Supplementary Material: Conditional Generation of Audio from Video via Foley Analogies

A.1. VQGAN codebook loss

We follow the formulation of Iashin and Rahtu [29], which adapts the VQGAN loss [13, 57] to spectrograms. The loss $\mathcal{L}$ is composed of three individual components: the reconstruction and codebook loss [34, 57] $\mathcal{L}_{VQVAE}$, the perceptual loss [32] $\mathcal{L}_{perceptual}$, and a patch-based discriminator loss [31] $\mathcal{L}_{disc}$. We present the final loss we used to learn the codebook:

$$\mathcal{L}_{VQGAN} = \mathcal{L}_{VQVAE} + \mathcal{L}_{perceptual} + \mathcal{L}_{disc} \quad (4)$$

We describe these losses next.

Reconstruction and codebook loss. Given the input waveform $\mathbf{a}$ and its mel spectrogram $\mathbf{S} = \text{MSTFT}(\mathbf{a}) \in \mathbb{R}^{T \times F}$, where T and F are the dimension for time and frequency, the encoder E will encode $\mathbf{S}$ into embeddings $\hat{\mathbf{z}}_{tf}$ and corresponding quantized code $q(\hat{\mathbf{z}}_{tf}) = \mathbf{z}_{tf}$. The decoder D will then reconstruct $\mathbf{z}_{tf}$ to a mel spectrogram $\hat{\mathbf{S}} = D(q(E(\text{MSTFT}(\mathbf{a}))))$. Following VQVAE, we minimize the reconstruction loss between $\mathbf{S}$ and $\hat{\mathbf{S}}$ as well as the *codebook loss* between $\hat{\mathbf{z}}_{tf}$ and $\mathbf{z}_{tf}$:

$$\mathcal{L}_{VQVAE} = \underbrace{||\mathbf{S} - \hat{\mathbf{S}}||}_{\text{recon. loss}} + \underbrace{||\hat{\mathbf{z}}_{tf} - \text{sg}[\mathbf{z}_{tf}]||_2^2 + ||\text{sg}[\hat{\mathbf{z}}_{tf}] - \mathbf{z}_{tf}||_2^2}_{\text{codebook loss}}. \quad (5)$$

where sg is the stop gradient operation.

Perceptual loss. We use the off-the-shelf VGGish-ish model [29], a variation of the VGGish model with VGG-16 backbone trained on the VGGSound dataset, to evaluate the perceptual loss. For the i^{th} level of features of the original and the reconstructed mel spectrogram, $\mathbf{S}^i$ and $\hat{\mathbf{S}}^i$ respectively, the corresponding perceptual loss is given by their squared L2 distance:

$$\mathcal{L}_{perceptual} = \sum_{i=1}^{N} \frac{1}{F^i T^i} ||\mathbf{S}^i - \hat{\mathbf{S}}^i||_2^2, \quad (6)$$

where N is the number of layers we selected to calculate the perceptual loss. We use $N = 5$ layers in our model, selected as in [29].

Patch-based discriminator loss. We use the discriminator loss introduced by Isola *et al*. [31]:

$$\mathcal{L}_{disc} = \log \mathcal{D}(\mathbf{S}) + \log(1 - \mathcal{D}(\hat{\mathbf{S}})), \quad (7)$$

where $\mathcal{D}$ is applied to a fully convolutional network at multiple scales as a discriminator.

A.2. Dataset

All data comes from the *Greatest Hits* dataset, a lab-collected dataset from Owens *et al*. [42], and the *CountixAV* [66] dataset. The *Greatest Hits* dataset is composed of 977 videos (11 hours) of a drumstick interacting with different objects in the scenes. The *CountixAV* [66] dataset is composed of 1483 videos (4.1 hours) of repeated actions including bouncing ball, skipping rope, and other actions in realistic scenarios.

To evaluate our conditional Foley generation task, we randomly sample 2-sec video clips from the test videos in the *Greatest Hits* [42] dataset as input videos. We randomly pair each silent input video with 3 conditional audio-visual clips from different test videos. We obtain the conditional Foley evaluation set of 582 input-condition pairs with 17 different materials and two action types.

We also evaluate our model on the *CountixAV* [66] dataset by randomly choosing 2 seconds clips from the videos. We apply a noise reduction algorithm [48] to improve the sound quality before training and evaluation. Considering the lack of the permission of the videos in the *CountixAV* [66] dataset, we demonstrate the result qualitatively on the publicly-sourced videos with proper permissions. We provide the credit to those videos in the Appendix A.11

A.3. Implementation details

To train our complete model, we first train the VQGAN, then train the audio predictor using its learned code. We trained the VQGAN for approximately 400 epochs with a batch size of 32 and a learning rate of 1.44×10^{-4} using Adam [33]. We trained the transformer for 50 epochs with a batch size of 8 and a learning rate of 1.6×10^{-4} using 4 NVIDIA A40 GPUs. The training of the VQGAN takes approximately 5 days and the training of the transformer takes roughly 20 hours.

Our model operates on 2 sec. video clips for both the conditional and input videos, at a 15Hz video sampling rate and 22.05 KHz audio sampling rate. The VQGAN codebook encoder produces a 12×5 time-frequency grid from a mel spectrogram, which is converted from a waveform using 80 mel bins and 1,024 Fourier filters. We use $d = 256$ for the codebook embedding dimension. The 2-sec. videos (30 frames) are randomly cropped and resized to 112×112, and are represented as 30 1024-dimensional feature vectors, obtained by performing a 1×1 convolution on the ResNet feature map. During training, we apply data augmentation in the form of frequency and temporal masking to the spectrogram prior to extracting the clips. The model is trained on both the *Greatest Hits* [42] dataset and the *CountixAV* [66]

dataset in the same manner.

During inference, we set the re-ranking tolerance to be $\tau = 0.2$.

A.4. Onset transfer baseline training details

For predicting onsets given a video, we used the same variation of ResNet (2+1)D-18 [53] visual network as our main model (i.e., after removing temporal striding). Our model outputs a vector of predictions using a fully connected layer after pooling (one onset prediction per input frame). We obtain the ground truth onset according to the timing label provided in the *Greatest Hits* [42] dataset by aligning it to the closest frame. We use a binary cross entropy loss, penalizing onset predictions that occur at incorrect times. Since each input video can have more than one sound event, we set the weight of each onset in the loss from the video equally according to the total number of onsets in the input so that the weight sums to one. The configuration of video and audio is the same as our model, we use a frame rate of 15Hz and an audio sampling rate of 22.05kHz. We train the model for 100 epochs with a batch size of 12 clips of 2 seconds and a learning rate of 1×10^{-4} using Adam [33]. The model is trained on a single NVIDIA RTX-2080 GPU.

The model is used to detect the onset in both input and conditional videos. We then randomly copy-and-paste the sound from the conditional video at onset timings. We included the onset transfer method to help analyze the generative models on Greatest Hits [40]. It simply detects onsets and copy-pastes sounds from the conditional example. By design, it (trivially) obtains near-perfect performance on the *material* metric when there is only one material and also performs well on onset metrics *because it is directly trained on onset labels*. We emphasize that this baseline is not generative and, in fact, we show that this method will fail when the action is different in the input and condition pairs (column 6 in Tab. 1). Qualitatively, we have found that it often completely fails on CountixAV [60] since there are no clear onsets. The number of onsets is rarely correct and the method fails due to background noise (*e.g.* row 1 in Fig. 5).

A.5. Sound classifier training details for quantitative experiment

We finetune two pretrained VGGish classifiers [25, 26] to predict the action or the material presented in the video based on the label provided in the *Greatest Hits* [42] dataset. With the same input video settings, we train both of the models with an early stopping criteria using Adam [33]. The model is trained on a single NVIDIA RTX-2080 GPU with a learning rate of 10^{-4} and batch size 32 video clips of 2 seconds. The trained model obtains a validation accuracy of 75.6% on the material task and 92.0% accuracy on the action prediction task.

Model	Window size		
	0.1-sec. AP (%)	0.2-sec. AP(%)	Avg. AP(%)
Style transfer* [18, 55]	46.9	46.7	58.3
Onset transfer	**71.9**	**78.4**	**76.5**
SpecVQGAN [29]	59.3	65.5	64.1
Ours - No cond.	59.3	73.2	69.8
Ours - Base	**60.0**	**74.0**	**70.0**

Table 3. **Onset synchronization window size evaluation.** We measure the average precision of onset predictions that are within different windows size of 0.1, and 0.2 seconds. We also measure the averaged AP score under different window sizes (from 0.10s to 0.25s with a step of 0.05s).

A.6. Onset detection experiment with different window size

We chose 0.1 seconds as the size of the detection window for the onset detection experiment following the standard value used for audio onset detection per `mir_eval` [46] and the *Greatest Hits* [42]. Alternatively, we have also experimented with different methods with alternative window sizes. We found that (Tab. 3) our method outperforms all the other generative baselines in the additional experiment. Meanwhile, the gap between our method and the onset transfer baseline is reduced when the window size is enlarged or averaged over different window sizes.

A.7. Generation of longer audio

Following Iashin *et al.* [29], we generate longer audio clips using a 2-sec. sliding window. During the process, we always keep the token to be generated in the center of all the visible tokens before feeding the tokens into the transformer. We have evaluated the onset sync. performance for the *Greatest Hits* dataset [40] for 454 pairs of videos. From Fig. 7, we note that our model continues to obtain strong performance for videos up to 6 sec. We provide examples of the generation of 4-sec., 6-sec., and 8-sec. in our project webpage.

A.8. Re-ranking qualitative result

We notice that the re-ranking improves the synchronization performance through the qualitative result in Fig. 6. Note the generated sound with re-ranking shows a much better synchronization performance in Row 2 and 3 in Fig. 6.

A.9. Human study details

For the human study, we recruited 609 participants from Amazon Mechanical Turk, in total. We filtered out 233 pieces of feedback according to the answers the participants

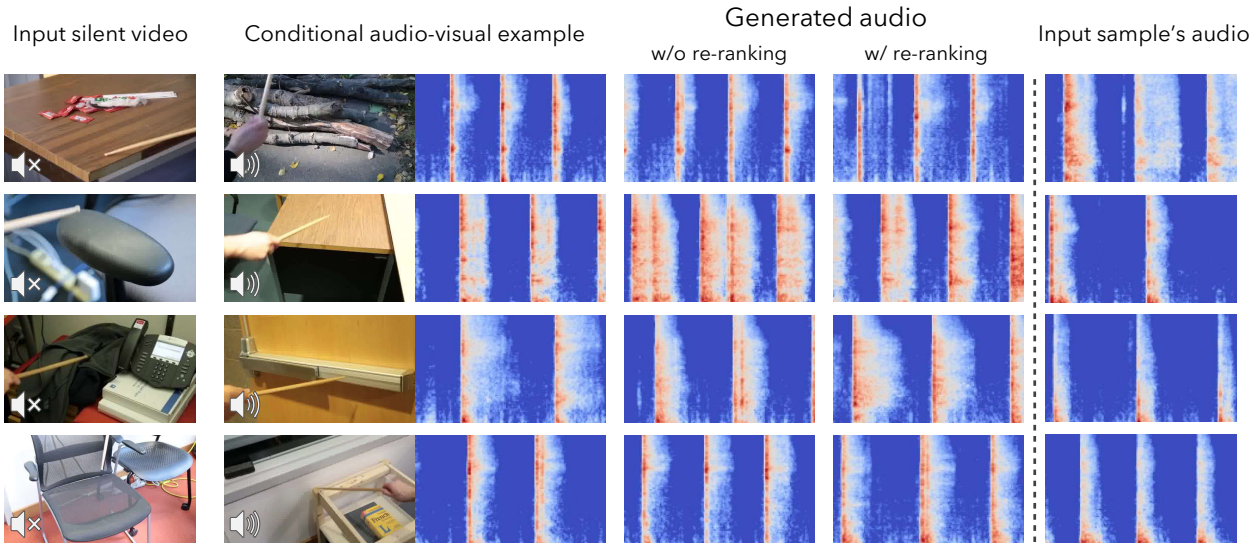

Figure 6. **Re-reranked comparison results**. We present 4 pairs of results to compare the effect of re-ranking qualitatively.

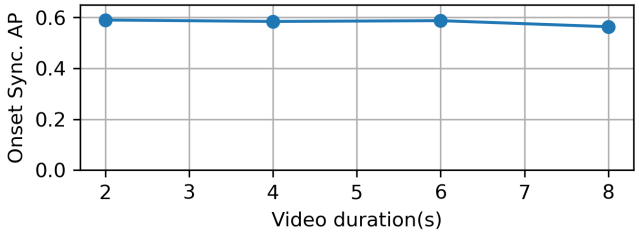

Figure 7. **Automated onset sync. evaluation for different durations**. We evaluate the performance of the onset synchronization for videos of different durations.

provided to a sentinel example, where one of the videos is attached with a clearly wrong sound that neither match with the action in the video, nor provide a timbre close to the chosen conditional video. We obtained 376 effective feedback in the end.

We provided them with the following instructions:

- Please use headphones for the test. You may hear some harsh or dissonant sounds, so please make sure to adjust your device's volume before the test.
- The task should take approximately 15 minutes to complete.
- You will take part in an experiment involving visual and hearing perception. To complete this task, you will need to watch and listen to 21 groups of videos and answer two questions for each group. Each group consists of the following videos (all videos have audio):
  - One input video: Video #1
  - Two output videos: Video #2 and Video #3
- Your task is to answer the two questions at the bottom:
  - One input video: Video #1. In which output Video (#2 or #3) is the audio most synchronized with the action in the video?
  - In which output Video (#2 or #3) does the audio sound most like the object or material in Video #1 according to the action in the output Video?
- You will complete a short practice of 5 groups of videos (about 3 minutes long) before starting the main task, so that you can get familiar with the interface.

We also provide the screenshot of the instruction page and main test page that the participant will see during the test in Fig. 8 and Fig. 9.

A.10. Randomly selected results

We provide randomly selected results in Fig. 10 (this supplemental). Please also see the provided video. It is notable that the generated audio provides a timbre very close to the conditional video in most situations. Meanwhile, the model can also generate a sound match with the timing of the actions in the input video usually.

A.11. Video Credit

We have obtained permission to use and edit the videos from the *Greatest Hits* dataset, and we have also recorded some of the videos ourselves. We provide here the credit for the publicly sourced and licensed videos that appear in the paper and the supplement.

1. fotorealis - https://stock.adobe.com/video/bambino-suona-bongo-percussioni-per-musicoterapia/536371955 - Adobe Stock Extended license.
2. FreeStockFootageClub - https://www.youtube.com/watch?v=zI1_fAtpHQc - YouTube Creative Commons CC BY license.
3. Ignat Gorazd - https://www.youtube.com/watch?v=1_ckbCU5aQs - YouTube Creative Commons CC BY license.

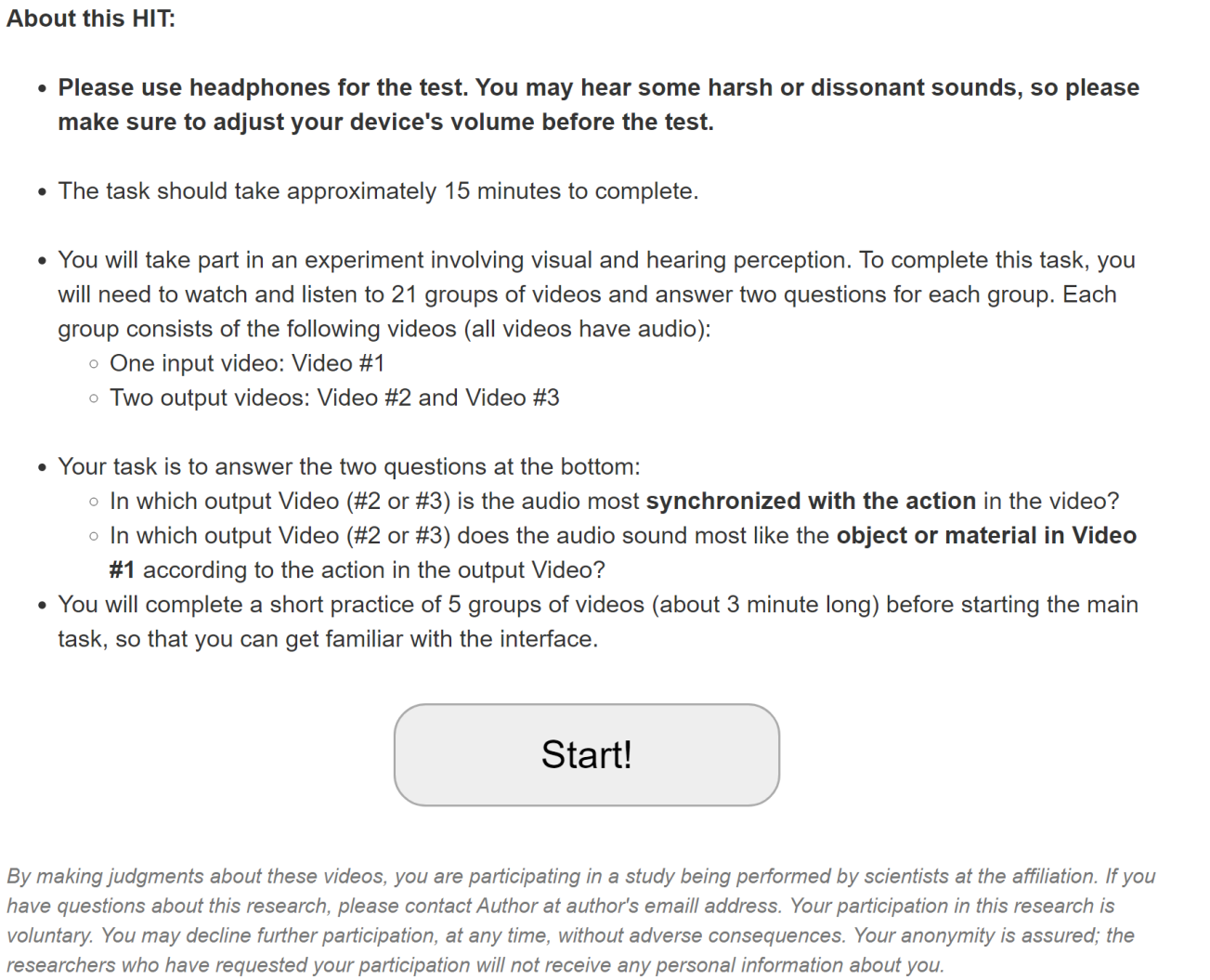

Figure 8. **Instruction page of the AMT test**. We present a screenshot of the instruction page that the participants will see at the beginning of the test. The sensitive information is removed from the image.

4. kriista - https://www.youtube.com/watch?v=6d1YS7fdBK4 - YouTube Creative Commons CC BY license.
5. Over & Out - https://www.youtube.com/watch?v=SExIpBIBj_k - YouTube Creative Commons CC BY license.
6. Over & Out - https://www.youtube.com/watch?v=XxmZxM8AtUc - YouTube Creative Commons CC BY license.
7. Percussion Play - https://www.youtube.com/watch?v=fcjfKvdkJyI - YouTube Creative Commons CC BY license.
8. Percussion Play - https://www.youtube.com/watch?v=xcUyiXt0gjo - YouTube Creative Commons CC BY license.
9. PhotoSerg - https://stock.adobe.com/video/kid-juggles-the-ping-pong-ball-green-screen/99378579 - Adobe Stock Extended license.
10. PUSAT E-PEMBELAJARAN UMS - https://www.youtube.com/watch?v=S6TkbV4B4QI - YouTube Creative Commons CC BY license.
11. Suliman Razvan - https://stock.adobe.com/video/christian-monk-hitting-a-large-wooden-piece-toaca-with-little-wooden-hammer-to-summon-the-breathren-to-prayer/93378558 - Adobe Stock Extended license.
12. Thomas Cremier - https://www.youtube.com/watch?v=GFmuVBiwz6k - YouTube Creative Commons CC BY license.

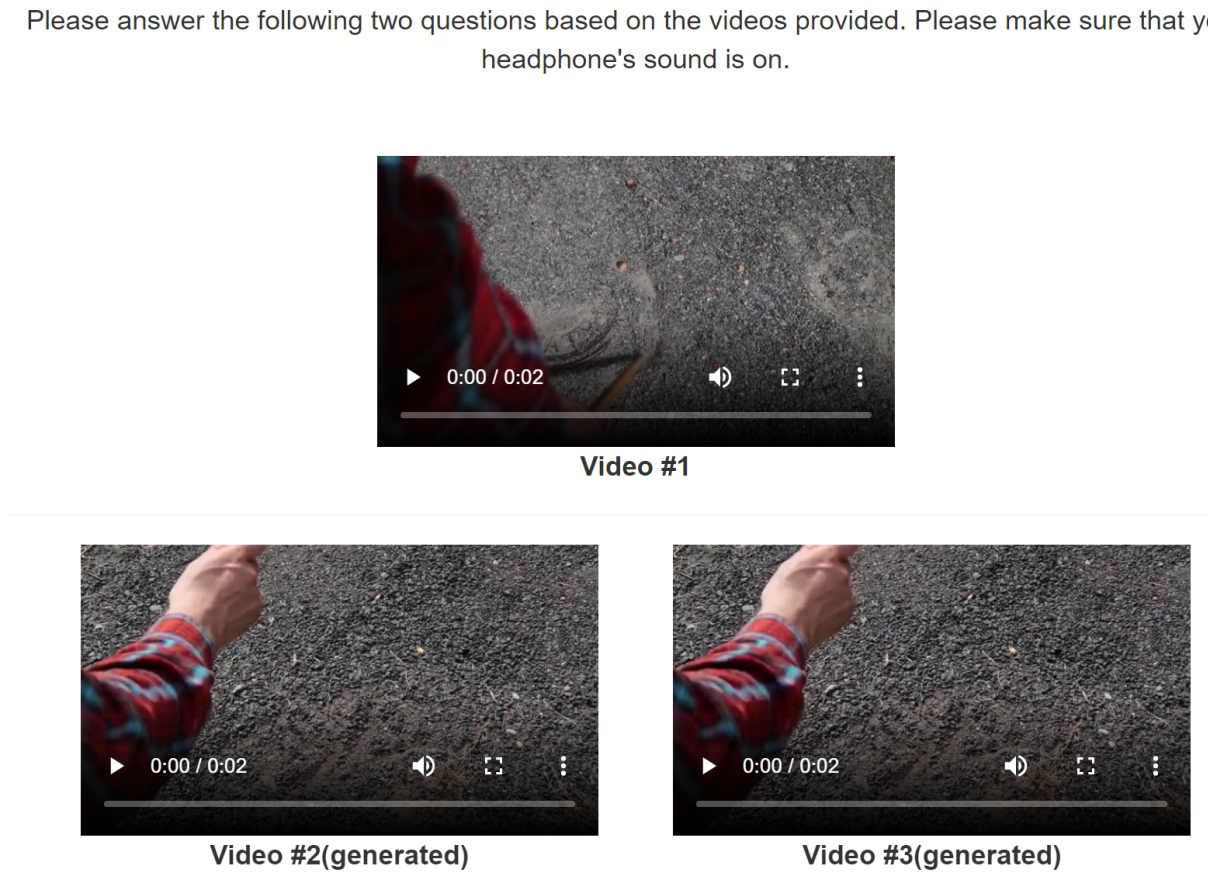

Figure 9. **Test page of the AMT test**. We present a screenshot of the main test page that the participants will see during the test. The participant need to answer both of the questions before moving to the next set of videos. Clicking on the "submit" button will navigate the participant to the next question.

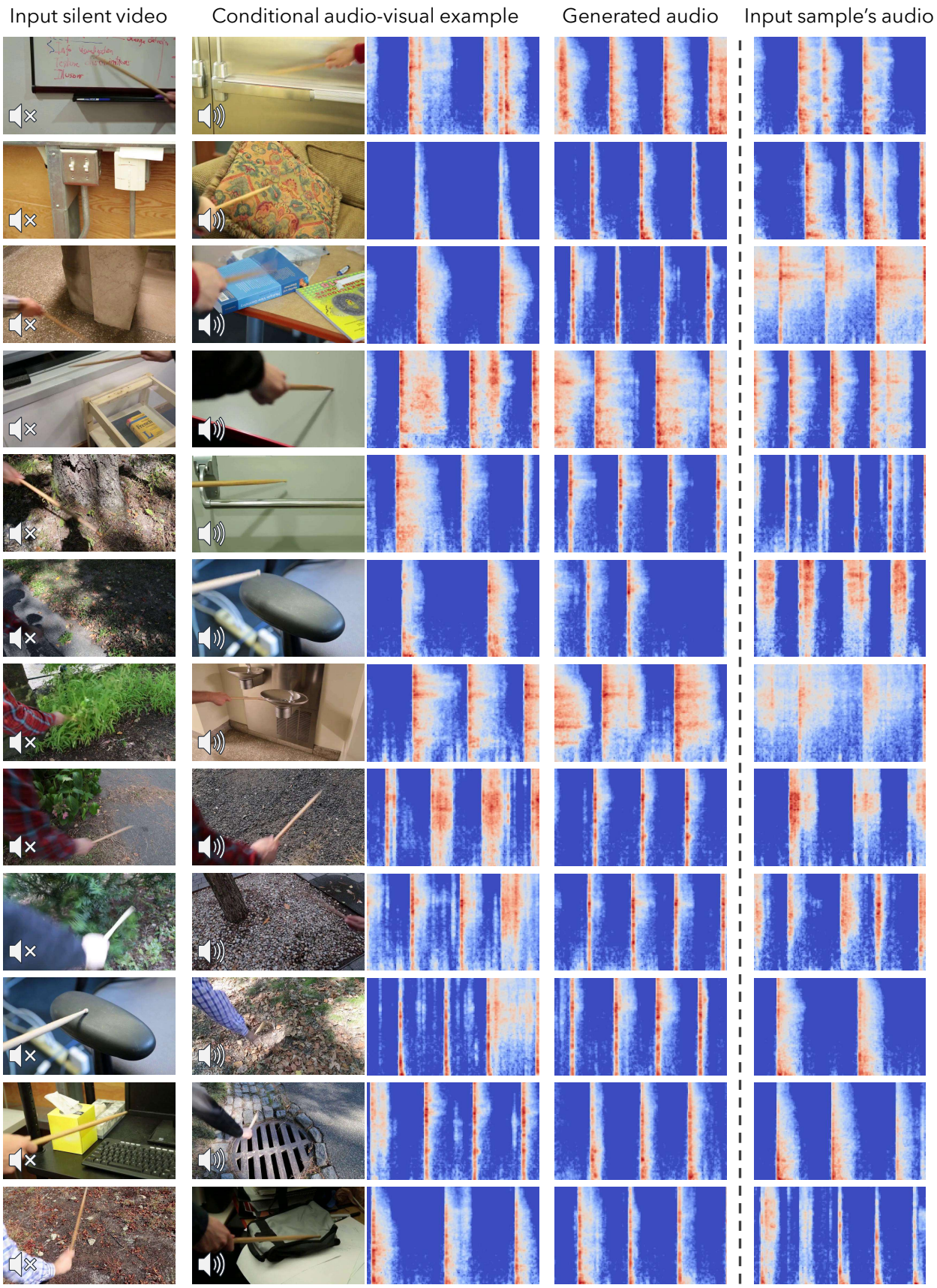

Figure 10. **Randomly selected results**. We present 12 results generated by our model. Please refer to the video to hear the results.

References

[1] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014. 1, 2

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2

[3] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. *arXiv preprint arXiv:2202.06875*, 2022. 2

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew 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. 7

[5] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. 2

[6] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *5th Annual Conference on Robot Learning*, 2021. 2

[7] Joon Son Chung, Bong-Jin Lee, and Icksang Han. Who said that?: Audio-visual speaker diarisation of real-world meetings. *arXiv preprint arXiv:1906.10042*, 2019. 2

[8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2, 4

[9] Chenye Cui, Yi Ren, Jinglin Liu, Rongjie Huang, and Zhou Zhao. Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement. *arXiv preprint arXiv:2211.10666*, 2022. 2

[10] Abe Davis and Maneesh Agrawala. Visual rhythm and beat. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2532–2535, 2018. 2

[11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *SIGGRAPH*, 2018. 2

[12] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017. 2

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2, 3, 4, 1

[14] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*, pages 758–775. Springer, 2020. 2

[15] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[16] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[17] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*, 2021. 2

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 5, 7, 8, 2

[19] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[20] Sanchita Ghose and John Jeffrey Prevost. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia*, 23:1895–1907, 2020. 2

[21] Sanchita Ghose and John J Prevost. Foleygan: Visually guided generative adversarial network-based synchronous sound generation in silent videos. *arXiv preprint arXiv:2107.09262*, 2021. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2

[23] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 4

[24] John Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001. 2

[25] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. *arXiv preprint arXiv:1609.09430*, 2016. 6, 2

[26] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. 6, 2

[27] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 1, 2, 8

[28] Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B Grosse. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620*, 2018. 2

[29] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021. 1, 2, 3, 4, 5, 7, 8

[30] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Sparse in space and time: Audio-visual synchronisation

with trainable selectors. *arXiv preprint arXiv:2210.07055*, 2022. 2, 4, 5, 7

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3, 1

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 3, 1

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. International Conference on Learning Representation, 2015. 1, 2

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[35] A Sophia Koepke, Olivia Wiles, Yael Moses, and Andrew Zisserman. Sight to sound: An end-to-end approach for visual piano transcription. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1838–1842. IEEE, 2020. 2

[36] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019. 4

[37] Timothy R Langlois and Doug L James. Inverse-foley animation: synchronizing rigid-body motions to sound. *ACM Transactions on Graphics (TOG)*, 33(4):41, 2014. 2

[38] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chan Young Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. *arXiv preprint arXiv:2112.00007*, 2021. 2

[39] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual associations. *arXiv*, 2022. 2

[40] Javier Nistal, Stefan Lattner, and Gael Richard. Drumgan: Synthesis of drum sounds with timbral feature conditioning using generative adversarial networks. *arXiv preprint arXiv:2008.12073*, 2020. 2

[41] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 2, 4

[42] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. *CVPR*, 2016. 1, 2, 5, 6, 7, 8

[43] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[44] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020. 2

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019. 4

[46] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. Mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, pages 367–372, 2014. 2

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 4

[48] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020. 1

[49] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[50] Kun Su, Xiulong Liu, and Eli Shlizerman. Multi-instrumentalist net: Unsupervised generation of music from body movements. *arXiv preprint arXiv:2012.03478*, 2020. 2

[51] Kun Su, Xiulong Liu, and Eli Shlizerman. How does it sound? *Advances in Neural Information Processing Systems*, 34, 2021. 2

[52] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. *CVPR*, 2022. 2

[53] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 4, 5, 2

[54] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 2

[55] Dmitry Ulyanov. Audio texture synthesis and style transfer. https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer, 2016. 2, 5, 7, 8

[56] Kees Van Den Doel, Paul G Kry, and Dinesh K Pai. Foleyautomatic: physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 537–544. ACM, 2001. 2

[57] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4, 1

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NIPS)*, 2017. 4

[60] Prateek Verma and Julius O Smith. Neural style transfer for audio spectograms. *arXiv preprint arXiv:1801.01589*, 2018. 2

[61] Yu Wang, Nicholas J Bryan, Justin Salamon, Mark Cartwright, and Juan Pablo Bello. Who calls the shots? rethinking few-shot learning for audio. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 36–40. IEEE, 2021. 2

[62] Yu Wang, Justin Salamon, Nicholas J Bryan, and Juan Pablo Bello. Few-shot sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 81–85. IEEE, 2020. 2

[63] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 2

[64] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2

[65] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 2

[66] Yunhua Zhang, Ling Shao, and Cees GM Snoek. Repetitive activity counting by sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14070–14079, 2021. 2, 3, 5, 6, 8, 1

[67] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2

[68] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3550–3558, 2018. 1, 2

[69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017. 2