Supplemental Material to: Sound-guided Semantic Video Generation

Seung Hyun Lee[1], Gyeongrok Oh[1], Wonmin Byeon[2], Jihyun Bae[1], Chanyoung Kim[1], Won Jeong Ryoo[1], Sang Ho Yoon[3], Hyunjun Cho[1], Jinkyu Kim[4*], and Sangpil Kim[1*]

[1]Department of Artificial Intelligence, Korea University
[2]NVIDIA Research, NVIDIA Corporation
[3]Graduate School of Culture Technology, KAIST
[4]Department of Computer Science and Engineering, Korea University

Overview

This supplementary material provides implementation details including the dataset details for sound-guided video generation in Section A. We also provide an additional ablation study with an effect of CLIP prior knowledge for sound inverting into the StyleGAN latent space and change in movement size according to hyperparameters as listed in Section B. Next, in Section C, we provide our detailed analysis of the movement of latent code's trajectory. Lastly, in Section D, we provide more diverse qualitative results with a variety of sound sources. Moreover, we also provide diverse examples in the following project website: https://kuai-lab.github.io/eccv2022sound/.

A. Implementation Details

CLIP-based Sound Representation Learning. We use the VGG-Sound [2] dataset to create Lee's [12] audio-visual embedding space. VGG-Sound is a large-scale audio-visual dataset including more than 310 classes with over 200,000 video clips. VGG-Sound dataset was collected from YouTube under unconstrained conditions with audio-visual correspondence. We train the encoder with VGG-Sound dataset to project audio embedding space into CLIP [15] embedding space. Thus, since the CLIP embedding space is aligned with the three modalities: text, image and audio, CLIP embedding space can be utilized as prior knowledge to create image frames with the semantic meaning of audio by handling StyleGAN inversion.

Dataset Details. Conventional audio-video pair public dataset does not consist with high quality audio and fidelity video [2,10]. Therefore, we curate and release a new high-fidelity landscape scene dataset for training our video generation model. Table 1 shows the number of clips per scene class and the distribution of

*Corresponding authors: Jinkyu Kim and Sangpil Kim.

clip resolutions. We manually make a list of Youtube videos following three principles. First, both the video and the sound should be relevant to the scene class. Second, there should be semantic consistency between the video and the audio. Third, the resolution should be higher than 1280×720. With the principles, we collect the high-fidelity audio-video dataset and the high-quality examples are shown in Fig. 1. In addition, Fig. 2a shows that the Landscape dataset has the best video resolution compared to the other audio-visual datasets, such as the VGG-Sound [2], Kinetics-400 [10].

Implementation Details for Quantitative Comparison. We implement Sound2sight [1] and CCVS [11] using the same settings based on the official implementation from each repository . We generate the videos by conditioning on an image with these baselines on the audio-visual datasets (High Fidelity Audio-Video Landscape dataset and the Sub-URMP dataset [13]). Our landscape dataset contains ten seconds clips, and we sample about ten frames from each clip. The Sub-URMP dataset contains 80,805 sound-image pairs. For both datasets, we resize all the frames to 256×256 and sample audio features with the one second window size and MFCC [4] from each clip. We generate 16 frames on each model for comparing our model. The details of the discriminator implementation are as follows. Following the PatchGAN [6] structure, 2D-Convolution, Instance Normalization [18], and Leaky ReLU layers are repeatedly stacked four times for D_I, and 3D-Convolution, Instance Normalization, and Leaky ReLU layers are repeatedly stacked four times for D_V.

StyleGAN Generator. We use the pre-trained StyleGAN3 [7] generator where its generated internal representations are known to be fully equivariant to subpixel-level translations and rotations, which should benefit video generation models. We have found earlier StyleGAN models [8,9] often fail to generate fine-grained consistent videos. In our experiment, the latent code is set to 16×512, and the output image resolution is set to 256×256. To generate high-quality landscape videos, our StyleGAN generator is pre-trained with Landscapes HQ (LHQ) [17] dataset, which provides over 90k high-resolution (at least 1024×1024) landscape images. Due to heavy computational burdens, images are resized to 256×256 by applying Lanczos interpolation following [17].

Sound Inversion Encoder. Audio inputs are first converted into the Mel spectrogram representation, which is then fed into a ResNet [5] backbone. As the input of the backbone network, the mel spectrogram is split by the number of frames of the video to be generated. Following recent pSp (pixel2style2pixel) [16] architecture, we extract feature maps from a standard feature pyramid over the backbone. To generate images with the finer details, we further use a small fully-convolutional mapping network trained to extract different levels of detail (or style). These are then fed into the StyleGAN [8] generator according to their resolution, which generates the final output image. We provide a more detailed process in the supplemental material.

---

https://github.com/16lemoing/ccvs
https://www.merl.com/research/license/Sound2Sight

Sound-based Frame Generator. Our video frame generator uses 5 GRU [3] cells to generate the latent code of StyleGAN for the next time step. The latent code of the next time step is created by adding the latent code inverted from the sound, and the vector predicted from the latent code of the previous time step. The length of the input sequence to the video generator is 10 and the batch size is 2. The dimension size of the hidden layer included in each GRU cell is 512. To disentangle the motion, the last fully connected layer feeds a 512-dimensional encoded sound vector and a 512-dimensional noise vector concatenated. After that, the final output has a value obtained by adding the input latent code and the output value of the layer.

B. Ablation Study

Effect of CLIP Prior Knowledge for Sound-guided Video Generation. We conduct an ablation study to demonstrate the effectiveness of applying CLIP Loss [15]. In the main manuscript, we demonstrate that CLIP Loss improves GAN inversion reconstruction performance. Leveraging the CLIP multi-modal embedding space helps to increase the video generation quality (see Fig. 3). For example, given an ocean wave sound as input, we expect to generate a video associated with the 'wave'. These show the video generation results with and without CLIP loss, respectively. Using CLIP loss makes the content of the generated video result more relevant to the audio.

Relationship between Identity and Moving Distance in StyleGAN Latent Space. There is a trade-off between the identity of the initial frame and variation on style change. We study the relationship through equations $\hat{\mathbf{w}}_a^{t+1} = (1-\alpha)\cdot\hat{\mathbf{w}}_a^t + \alpha\cdot\hat{\mathbf{w}}_a^{t+1}$. We compare the results generated by changing the hyperparameter value α. For example, for a given rain sound, as the α increases, there are more dark clouds in the next generated frame, and the style change increases. Movement size in the StyleGAN latent space and identity of the latent code predicted by the recurrent block has a tradeoff relationship (see Fig. 4).

C. Direction Analysis of Sound-guided Latent Code

Sound-guided Latent Code Trajectory. In our model, the sound provides appropriate guidance for video generation in StyleGAN [8] latent space. This guidance provides generated video that is consistent with the meaning of sound. To demonstrate this, we compare the video generation results with random noise input instead of sound input to our model (see Fig. 5). The temporal consistency of the video is maintained when the latent code guides the direction of movement with sound. However, in random directions, even though the initial latent code has clear content, the temporal consistency of the video decreases. We also observe that series of different chunks of sounds would provide distinctive guidance based on embedded semantics (see Fig. 6).

Table 1: High Fidelity Audio-Video Landscape dataset statistics. The number of clips per scene class and the distribution of clip resolutions.

		Thunderstorms	Waterfall Burbling	Volcano Eruption	Squishing Water	Wind Noise	Fire Crackling	Raining	Underwater Bubbling	Splashing Water	Total
Resolution	3840×2160	10	10	0	133	105	175	10	70	70	583
	2560×1440	230	200	60	1,342	1,665	1,014	1,674	190	1,650	8,025
	1920×1080	10	10	0	15	10	0	0	0	0	45
	1280×720	30	50	0	120	55	176	141	0	55	627
Sum		280	270	60	1,610	1,835	1,365	1,825	260	1,775	9,280

Distribution of CLIP Embeddings in Images Generated from Sound-Inverted Latent Code. We qualitatively evaluate whether the video generated from a given sound is semantically consistent. For this, we generate 120 frames from the sound and visualize the CLIP image embeddings extracted from the first and last frames. Fig. 7 shows visualization results of the extracted embeddings with t-SNE [14]. We show that videos generated from different classes of sound are semantically separated.

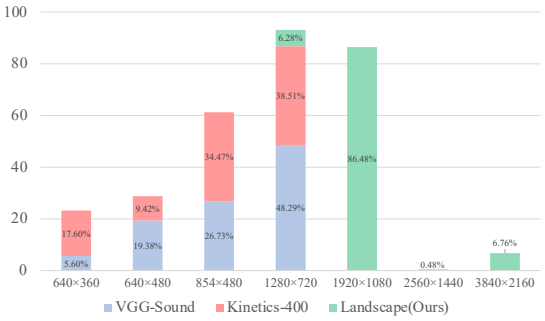
D. Qualitative and Quantitative Results

Additional User Study. We conduct an additional user study on the Sub-URMP dataset [13]. As mentioned in the paper, we recruit 25 participants from Amazon Mechanical Turk (AMT) to evaluate our proposed model. As shown in Fig. 10, our method significantly outperforms other state-of-the-art approaches (*Realness*, *Naturalness*, *Semantic Consistency*).

Sound-guided Image Editing. By using the mapping function only one time step, our mapping function is effective for sound-guided image editing. Lee *et al.* [12] determines the direction of the latent code using CLIP's prior knowledge via the latent vector optimization process, which requires 200-300 iterations of about 1 minute for editing one image. On the other hand, we observe that it takes 6 minutes 10 seconds for our model to edit 1000 images with sound as input (see Fig. 11.). The implementation details of the sound inversion encoder for sound-guided image editing are as follows. The hyperparameter λ_b to minimize the loss $\mathcal{L}_{\text{enc}}$ is set to 1.0.

Additional Qualitative Examples. We show more diverse results with the Sub-URMP dataset and our High Fidelity Audio-Video Landscape dataset in Fig. 8 and Fig. 9. In particular, the video results generated by our model show that sound and video are semantically consistent. Fig. 12 shows that our model is applicable to many cases like face video generation. Note that we used a small face video set different from the dataset we used for StyleGAN pre-training.

Fig. 1: Examples of sampled frames illustrating the high video quality of our High Fidelity Audio-Video Landscape Dataset. We sample 15 frames from each video. For the bottom left of each frames, f denotes the order of frame in the video.



(a)

Resolution	VGG-Sound [2]	Kinetics-400 [10]	Landscape (Ours)
640 × 360	9,892	52,333	0
640 × 480	34,187	27,999	0
854 × 480	47,166	102,508	0
1280 × 720	85,197	114,504	583
1920 × 1080	0	0	8,025
2560 × 1440	0	0	45
3840 × 2160	0	0	627

(b)

Fig. 2: Comparison between our Landscape dataset and other audio-visual datasets, such as VGG-Sound [2] and Kinetics-400 [10], in terms of the video resolution. Because all the datasets contain 10-seconds clips, the duration of the dataset is proportional to the number of the clips. (a) Distribution of video resolution of the datasets. We report the percentage of clips corresponding to the resolution for each dataset. (b) For the datasets, we report the number of clips for each resolution.

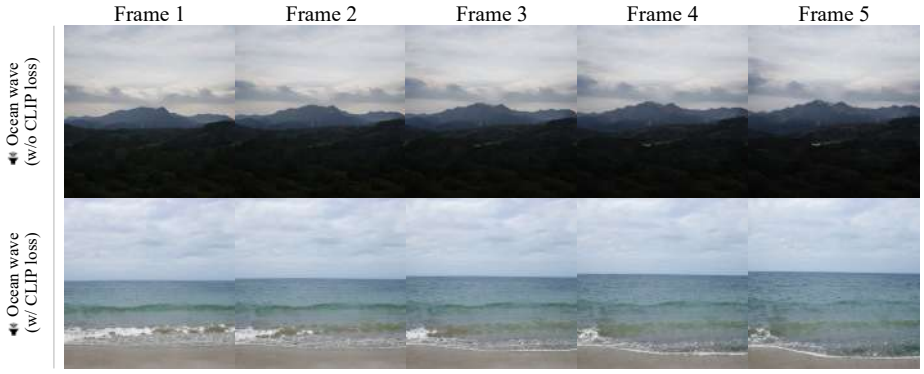


Fig. 3: Ablation study according to the CLIP [15] loss. CLIP loss can produce video with sound and video semantic consistency. The time interval between each column is five frames.

Fig. 4: Ablation study according to the hyperparameter α from the equation $\hat{\mathbf{w}}_a^{t+1} = (1-\alpha) \cdot \hat{\mathbf{w}}_a^t + \alpha \cdot \hat{\mathbf{w}}_a^{t+1}$. As α increases, style change in latent space increases, and as α decreases, movement in latent space becomes smaller, resulting in less style change. The time interval between each column is five frames.

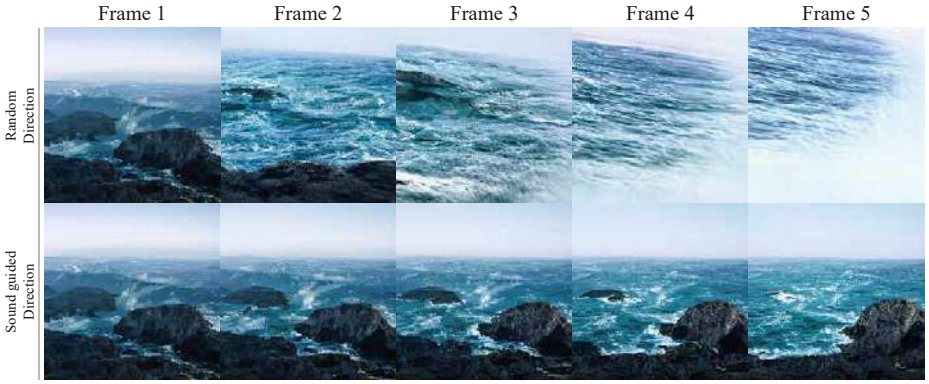


Fig. 5: Comparison of video generation results with and without sound guidance. When sound is not given as guidance and the direction of movement of the latent space is randomly given, frame consistency is reduced. The time interval between each column is five frames.

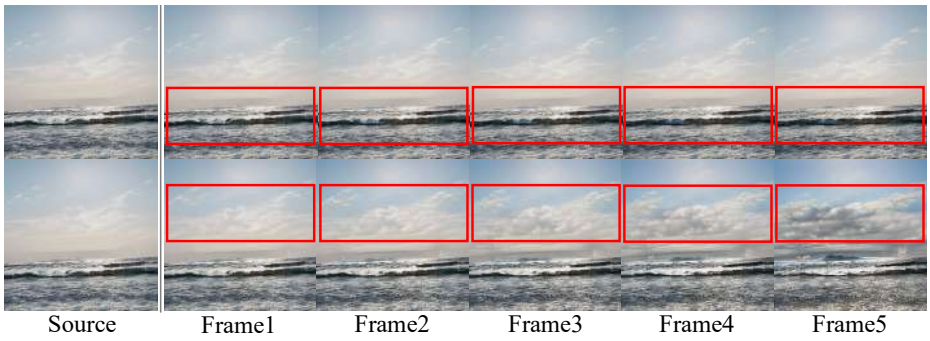


Fig. 6: Sounds with one or more semantic transitions. The images in the first row are guided by the sound of waves whereas the second row by the wind.

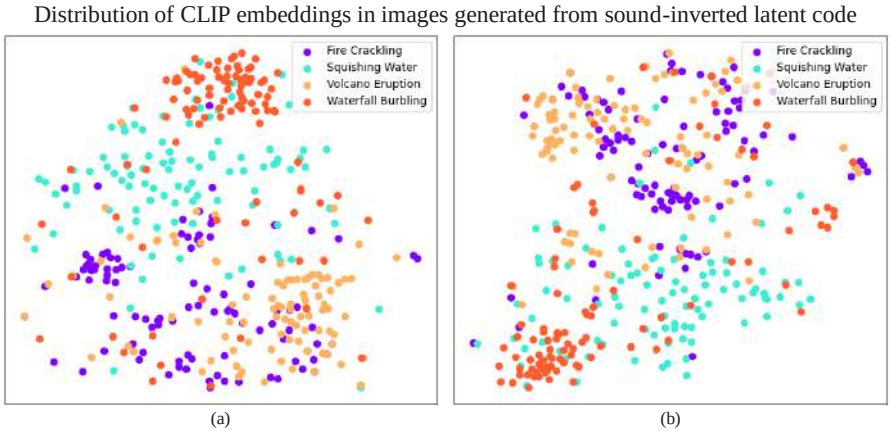


Fig. 7: Visualization of CLIP [15] embeddings in images generated from sound-inverted latent code with t-SNE [14].

Fig. 8: Example of sound-guided semantic video generation from the Sub-URMP [13] Dataset. Our method takes audio as a given input feature (left) and produces a video with audio semantics and temporally consistent (right). The time interval between each column is five frames.

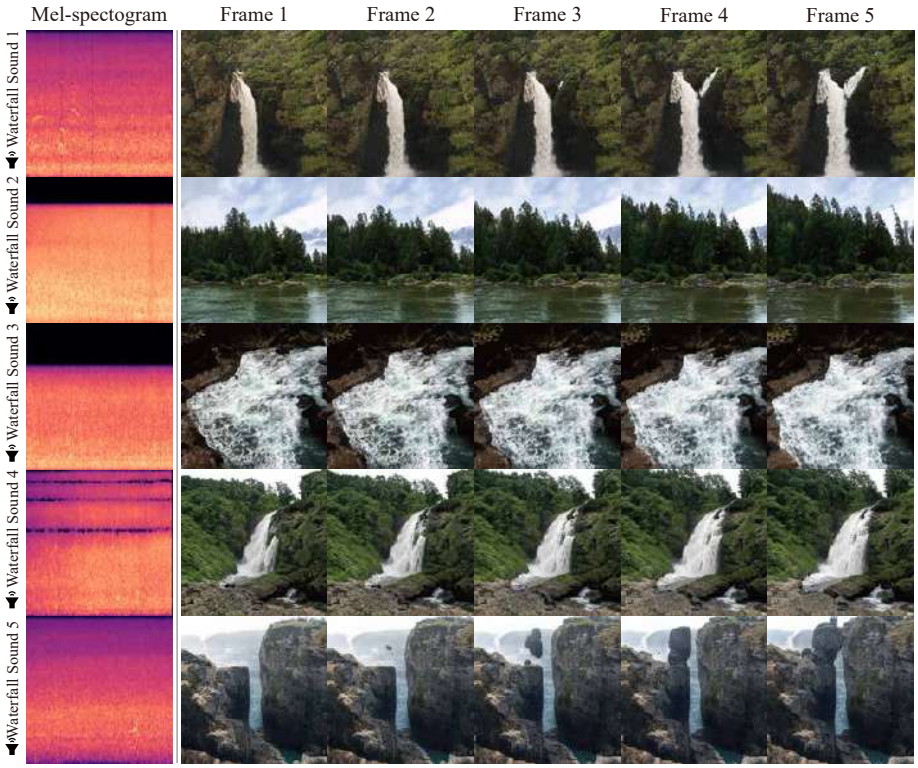
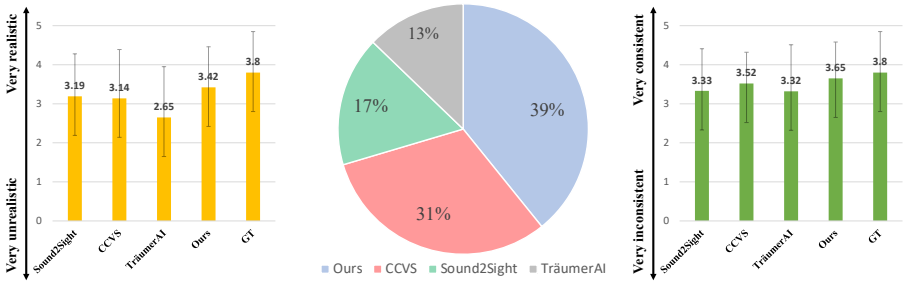


Fig. 9: Example of sound-guided semantic video generation from Landscape Dataset. Our method takes audio as a given input feature (left) and produces a video with audio semantics and temporally consistent (right). The time interval between each column is five frames.



(a) Realness. (b) Naturalness. (c) Semantic Consistency.

Fig. 10: User Study Results on the Sub-URMP dataset [13]. (a) Realness. (b) Naturalness. (c) Semantic Consistency. We use a 5-point Likert scale for (a) and (c).

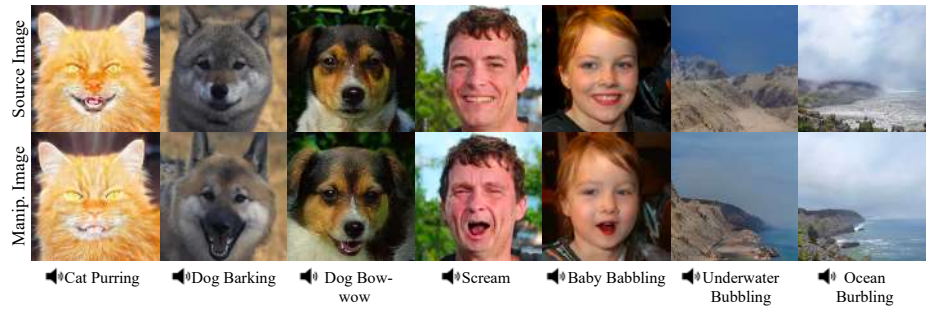


Fig. 11: Examples of images edited with the sound inversion module. Our inversion module produces various image editing results based on the input sound.

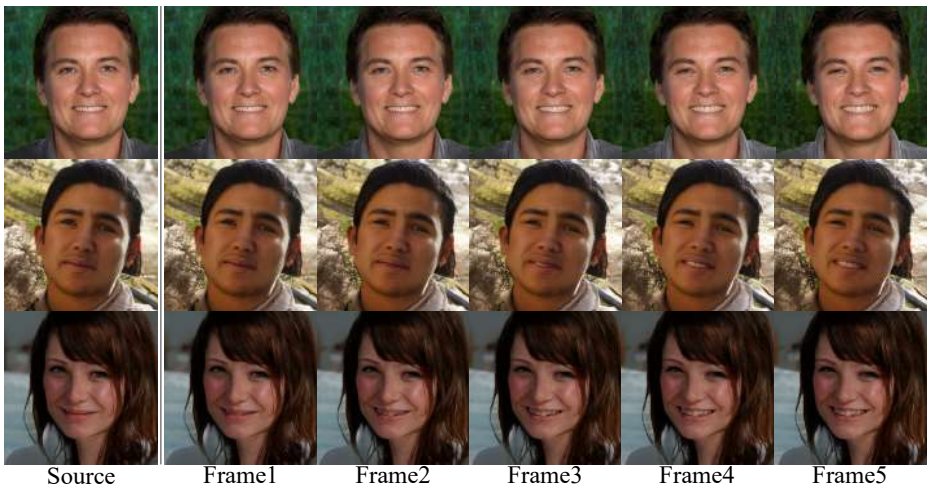


Fig. 12: Examples of sound-guided face video generation. The images are generated by the sound of laughing.

References

1. Chatterjee, M., Cherian, A.: Sound2sight: Generating visual dynamics from sound and context. In: European Conference on Computer Vision. pp. 701–719. Springer (2020)
2. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020)
3. Chung, J., Çaglar Gülçehre, Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv **abs/1412.3555** (2014)
4. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
7. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34** (2021)
8. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
9. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
10. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
11. Le Moing, G., Ponce, J., Schmid, C.: Ccvs: Context-aware controllable video synthesis. Advances in Neural Information Processing Systems **34** (2021)
12. Lee, S.H., Roh, W., Byeon, W., Yoon, S.H., Kim, C.Y., Kim, J., Kim, S.: Sound-guided semantic image manipulation. arXiv preprint arXiv:2112.00007 (2021)
13. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Transactions on Multimedia **21**(2), 522–535 (2018)
14. van der Maaten, L., Hinton, G.E.: Visualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (2008)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. Image **2**, T2
16. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021)
17. Skorokhodov, I., Sotnikov, G., Elhoseiny, M.: Aligning latent and image spaces to connect the unconnectable. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14144–14153 (October 2021)

18. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. CoRR **abs/1607.08022** (2016), http://arxiv.org/abs/1607.08022