Supplementary Material: Generating Realistic Images from In-the-wild Sounds

Note: We provide the implementation details, wild audio files used in the figures on the main paper, and additional audio files and generated image samples.

1. Implementation Details

We used single NVIDIA A100 80G for our experiments, and performed 10 epochs of direct sound optimization using the Adam optimizer with λ_{aCLIP} = 0.9, λ_{CLIP} = 1.0, λ_{L2} = 0.01 and set the learning rate for W_n to 0.01 and learning rate for W_{wn} to 0.001. We used a PLMS sampler and we performed the DDIM step 40 time. Also it took about 10 seconds per epoch. The W2c-vqgan we used as the baseline consists of two models: sound feature extraction and image generation. For extracting sound features in W2-vqgan, the pre-trained model proposed in Wav2clip [7] was used. For generating images from sound features, the VQGAN-CLIP [1] pre-trained on ImageNet [2] was used. Wan *et al*. [6] was trained under the same conditions using the 10,701 sound-image paired dataset as proposed in their paper. The code for this approach is available at Github. Our model and W2c-vqgan have an image resolution of 512x512, while Wan *et al*. [6] has an image resolution of 64x64.

2. Audio Files for the Generated Images in the Main Paper

Our task is to generate realistic images from in-the-wild sounds. We provide the wild audio files used in the figures of our main paper and you can listen to them by referring audio file folder. The audio folder names correspond to the figure numbers of the main paper. For example, if you want to listen to the audio for **'Figure 3'** of the main paper, you will find the audio files within the **'figure_3'** folder. The audio files from the left in the figure are matched with audio_01.mp4, audio_02.mp4, and so on.

3. Audio Files and Generated Images (not Included in the Main Paper)

We provide additional generated images from both single category sounds and wild sounds that we could not include in the main paper due to space constraint.

Figure 1 and 2 show the generated images from the Urbansound8K[5] audio dataset. We observe that each of the images is generated appropriately for the corresponding sound, generating rich and realistic images according to the intensity of the sound. The audio files are included in **'sup_figure'** folders. For instance, if you listen to the sounds (**sup_figure/sup_figure_1_urban8K/audio_03.mp4, audio_04.mp4**) of the third and fourth images from the left in Figure 1, you can observe that the number of cars generated varies depending on the intensity of the car horn sound. Furthermore, if you listen to sounds (**sup_figure/sup_figure_2_urban8K/audio_04.mp4, audio_05.mp4**) of the fourth and fifth images from the left in Figure 2, you can observe that the generated images differ depending on whether the sounds are continuous gunshots or a single gunshot.

Figure 3 contains additional images generated from Clotho[3], Figure 4 contains additional images generated from Audiocaps[4] and Figure 5 contains additional images generated from Multi-ESC50.

References

[1] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.

[4] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

Figure 1. **Comparison of sound-guided image generation result on Urbansound8K.** Class is ground-truth audio class of Urbansound8K [5]. The audio files corresponding to the figure are located in the **'sup_figure/sup_figure_1_urban8K'** folder, and the audio files from the left in the figure are matched with audio_01.mp4, audio_02.mp4, and so on.

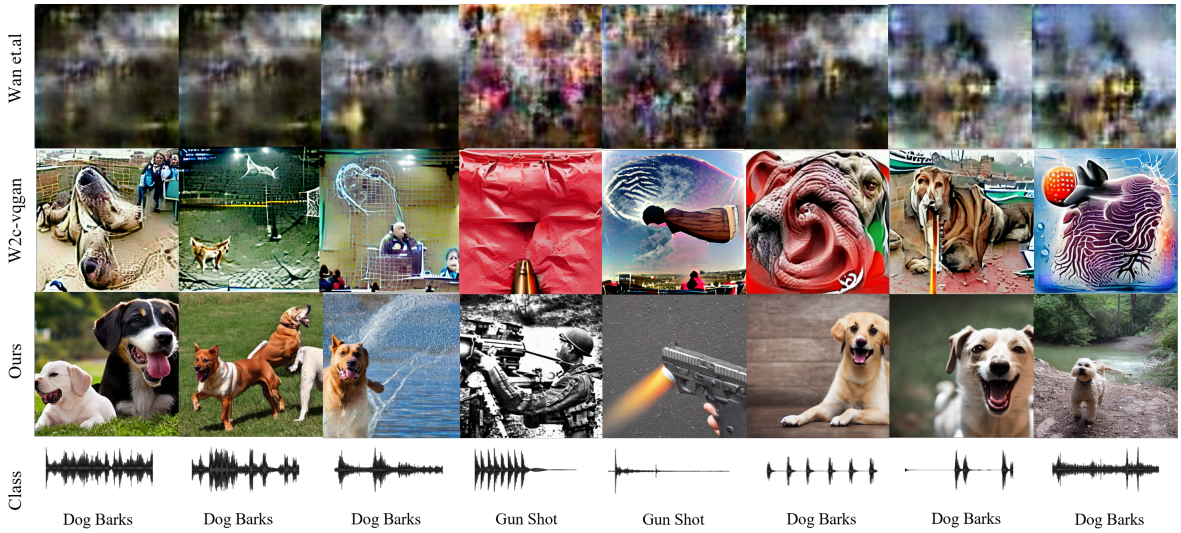

Figure 2. **Comparison of sound-guided image generation result on Urbansound8K.** Class is ground-truth audio class of Urbansound8K [5]. The audio files corresponding to the figure are located in the **'sup_figure/sup_figure_2_urban8K'** folder, and the audio files from the left in the figure are matched with audio_01.mp4, audio_02.mp4, and so on.

[5] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[6] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500. IEEE, 2019.

[7] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.

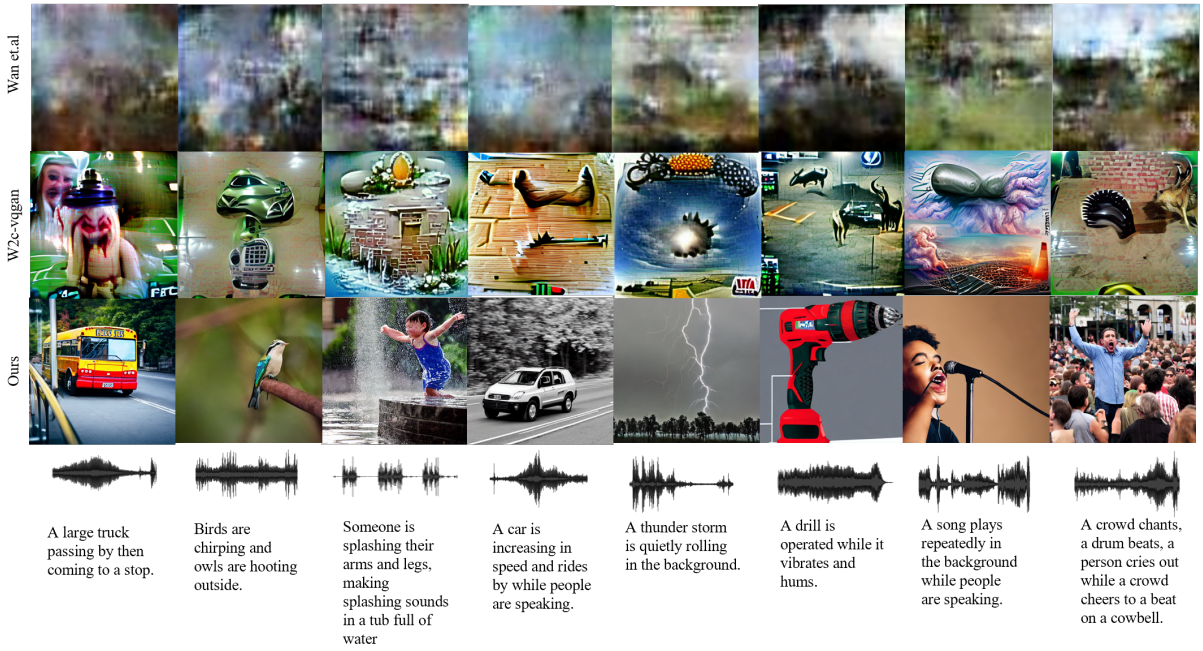

Figure 3. **Comparison of sound-guided image generation result on Clotho.** Text is ground-truth(GT) audio caption of audio datasets [3]. The audio files corresponding to the figure are located in the **'sup_figure/sup_figure_3_clotho'** folder, and the audio files from the left in the figure are matched with audio_01.mp4, audio_02.mp4, and so on.

Figure 4. **Comparison of sound-guided image generation result on Audiocaps.** Text is ground-truth(GT) audio caption of audio datasets [4]. The audio files corresponding to the figure are located in the **'sup_figure/sup_figure_4_audiocaps'** folder, and the audio files from the left in the figure are matched with audio_01.mp4, audio_02.mp4, and so on.

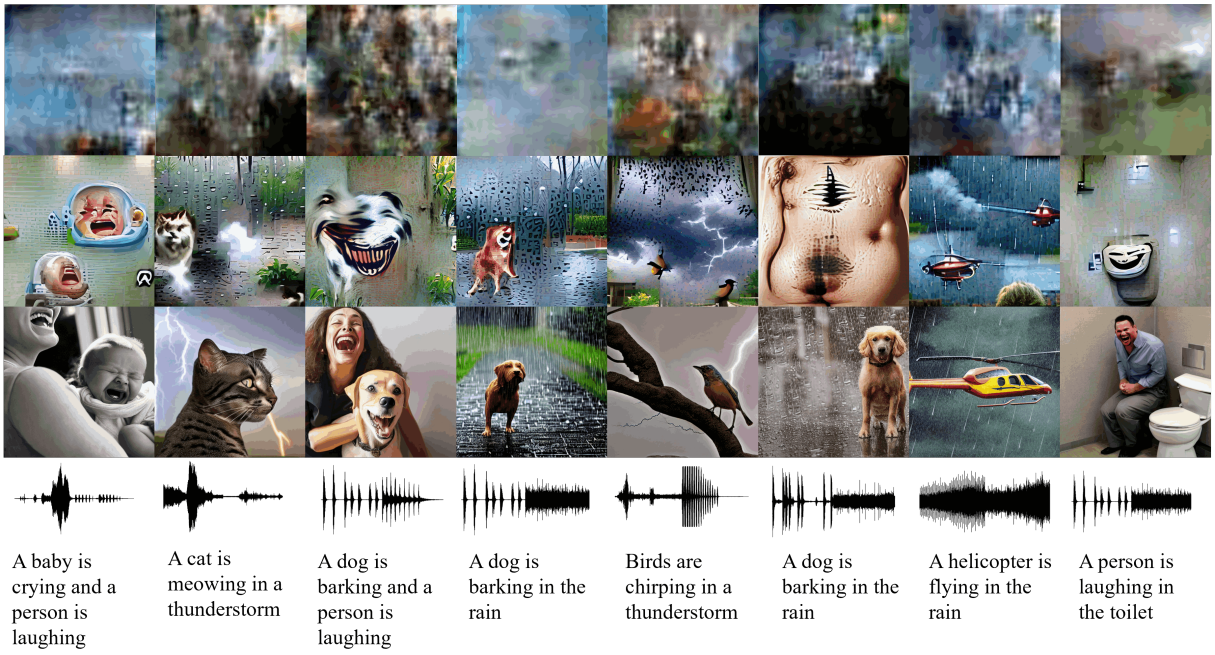

Figure 5. **Comparison of sound-guided image generation result on Multi-ESC50.**. The audio files corresponding to the figure are located in the **'sup_figure/sup_figure_5_multi'** folder, and the audio files from the left in the figure are matched with audio_01.mp4, audio_02.mp4, and so on.