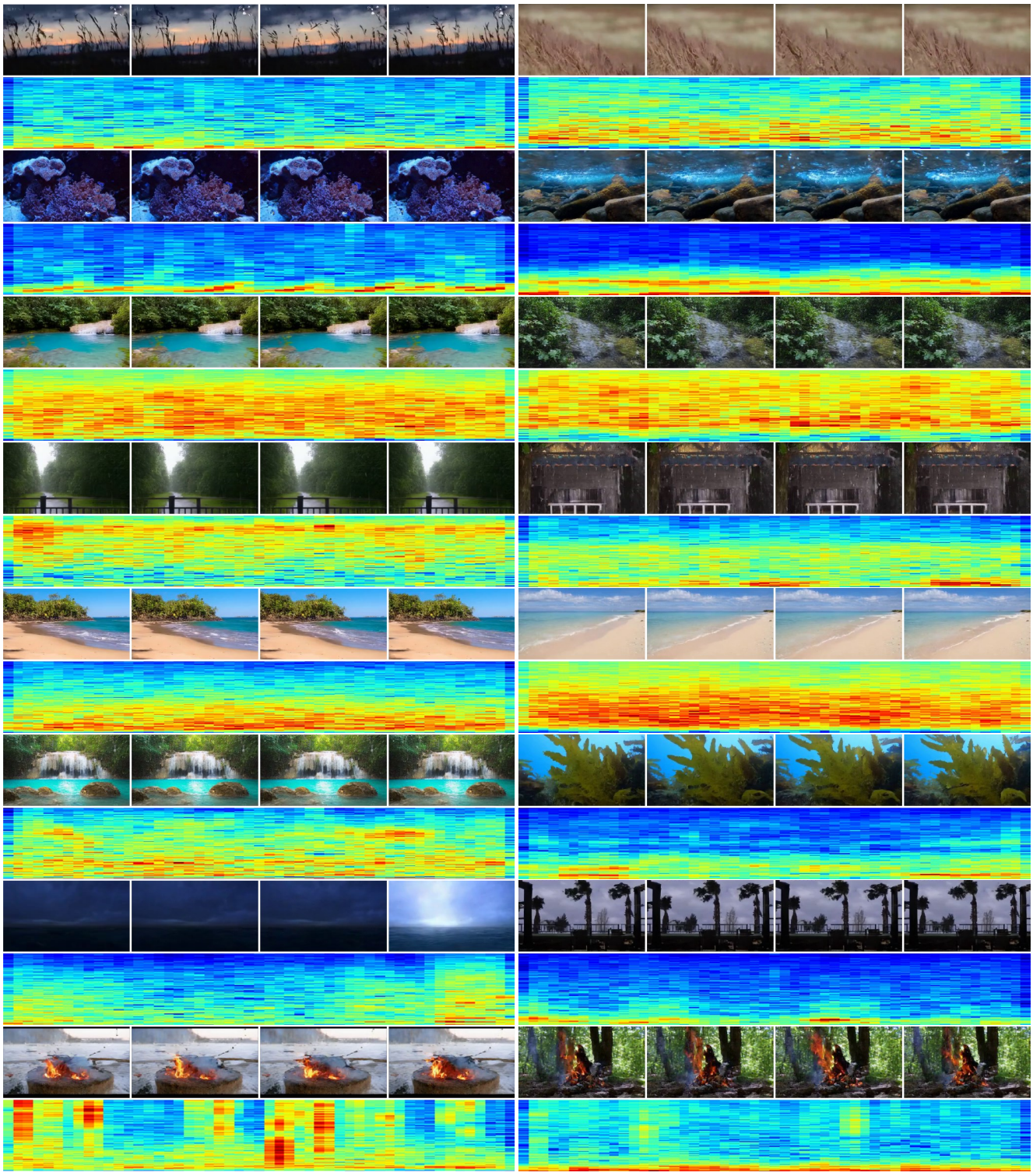


Figure 1. More generation results from Landscape of our MM-Diffusion. The given cases show the scenes of blowing wind, underwater, splashing water, raining, squashing water, waterfall, thunder, and fire cracking respectively.

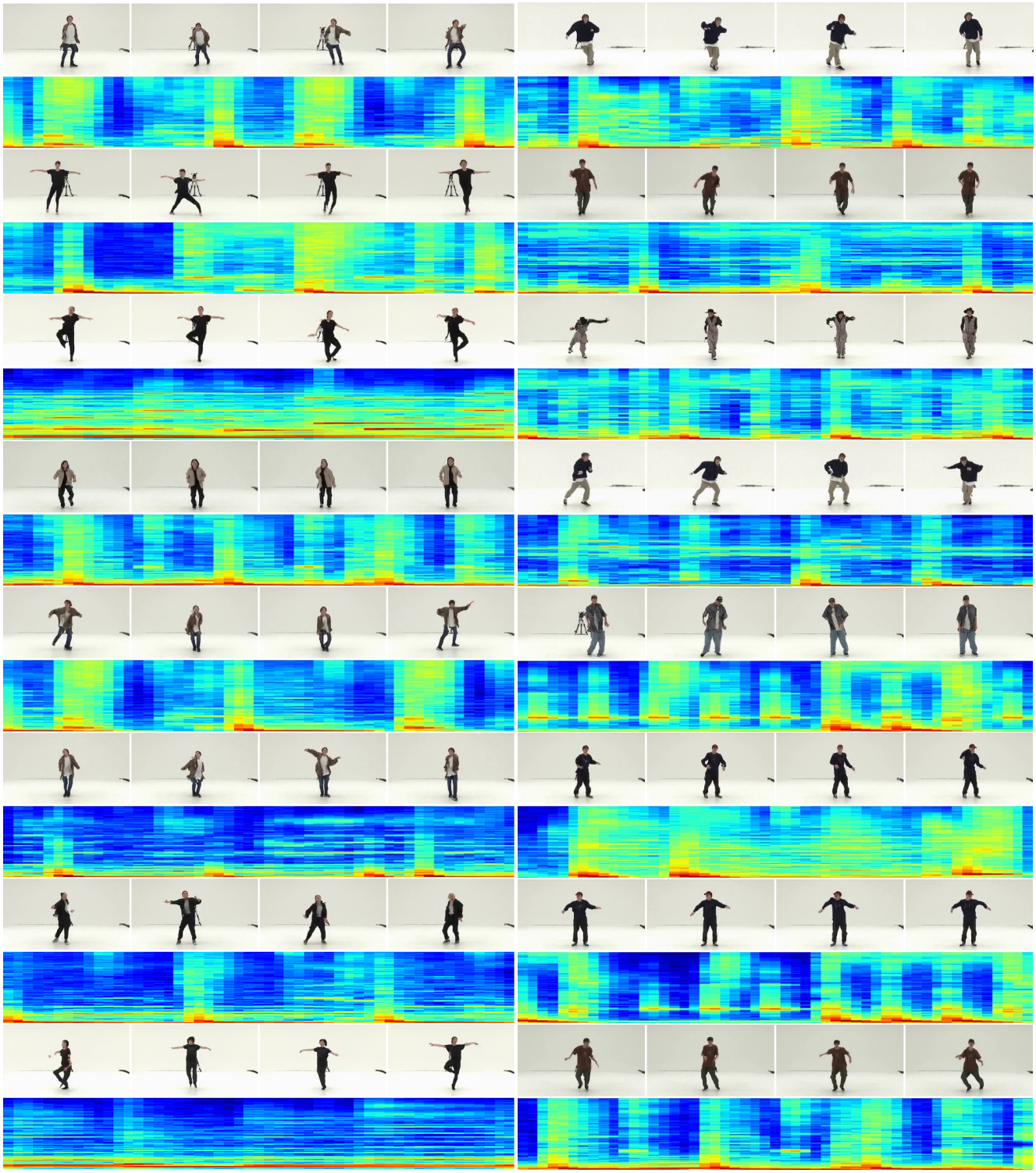


Figure 2. More generation results from AIST++ of our MM-Diffusion. Matched audio is generated with video appearances (e.g., the periodical rhythm for dancers).

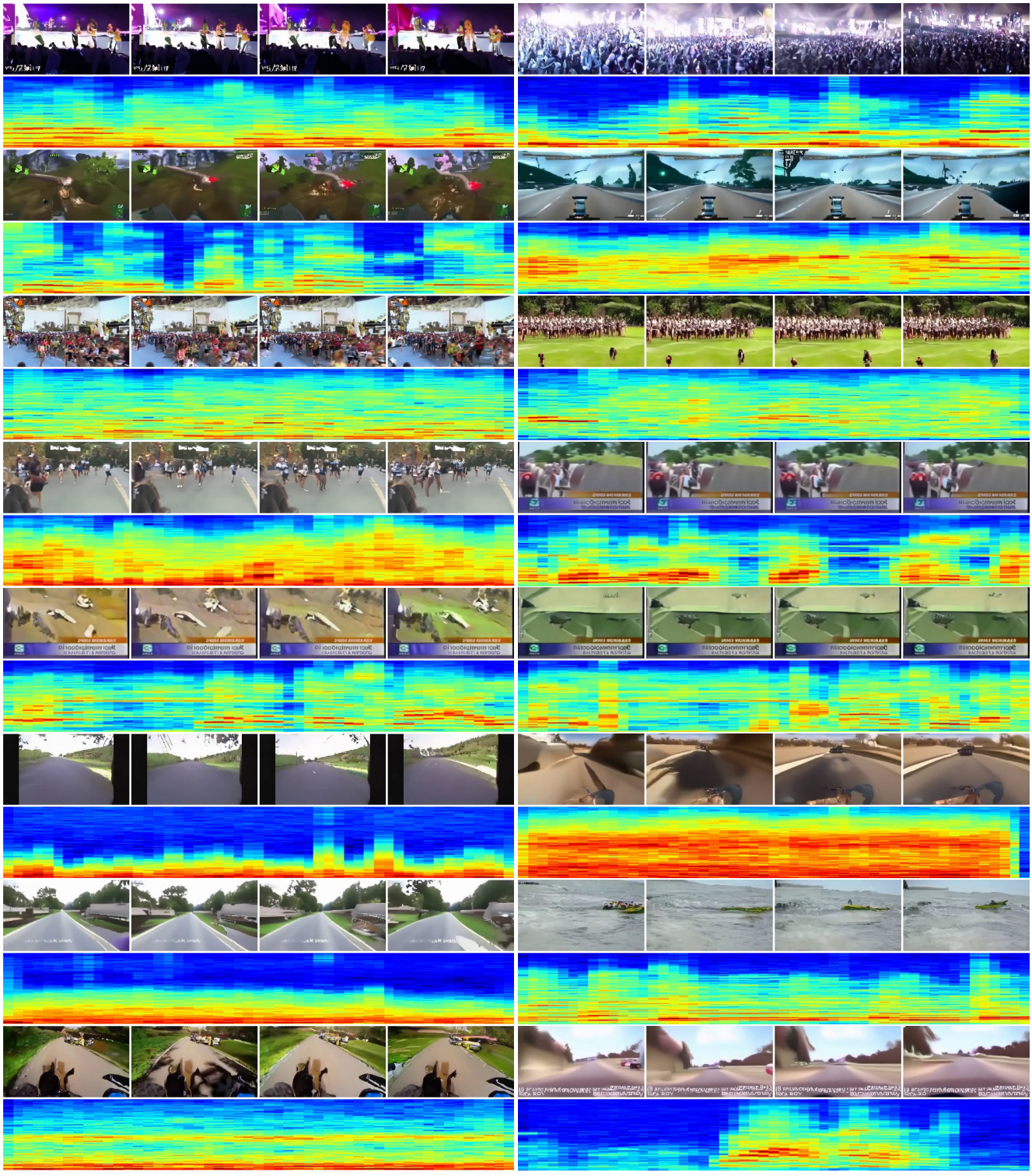
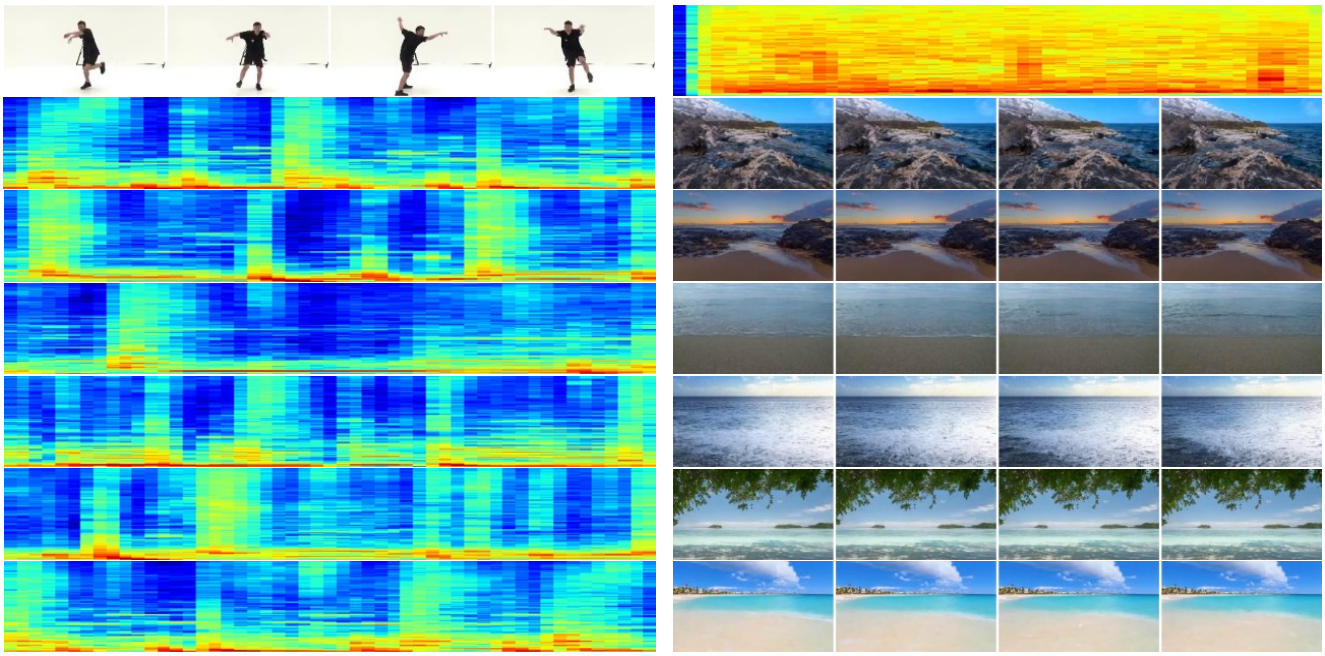


Figure 3. More generation results from open domain (AudioSet) of our MM-Diffusion.The given cases show the scenes of concert, game streaming, marathon, news playback, surfing and driving in first-person perspective respectively



(a) Video to Audio (b) Audio to Video

Figure 4. More visual examples of zero-shot conditional generation with our MM-diffusion.