## Supplementary Material

**Overview** This supplementary material provides implementation details including the dataset details for learning multi-modal representations and optimizing image manipulation (Section A). We also provide an additional ablation study with a different set of hyperparameters to see their effects on the quality of image manipulation (Section B). Next, in Section C, we provide our detailed analysis of the manipulated latent code. Lastly, in Section D, we provide (i) details of our user study and (ii) more diverse qualitative results with a variety of sound sources. We discuss about societal impact in Section E. Moreover, we also provide diverse examples in the following *anonymized* project website: https://kuai-lab.github.io/cvpr2022sound/.

## A. Implementation Details

**Dataset Details**. For learning multi-modal representations of audio, text, and images, we use large-scale audio and video benchmarks [3, 1] as training datasets. As an evaluation dataset, the zero-shot classification accuracy is measured using the audio classification benchmark [11, 9]. For image manipulation evaluation, we use partial datasets [3, 1] of audio and video which were not included in the training dataset.

To establish a multi-modal embedding space, we use the audio-text pair and video datasets named The Audio Set [3] which is consists of 632 audio event categories and a collection of 2,084,320 human-labeled 10-seconds sound clips from YouTube videos. Since there are some missing associate urls on YouTube, we use 17,153 audio and text pairs out of 20,371 pairs in balanced subsets, and 1,617,939 pairs out of 2,041,789 pairs as unbalanced subsets for training. We also utilized VGG-Sound [1] which consists of 310 event classes and a collection of over 200,000 human-labeled 10-second video clips pulled from YouTube videos. We use 182,342 videos for training in the dataset. For each video clip, we capture the middle frame and use it as an input image. Here, the corresponding text prompts are extracted from the ground truth labels.

To evaluate zero-shot transferability of the audio embedding from our proposed model, we use the audio classification benchmark: (i) Environment Sound Classification dataset (ESC-50) [9], which comprises of 2000 clips from 50 classes. Note that each of these clips are sampled at 44.1 kHz, with a length of 5 seconds. (ii) The UrbanSound8k dataset [11] contains 8732 clips from 10 classes. Each audio is less than 4-second long and sampled at frequencies of 16 to 48 kHz.

**Implementation Details for StyleGAN2 Generator**. We implement StyleGAN generator [5] based on the official PyTorch implementation from StyleGAN2-ADA [1]. We manipulate the images with the pre-trained generator with high-resolution image datasets [6, 12, 16]. Flickr-Faces-HQ (FFHQ) [6] contains the 70k high-quality human face images in resolution of $1024 \times 1024$. The Large-scale Scene Understanding (LSUN) Challenge [16] contains the church images in resolution of $256 \times 256$ and the car images in the resolution of $384 \times 512$. WikiArt [12] contains the painting images in resolution of $1024 \times 1024$ drawn by 195 different artists. The Landscapes High-Quality (LHQ) [13] contains nature landscape images in resolution of $256 \times 256$.

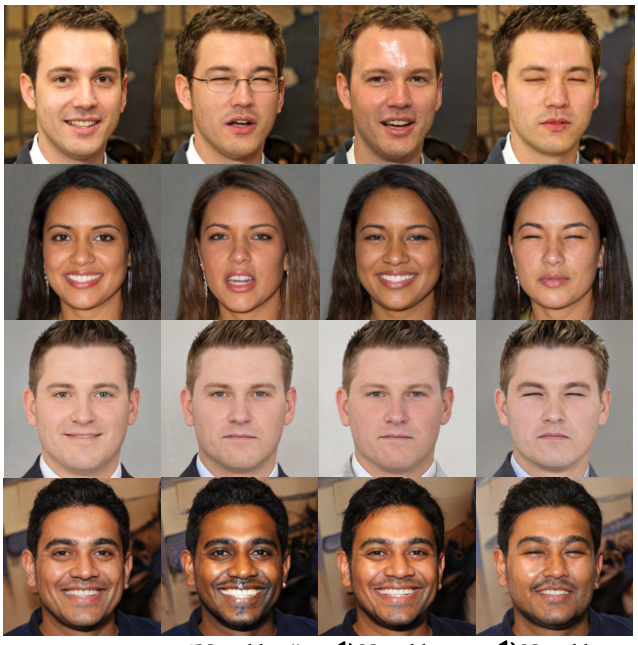


Figure 1: Ablation study of self-supervised representation learning for audio inputs.

## B. Ablation Study

**Effect of Self-supervised Contrastive Representation Learning for Audio Inputs.** We conduct an ablation study on applying self-supervised representation learning. We already demonstrate that the self-supervised representation learning of audio inputs improves the zero-shot audio classification performance. Furthermore, rich audio representation obtained from the self-supervised learning improves sound-guided image manipulation quality (see Fig. 1). These are manipulation results without $\mathcal{L}_{nce}^{(a\leftrightarrow\hat{a})}$ and with $\mathcal{L}_{nce}^{(a\leftrightarrow\hat{a})}$, respectively. In the absence of the self-supervised learning of audio inputs, the manipulation results are similar to StyleCLIP [8] because the representation in the latent space of audio is CLIP [10]-dependent. The self-supervised

[1] https://github.com/NVlabs/stylegan2-ada-pytorch

learning of audio inputs reflects vivid emotions in image manipulation that can be only done with audio. In addition, Fig. 3 shows that our method produce more realistic manipulation results than AudioCLIP [4].

**Effect of Identity Loss.** We compare the manipulation results while varying hyperparameters $\lambda_{sim}$ and $\lambda_{ID}$ (see Fig. 2). Changes in brightness and saturation can be controlled by those hyperparameters. High values of $\lambda_{sim}$ and $\lambda_{ID}$ lead to maintaining the content of the source image, while low values distort the content.

**Effect of Symmetric Contrastive Loss Form.** We identify the effectiveness of the symmetric contrastive loss form. Our model trained with the symmetric loss shows 4.05 % higher in the ESC-50 [9] dataset and 3.1 % higher in the UrbanSound8k [11] dataset than those with the non-symmetric form. These experimental results imply that loss function with symmetric form improves the zero-shot audio classification accuracy.

## C. Manipulated Latent Code Analysis

**Manipulated Latent Code Interpolation.** Our model allows different modalities (audio, text and image) to share the same embedding space using our multi-modal contrastive loss. In order to illustrate that the source latent code is guided in the same embedding space even if the modality is different, we interpolate text and sound-guided latent code (see Fig. 10 and Fig. 11). The interpolated latent code $w$ is the weighted sum of $w_t$, which is a text-guided latent code, and $w_a$, which is a sound-guided latent code. The expression $w = (1-\alpha) \cdot w_a + \alpha \cdot w_t$ is obtained. It shows the result of generating an image by linearly interpolating the latent code guided by text and audio. The generated result from the interpolated latent code contains the intermediate meaning of the two modalities continuously.

**Distribution of Manipulation Direction.** We analyze the distribution of the manipulation direction that text and sound are intended to guide. We perform audio-driven image manipulation of 150 latent codes with audio that is not used for learning among the attributes in the category of VGG-Sound [1]. As shown in Fig. 4, sound manipulates images with more diverse guidance than text. When manipulating an image with six attributes with the FFHQ [6] dataset, the sound is farther from the average of the manipulation direction than text. The mean and variance for the manipulation direction are numerically summarized in Table 1.

We illustrate that the direction of sound-guided manipulation is more diverse than that of text-based manipulation. Our method encourages audio representations of the same-class different views to be closer in the embedding space.

## D. Qualitative and Quantitative Results

**User Study Details.** Detailed results of downstream tasks are described in Table 2. We manipulate the source image with a total of 8 attributes. In the FFHQ [6] dataset, images are manipulated by three attributes including giggling, sobbing and nose blowing. The remaining wind noise, underwater bubbling, explosion, and thunderstorm are used to manipulate the image generated by the LSUN [16] dataset. Our model manipulates images with the meaning of attributes better than text-based manipulation methods such as TediGAN [15] and StyleCLIP [8]. Among three models, our model best reflects the meaning of attributes in image manipulation.

Additionally, Amazon Mechanical Turk (AMT) participants select the manipulation results generated by our model match well with the ground truth. Fig. 5 also shows the incorrect answer ratio chosen by the participants in the user study for each question. In text-based manipulation methods, most people perceive the results created with "sob" and "scream" as "giggle".

**Additional Qualitative Examples.** We show more comparison results in Fig. 6 and Fig. 7. The sound-guided image manipulation illustrates more radical results than text-guided image manipulation (StyleCLIP [8]).

Sound controls signal intensity by adding or subtracting decibels. In Fig. 8, we show that the semantic reflected in the image does not change significantly with scaled sound. Still, the detail within the image changes. We demonstrate more diverse examples in Fig. 9, Fig. 12 and Fig. 13. All manipulated results in Fig. 13 are obtained from the pretrained StyleGAN2 [7] with the LHQ dataset [13].

## E. Societal Impact

The proposed method of the sound-guided image manipulation is based on CLIP's knowledge which may have social prejudice. In the CLIP paper, the author says *"CLIP is trained on text paired with images on the internet. These image-text pairs are unfiltered and uncurated, resulting in learning online social biases."*. Therefore, there is a possibility that certain social bias may appear when editing people's faces with sounds related to a thief, prisoner, criminal, and suspicious people, such as guns and fighting sounds. Also, when manipulating a human face with a sound such as a vacuum cleaner or women speaking, manipulation results with social prejudice like housekeepers may appear in the manipulated images.

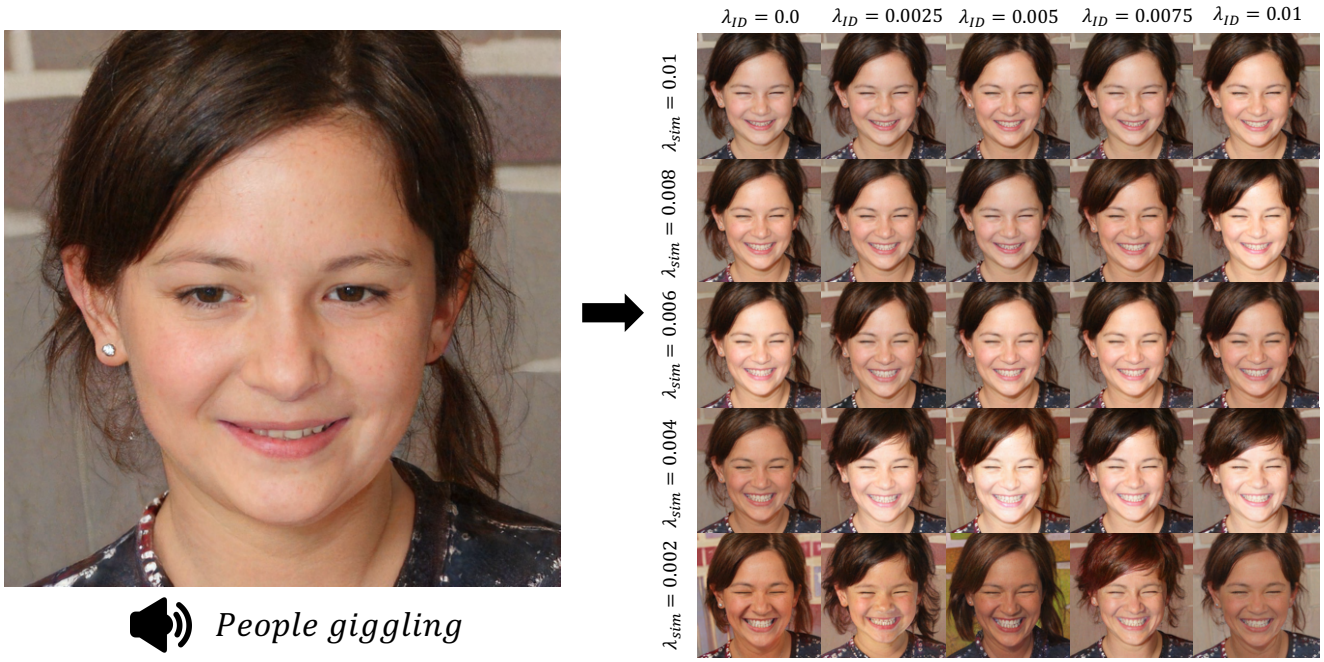


Figure 2: Ablation study of the hyperparameters for sound-guided image manipulation. In the direct latent code optimization step, the row below means that $\lambda_{sim}$ is low and the column to the right means that $\lambda_{ID}$ is high.

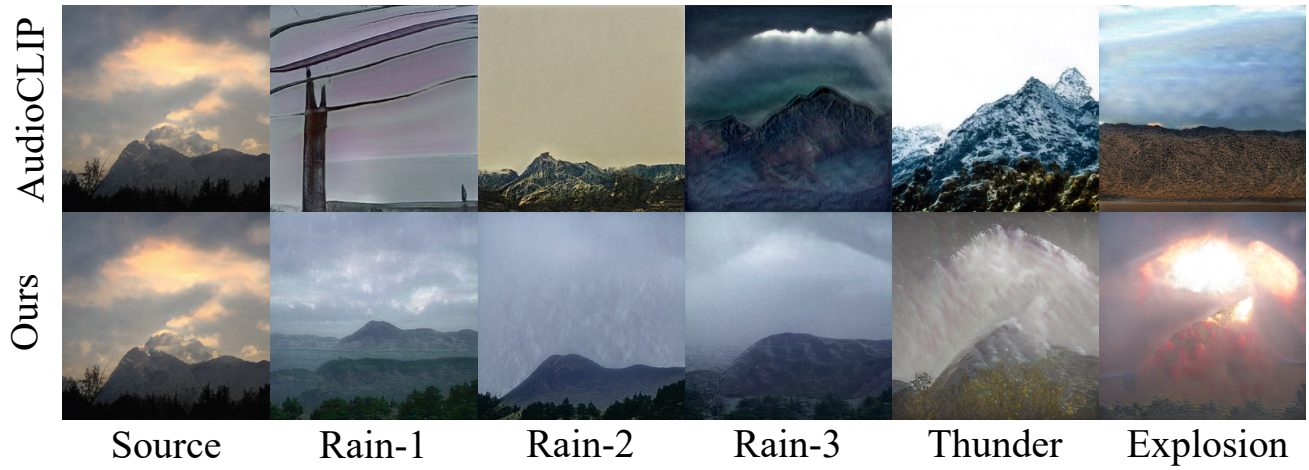


Figure 3: Qualitative comparison of manipulation results between ours and AudioCLIP [4].

Table 1: Cosine similarity between text-guided and sound-guided latent code. $w_s$ is source latent code, $w_a$ is audio-driven latent code, $w_t$ is text-driven latent code. $w_t$ is the latent code guided by StyleCLIP [8]'s text-driven latent optimization.

| | | Attribute | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Latent code | Metric | Giggling | Sobbing | Nose blowing | Fire crackling | Wind noise | Underwater bubbling | Explosion | Thunderstorm | Average |
| ($w_s$, $w_a$) | Mean (↓). | 0.99528 | 0.99510 | 0.99448 | 0.97907 | 0.97679 | 0.98040 | 0.97956 | 0.98479 | 0.98568 |
| | Std (↑). | 0.00178 | 0.00162 | 0.00242 | 0.00786 | 0.00829 | 0.00614 | 0.00698 | 0.00512 | 0.00502 |
| ($w_s$, $w_t$) | Mean (↓). | 0.99866 | 0.99849 | 0.99779 | 0.99002 | 0.99294 | 0.99024 | 0.99117 | 0.99136 | 0.99383 |
| | Std (↑). | 0.00067 | 0.00065 | 0.00097 | 0.00379 | 0.00322 | 0.00318 | 0.00321 | 0.00340 | 0.00238 |
| ($w_a$, $w_t$) | Mean (↓). | 0.99554 | 0.99510 | 0.99448 | 0.97511 | 0.97307 | 0.97695 | 0.97497 | 0.97843 | 0.98295 |
| | Std (↑). | 0.00166 | 0.00162 | 0.00242 | 0.00972 | 0.00935 | 0.00703 | 0.00827 | 0.00713 | 0.00590 |

Table 2: More detailed downstream task evaluation to compare the quality of representations between ours and text-driven manipulation approaches on the FFHQ [6] dataset and the LSUN [16] dataset. This table shows how much the meaning fits the user-provided input after sampling 150 latent codes and manipulating them with TediGAN [15], StyleCLIP [8], and Ours. Image features are extracted with an image encoder trained with multi-modal latent representation learning, and these are classified by logistic regression, a linear classifier.

| | Attribute (↑) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Giggling | Sobbing | Nose blowing | Fire crackling | Wind noise | Underwater bubbling | Explosion | Thunderstorm | Average |
| TediGAN [15] | 0.967 | 0.940 | 0.947 | 0.686 | 0.940 | 0.727 | 0.858 | 0.720 | 0.848 |
| StyleCLIP [8] | 0.933 | 0.913 | 0.866 | 0.846 | 0.953 | 0.933 | 0.987 | 0.987 | 0.927 |
| Ours | **0.987** | **0.993** | **0.966** | **0.953** | **0.993** | **0.987** | **0.993** | **0.993** | **0.983** |

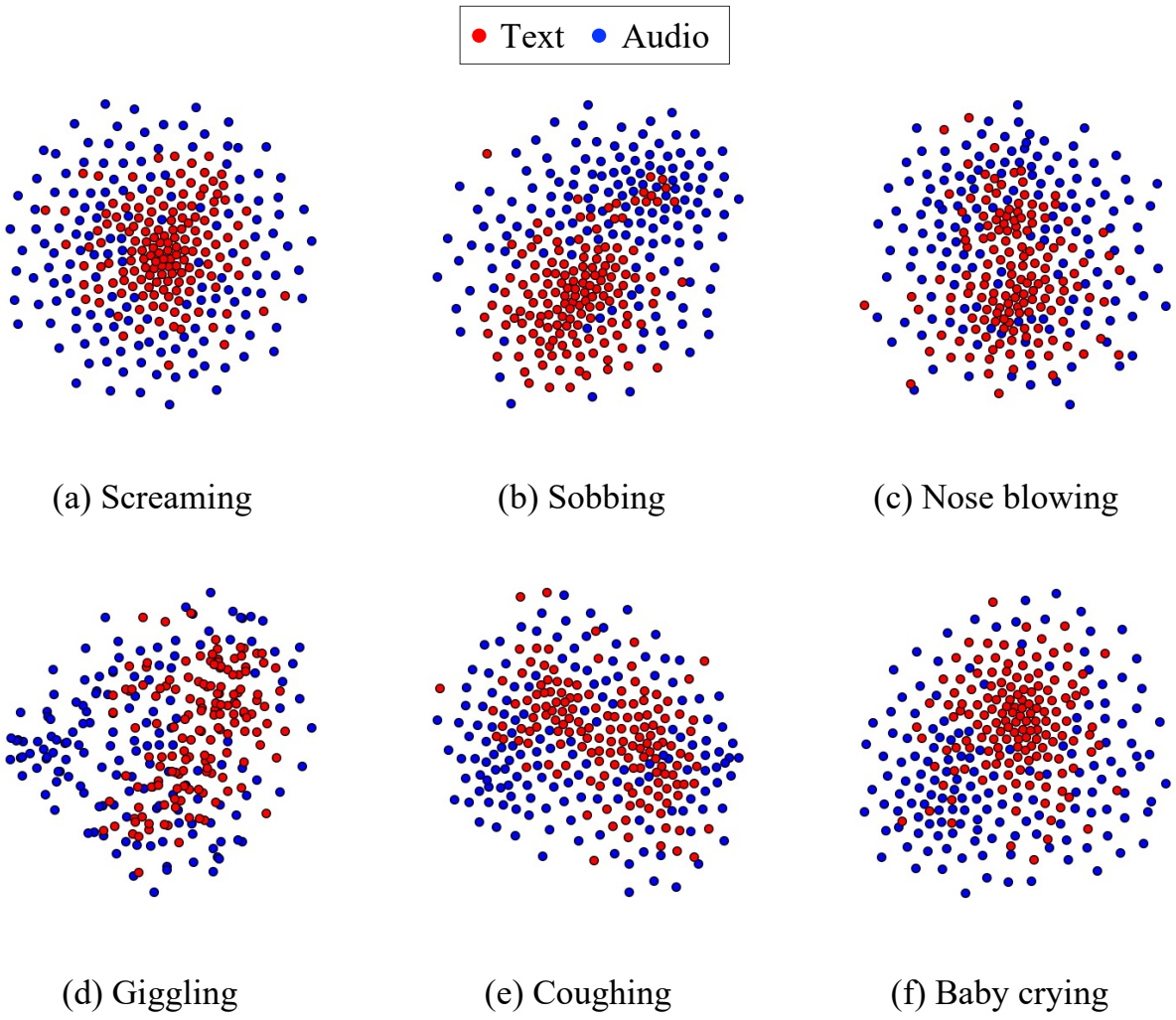


Figure 4: Visualization of audio and text-guided manipulation directions with t-SNE [14] (red: text, blue: audio).


Figure 5: Further user study to demonstrate whether the image manipulated by each model matches the ground truth. These results are reported in the percentage.

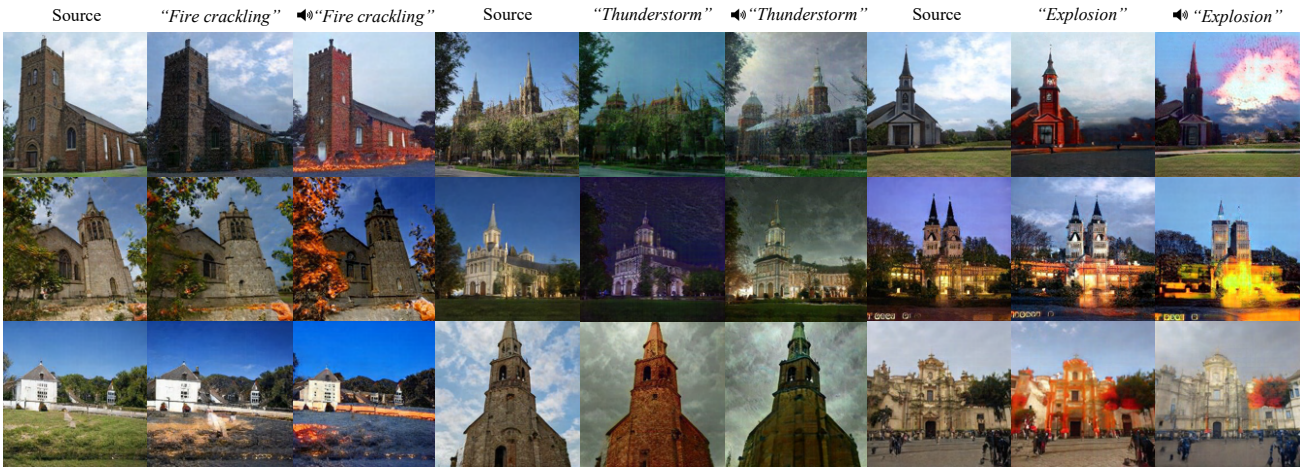


Figure 6: Comparison of the text-driven manipulation and audio-driven manipulation results from the LSUN dataset [16].


Figure 7: Comparison of the text-driven manipulation and audio-driven manipulation results from the FFHQ dataset [6].


Figure 8: Ablation study according to the volume of sound. A change in the size of the volume affects the detailed style, but the overall meaning does not change.

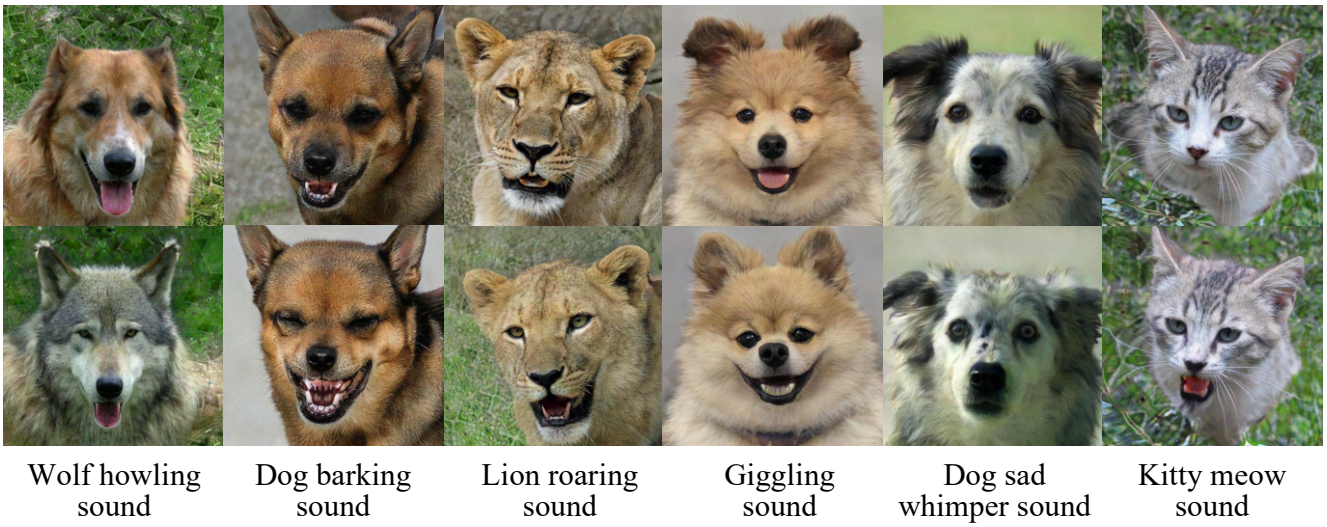


Figure 9: Additional results of sound-guided image manipulation from the AFHQv2 dataset [2]. The first row is the input image, the second row is the manipulation results.

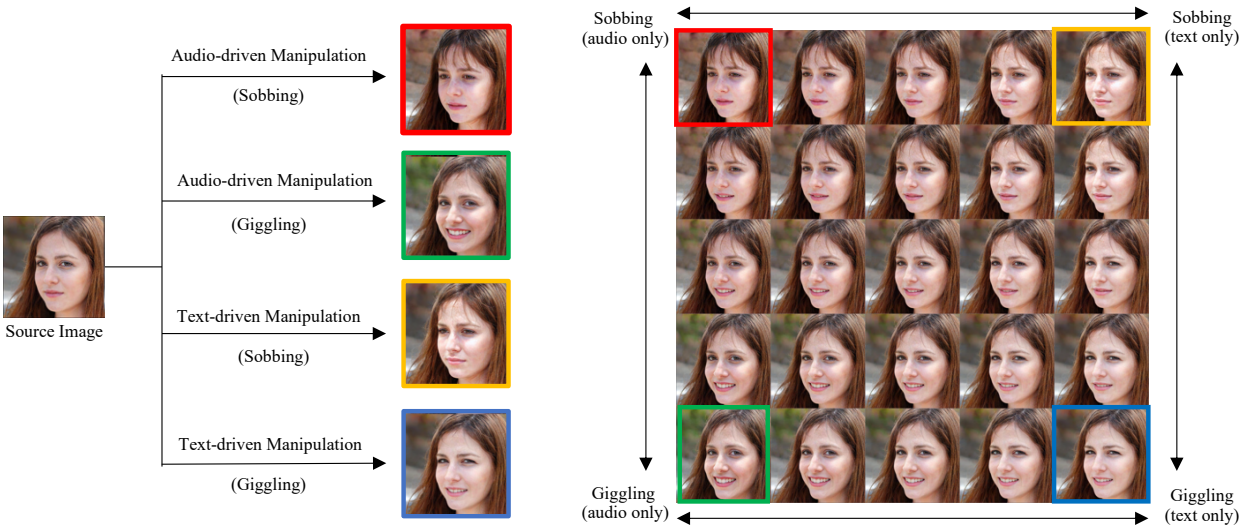


Figure 10: The manipulation results of interpolation between text and audio-guided latent codes. Even if the source latent code is guided by different modality, optimization occurs in the same latent space of StyleGAN, which is pre-trained with the FFHQ [6] dataset.

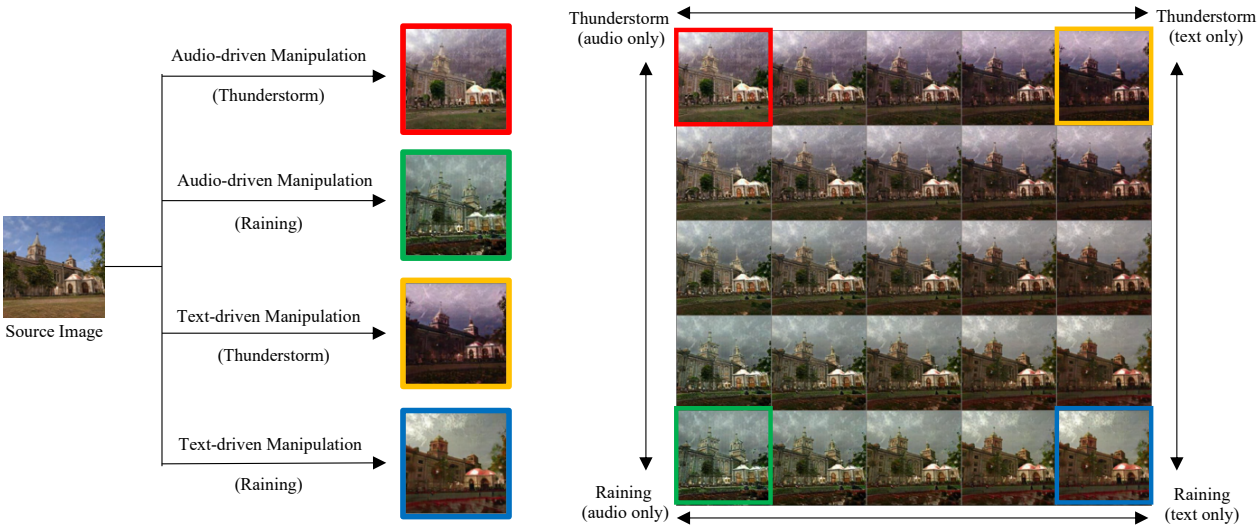


Figure 11: The manipulation results of interpolation between text and audio-guided latent codes. Even if the source latent code is guided by different modality, optimization occurs in the same latent space of StyleGAN, which is pre-trained with the LSUN [16] dataset.

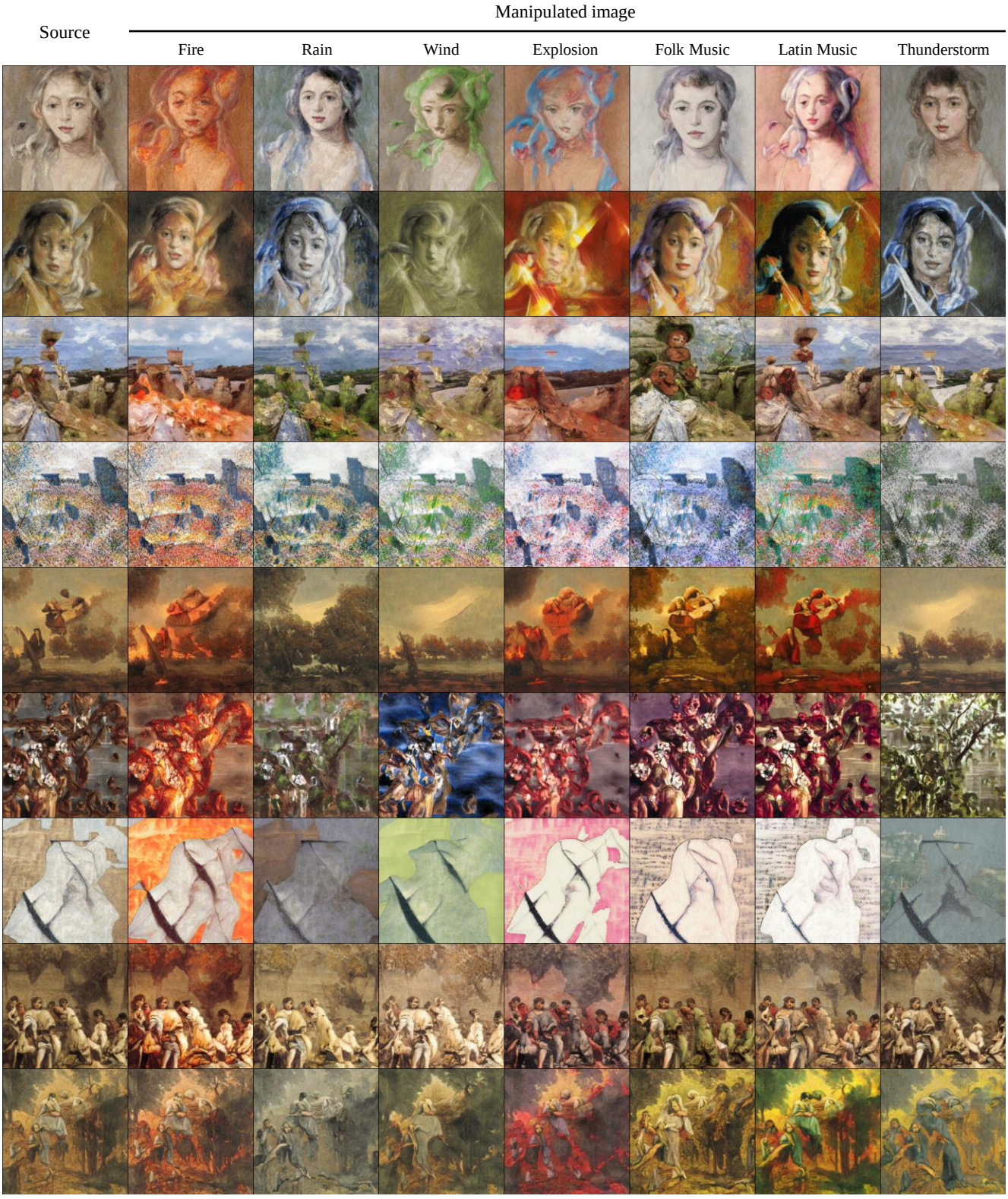


Figure 12: Additional results of sound-guided artistic paintings manipulation from the Wikiart dataset [12].


Figure 13: Additional results of sound-guided image manipulation from the LHQ dataset [13].

## References


[1] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 2

[2] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 7

[3] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[4] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 3

[5] D. Jeong, S. Doh, and T. Kwon. Träumerai: Dreaming music with stylegan. *arXiv preprint arXiv:2102.04680*, 2021. 1

[6] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 4, 6, 8

[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2

[8] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 1, 2, 3, 4

[9] K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 1, 2

[10] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1

[11] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014. 1, 2

[12] B. Saleh and A. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), 2016. 1, 9

[13] I. Skorokhodov, G. Sotnikov, and M. Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14144–14153, October 2021. 1, 2, 10

[14] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-629 sne. *Journal of Machine Learning Research*, 9(2579-2605):630, 2008. 4

[15] W. Xia, Y. Yang, J.-H. Xue, and B. Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 2, 4

[16] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 2015. 1, 2, 4, 6, 8