– Supplementary Material –

Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment

Anonymous CVPR submission

Paper ID 6866

The contents in this supplementary material are as follows: A. Implementation details (Sec. A), B. Suitable categories for sound-to-image (Sec. B), C. Comparison with the prior arts (Sec. C), D. Additional qualitative analysis and results (Sec. D), and E. Details of the user study (Sec. E). We also recommend watching the supplementary video, containing generated images and corresponding input sounds.

A. Implementation Details

Audio pre-processing. The input for the audio encoder is 1004×257-dimensional log-spectrogram, converted from 10 seconds of audio. We first extract up to 10 seconds of audio from the beginning of each video. If the video clip is shorter than 10 seconds, we repeat the audio to have the expected input length. Then, we resample the audio waveform at 16kHz and convert it into frequency domains by constructing a spectrogram. The spectrogram is passed through a logarithm function before using it as input.

Evaluation metric. In the main text, we use Fréchet Inception Distance (FID) [6] and Inception Score (IS) [10] metrics to evaluate the quality and diversity of the generated images. To measure both of the metrics, the Inception-V3 [12] model is required. We fine-tune the Inception-V3 model on VGGSound [3] and compute FID and IS with 30k generated images from the test set.

B. Suitable Categories for Sound-to-Image

Not every category is suitable to be used to infer visual scenes from sounds. In this section, we analyze which categories are not only audio-visually well-corresponded but also suitable for sound-to-image translation.

As described in [11, 15], despite the multi-modality of video datasets, not every class is audio-visually well-corresponded, *e.g.*, Kinetics [7] is visual modality biased. Although several datasets are introduced as audio-visual datasets, many of the categories of these datasets may not be audio-visually correlated, such as "civil defense siren", "wind noise", or "reversing beeps". Moreover, even though categories are audio-visually correlated, they may not contain sufficiently dominant semantic signals that can properly bridge the sound-to-image generation. For example, "people slurping", "people eating", or "people sneezing" are similar in terms of containing human instances regardless of the category, while they completely differ in the audio modality. Such misalignment or weak correspondence in audio-visual modality may act like outliers and disturb the model learning to generate an image from the sound. Thus, we conduct an analysis to identify which categories of audio-visual events are proper for the sound-to-image generation task.

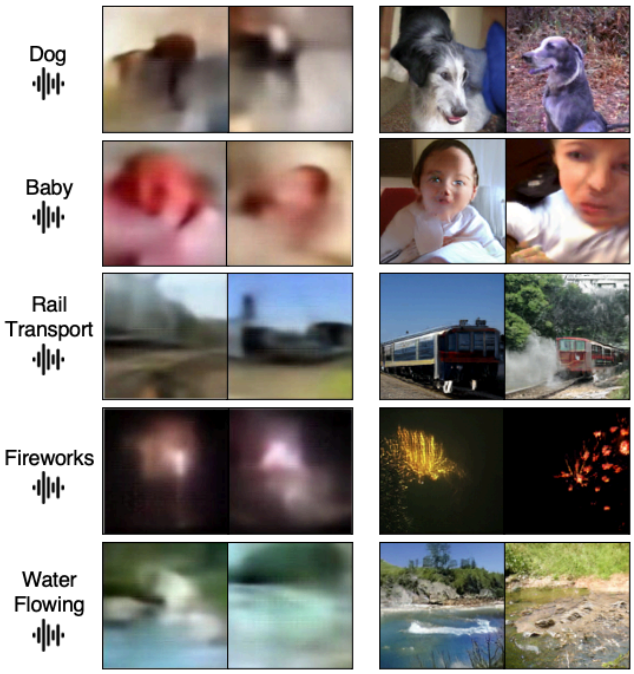
VGGSound (50 classes)

airplane flyby	cow lowing	playing timbales	underwater bubbling
ambulance siren	dog barking	printer printing	railroad car, train wagon
baby crying	elk bugling	scuba diving	train wheels squealing
baby laughter	fire truck siren	sea waves	baltimore oriole calling
car engine idling	gibbon howling	sheep bleating	car engine knocking
cat purring	hail	singing choir	driving snowmobile
chainsawing trees	lawn mowing	skiing	wood thrush calling
church bell ringing	volcano explosion	slot machine	hedge trimmer running
train horning	stream burbling	snake hissing	ice cream truck, ice cream van
train whistling	waterfall burbling	tractor digging	woodpecker pecking tree
playing bassoon	people cheering	orchestra	playing harpsichord
playing drum kit	people crowd	owl hooting	playing snare drum
playing harp	people marching		

Figure S1. **Selected audio-visual event categories.** We select 50 classes among VGGSound [3] for sound-to-image generation.

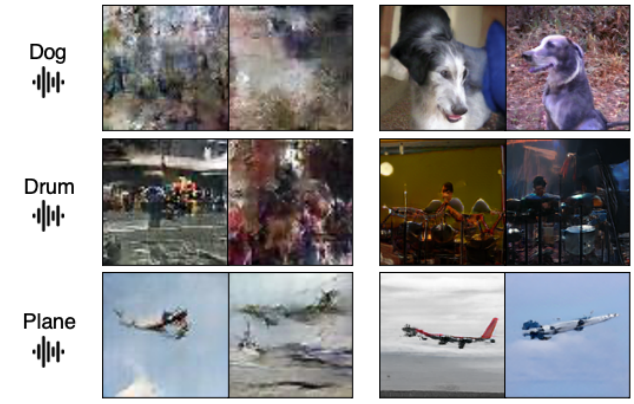
We analyze the VGGSound [3] dataset to find proper categories for the sound-to-image generation task, as large-scale benchmark datasets contain many in-the-wild videos and categories with very different characteristics. For the analysis, we first train our model with all the categories in the VGGSound dataset. Then, we evaluate the $R@1$ of the generated images for each category using the CLIP [8] retrieval metric introduced in the main paper. The categories above a certain threshold in terms of $R@1$ performance naturally reveal plausible image generation quality.

We discover that the categories related to action scenes are mostly excluded since our work focuses on a single frame generation task, which is more sensitive to the instance itself than the action in the scene. In addition, we find that as we increase the number of categories, the image quality generated by our model degrades as it brings a high chance



(a) S2I [5] (b) Ours

Figure S2. **Qualitative comparison to S2I [5].** We compare the generated images between (a) S2I and (b) Ours.



(a) Wan *et al*. [14] (b) Ours

Figure S3. **Qualitative comparison to Wan *et al*. [14].** We compare the generated images between (a) Wan *et al*. and (b) Ours.

of including improper categories.

Given the analysis, we select the categories from VGGSound by sorting in $R@1$ performance and human perception. We show the top-50 selected categories in Fig. S1 that are suitable for the sound-to-image generation task. Since our analyses in the main paper are to see the quality of sound-to-image generation with much more diverse classes than the prior arts, we use those selected classes in all the experiments of the main paper. It contains more diverse categories with different levels of audio-visual correspondences compared to the existing methods that come with a small number of categories in which images and sounds are closely correlated.

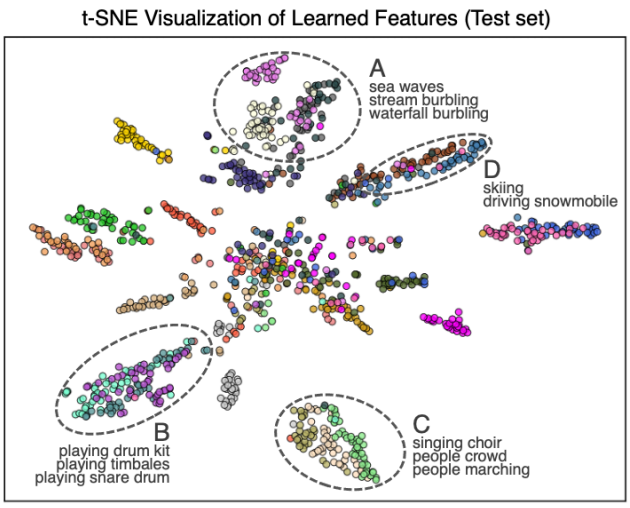


Figure S4. **t-SNE visualization [13] of learned features.** We sample 25 classes from VGGSound and visualize the learned audio features of the test set. For visualization purposes only, we color the features in terms of class labels, and no labels are used for training.

Discovering more proper videos and filtering outliers to enhance the sound-to-image generation task is an interesting research direction and needs further investigation.

C. Comparison with the Prior Arts

We show qualitative comparison with our model and prior arts, Sound-to-Imagination (S2I) [5] and Wan *et al*. [14]. We obtain the generated images of the prior work directly from their published results. Thus, the input audio for each generated image and the training dataset for each model is different. However, the purpose of this comparison is only to show how well our model and existing methods can generate images for given categories. The image size varies depending on the models; our model generates 128×128, S2I generates 96×96, and Wan *et al*. generate 64×64 pixels images.

The comparison results of the overlapping sound categories for S2I and Wan *et al*. are shown in Fig.S2 and S3, respectively. While S2I preserves the coarse shapes of dogs or babies and contains scenes that are relevant to the input sound, the images are too blurry to clarify detailed depictions. Wan *et al*. also produces the coarse shape of the plane but fails to produce informative images on other categories. In contrast, our model consistently generates visually plausible and detailed images aligned with the given sound category.

D. Additional Qualitative Analysis and Results

t-SNE visualization of audio features. We show t-SNE visualization [13] of the learned features of our audio encoder in Fig. S4. As shown, our audio encoder segregates input audios into clusters correlated with their audio-visually related classes. For example, three water-related clusters, which include similar visual scenes and also the sounds, are located

Figure S5. **Qualitative results of sampled categories in VGGSound [3] test set.** Sound2Scene can generate visually plausible images from diverse in-the-wild sounds. Note that our method does not use any class information during training and inference.

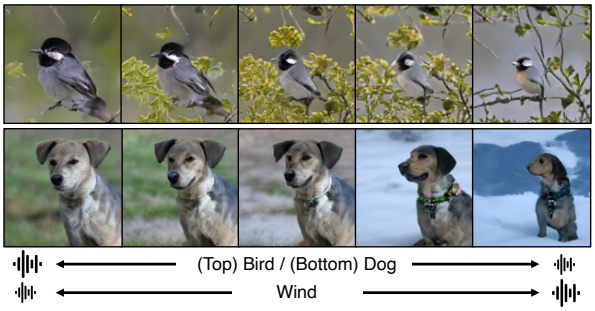


Figure S6. **Generated images by mixing multiple audios with volume changes in the *waveform space*.**

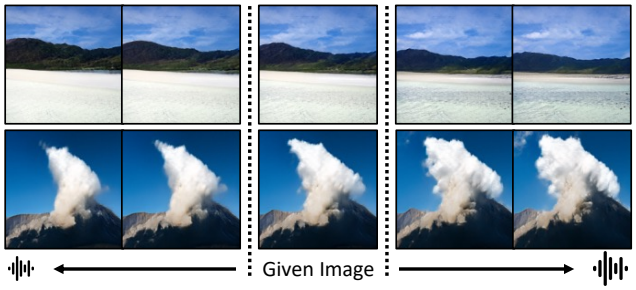


Figure S7. **Image editing by volume changes in the *latent space*.** We move the extracted visual feature in the direction of the volume difference between the two audio features.

closely (**A** in Fig. S4). In addition, drum-related videos are also similar in terms of audio-visual information and are located closely (**B** in Fig. S4). Although our model mostly embeds the input audios to audio-visually related clusters, several clusters are closely located in terms of visual information. This is expected as no class-level supervision is provided, but only the visual features are used. For example, the sound of "singing choir", "people crowd", and "people marching" are different from each other and clustered separately, but they are closely located in terms of their visual similarity (**C** in Fig. S4), and the similar results are shown with "skiing" and "driving snowmobile" (**D** in Fig. S4).

Additional generated images from different sounds. Additional qualitative results for generating images from single waveform are shown in Fig. S5. Each image is generated from different sounds without providing any class information to the model. As shown, our model can handle different categories of sounds, such as from animals, and vehicles, to diverse sceneries, and generate plausible results conditioned on the given sound. Generated images generally preserve the semantics of the scenes properly, such as the "Chainsawing" action appearing in the middle of the forest scene or "Lawn Mowing" images on grass instead of asphalt roads.

We further show the generated images by mixing multiple audios with volume changes in Fig. S6. For example, by decreasing the volume of the "Dog" while increasing the "Wind" sound, a close-up shot of the dog starts disappearing and a wide-shot in the snowy environment (windy) with a smaller dog appears gradually.

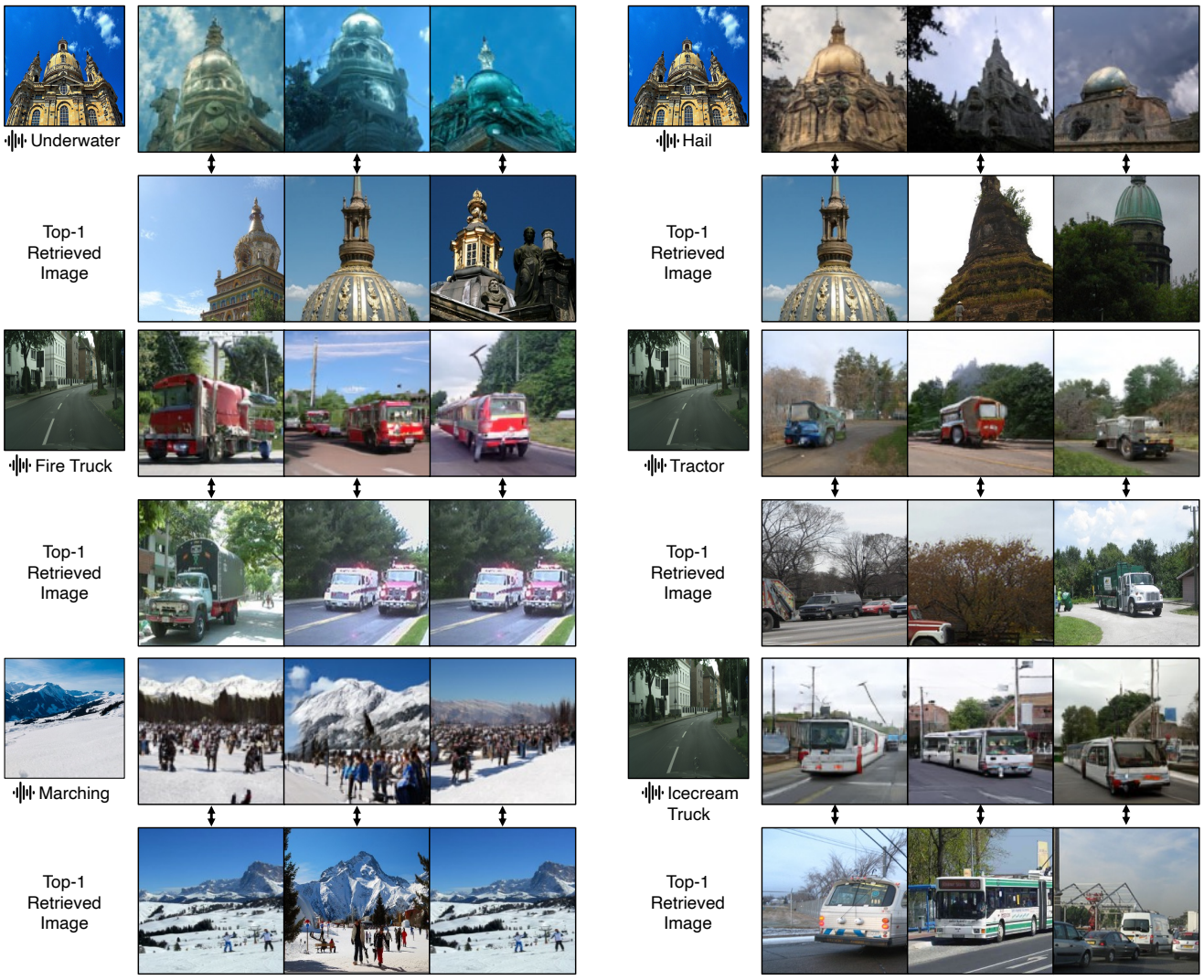
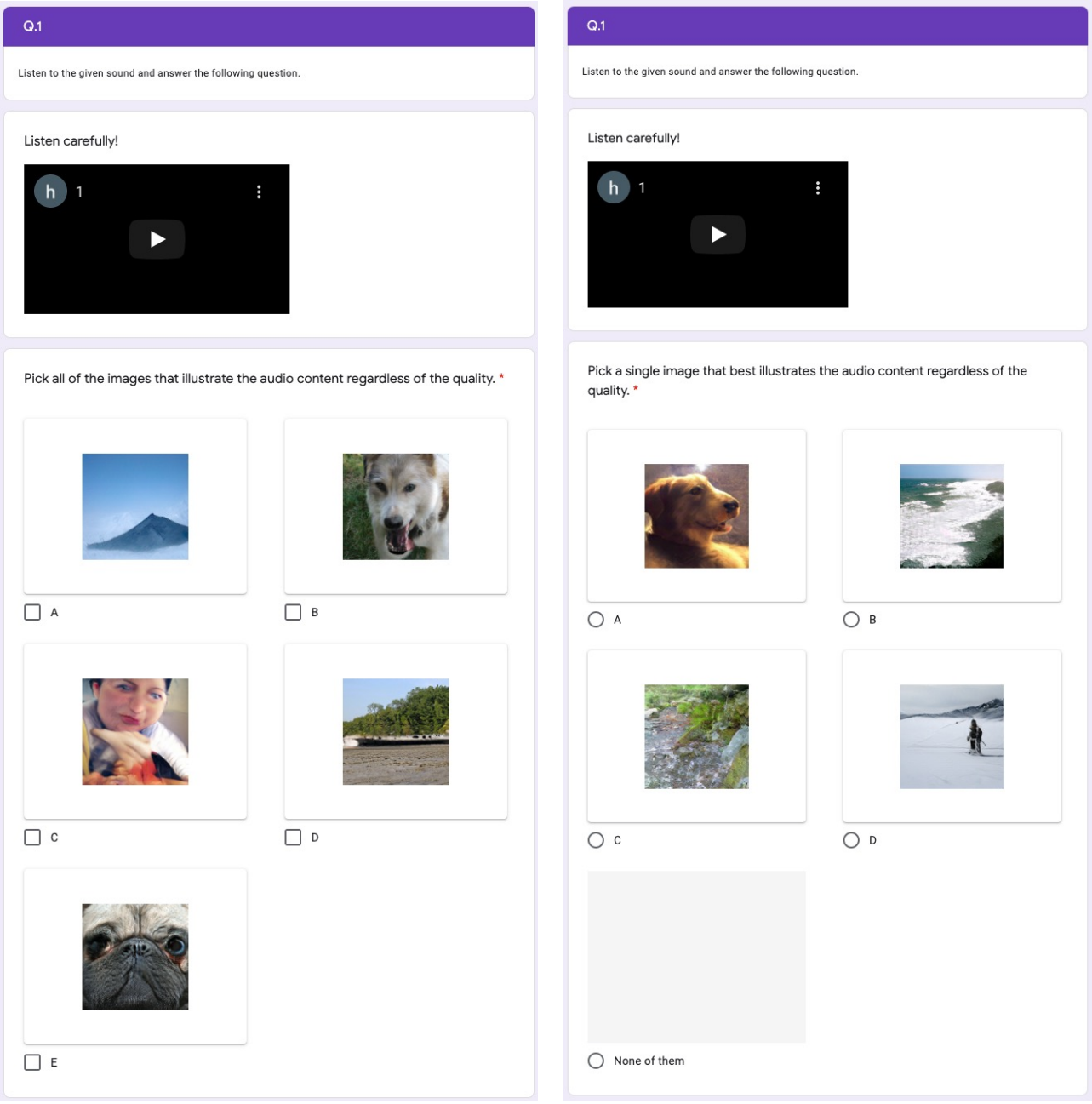


Figure S8. **Generated images conditioned on sound and image.** We simply interpolate between a visual feature and an audio feature in the joint latent embedding space. This interpolated feature is then fed to the image generator to generate a novel image. The first row per sample shows the generated images, while the row below contains the images from the ImageNet [4] training set, which are retrieved by measuring cosine similarity in the latent space of the image encoder, $f_V(\cdot)$.

Additional generated images from latent manipulations. As introduced in the main text, Sound2Scene provides latent space manipulations to generate images conditioned on both audio-visual signals. For the first approach, we can edit the given image by the given paired audio. We extract a visual feature and the noise vector by GAN inversion [1, 9] and move the visual feature toward the volume change direction of the audio features. Then, the manipulated visual feature and the noise vector are fed to the image generator, $G(\cdot)$, to generate an edited image. As shown in Fig. S7, the explosion of the given image gets smaller while we move the visual feature toward the volume-decreasing direction while getting bigger by moving toward the volume increase direction.

Furthermore, as a second approach, we can simply interpolate between the audio and visual features and generate a novel image by conditioning on both audio-visual signals, as shown in Fig. S8. We can stylize the building to be on a cloudy day or underwater, insert diverse vehicles on the road, or even insert people in the snowy fields. Moreover, we compare the generated images with the closest samples in ImageNet [4]. For example, we see that for generated images conditioned on the building and "Underwater" sound, no buildings in the closest images are underwater but with the blue sky; or we observe that the closest images rarely contain many people in the scene for the generated images conditioned on a snowy mountain and "Marching" sound. These examples show that our model generates new unique images rather than memorizing the training set.



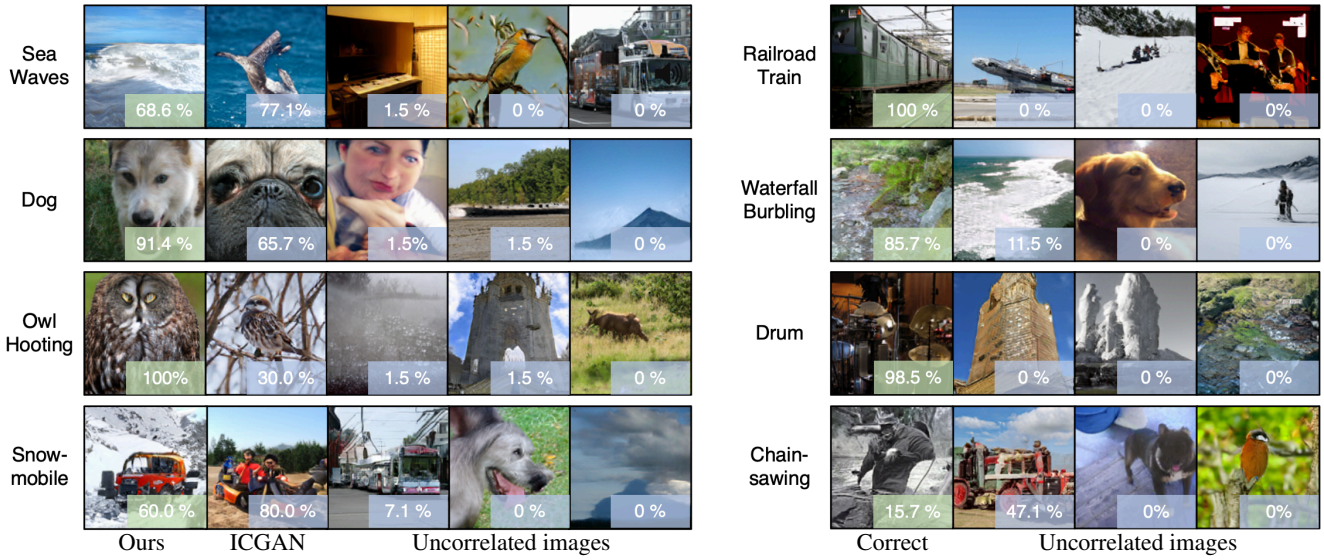
(a) Comparison to ICGAN [2]

(b) Validation of proper image generation

Figure S9. **Examples of the user study.** We conduct the user study by comparing with ICGAN in (a) and validating the proper image generation in (b). The images provided in the user study are randomly ordered.

E. Details of the User Study

We conduct a user study to analyze the performance of our method from the human perception perspective. The user study questionnaire interface is shown in Fig. S9. Users listen to the given audio, see the generated images and make a selection without any time limitation. This user study contains two experiments with 20 questions in each, as described in the main paper. The first experiment is about the comparison to ICGAN [2]. Audio and five images are given to the participants as Fig. S9 (a). Among five images, two are generated by our model and ICGAN, respectively, from the given sound or its paired image. The rest are generated images from random categories of the sounds. The users are asked to pick all the images that illustrate the given sound. As shown in



(a) Comparison to ICGAN [2]

(b) Validation of proper image generation

Figure S10. **Samples of human evaluation results.** Each row denotes each question, and the corresponding audio is described with text. We compare the generated images with ICGAN in (a). Conditioned on the given audio-image pair (audio for ours and image for ICGAN), the two left images are generated by our model and ICGAN, respectively. Three remaining images are generated by our model but conditioned on uncorrelated category audios. Each percentage in the colored box states the recall probability of the generated image. We validate the proper image generation of our method in (b). All images are generated by our model, but only the first column is conditioned on the given sound. Each percentage in the colored box states the selection ratio of the participants.

Fig. S10 (a), our generated images are more preferred to ICGAN. However, there are interesting results showing that the user study is highly subjective. For example, in the bottom row, even though only the image generated from our model can be considered as a snowmobile, users tend to pick the option that is more familiar than the snowmobile, a car-looking object, as the given audio is engine-like.

In the second experiment, we validate how properly our model generates images for given audio. Audio and four images are provided to the participants. Our model generates all four images, but only one image is from the given sound. Participants are asked to choose one image that best illustrates the given sound or check the {None of them} as in Fig. S9 (b). The selection ratio in Fig. S10 (b) clearly shows that our model generates highly-correlated images to the given sounds from the human perspective. We observe several interesting user subjectivities for making a selection. In the last row of Fig. S10 (b), among four images, even though the generated image in the first column seems to be more related to the given sound (looks like a human is in the position of using a chainsaw), users select the second image containing a vehicle in the scene. We assume that for people who are not experts in the audio-visual domain, it may be challenging to differentiate similar sounds (engine-like), *e.g.*, chainsaw, tractor, and truck sounds. Nonetheless, the overall user studies support that our model generates visually plausible images corresponding to the given sound.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[2] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5, 6

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 1, 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 4

[5] Leonardo A Fanzeres and Climent Nadeu. Sound-to-imagination: Unsupervised crossmodal translation using deep dense network architecture. *arXiv preprint arXiv:2106.01266*, 2021. 2

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman,

and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[9] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1

[11] Arda Senocak, Junsik Kim, Tae-Hyun Oh, Hyeonggon Ryu, Dingzeyu Li, and In So Kweon. Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 1

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008. 2

[14] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. Towards audio to scene image synthesis using generative adversarial network. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019. 2

[15] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. In *British Machine Vision Conference (BMVC)*, 2020. 1