This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Hearing Hands: Generating Sounds from Physical Interactions in 3D Scenes

Yiming Dou[1] Wonseok Oh[1] Yuqing Luo[1] Antonio Loquercio[2] Andrew Owens[1]
[1]University of Michigan [2]University of Pennsylvania



Figure 1. **What sound does this object make when you strike it with your hand?** We capture a 3D scene representation that can be used to simulate the sound that would result from a given hand motion. We reconstruct the scene Gaussian Splatting [20], then manipulate objects in the scene with hands, obtaining a sparse set of action-sound pairs. We use these examples to train a rectified flow model to map 3D hand trajectories at given position in a scene to a corresponding sound. At test time, a user can query an arbitrary 3D hand action and the model will estimate the resulting sound. Here we show several captured hand and audio pairs for two scenes (with representative video frames).

Abstract

We study the problem of making 3D scene reconstructions interactive by asking the following question: can we predict the sounds of human hands physically interacting with a scene? First, we record a video of a human manipulating objects within a 3D scene using their hands. We then use these action-sound pairs to train a rectified flow model to map 3D hand trajectories to their corresponding audio. At test time, a user can query the model for other actions, parameterized as sequences of hand poses, within the scene to estimate their corresponding sounds. In our experiments, we find that our generated sounds accurately convey material properties and actions, and that they are often indistinguishable to human observers from real sounds. Project page: https://www.yimingdou.com/hearing_hands/.

1. Introduction

Today's 3D reconstruction methods [20, 31] generally represent scenes as collections of static objects. While these representations are well-suited to many computer vision applications, they lack the ability to model *physical interactions*, such as what would happen if we struck an object with our hands. Modeling these interactions is a core challenge in a number of domains, ranging from AR/VR to robotics.

An emerging line of work aims to address this problem, particularly by modeling action-conditioned visual dynamics, resulting in reconstructions where one can open and close a microwave, operate scissors, or animate an object [7, 18, 22, 23, 44, 47]. While these approaches have been effective, they primarily focus on the visual and structural changes that objects undergo, and may not always be applicable to all objects, such as those that do not articulate or deform.

We focus instead on an aspect of interaction for 3D reconstruction that is complementary to these approaches: predicting the sound that an action would make if it were performed in a scene. Beyond making scenes more immersive and the interaction more realistic, studying the sounds of actions could provide a more complete understanding of the scene, beyond what's accessible from only its visual appearance [19, 35]. For instance, the sound we obtain from interacting with a surface can tell us whether it is hard or soft, smooth or rough, and hollow or dense. In addition, by predicting sound, one can implicitly model highly dynamic effects, such as vibrations or deformations of objects [6, 7, 50].

We aim specifically to create 3D reconstructions that enable us to predict what sounds a human hand will make when it interacts with the scene. We choose to parameterize our actions using hands, rather than alternatives such as drumstick [35] or hammer strikes [13], since they can execute a highly diverse range of actions by hitting, scratching, and manipulating objects. Hand sounds are also crucial for simulating interactions that a human might make in a virtual world application [33]. Finally, the actions that a hand makes can be parameterized using trajectories of 3D hand reconstructions, which can easily be captured using ordinary video cameras [14, 36, 41].

We take advantage of the link between a material's visual appearance and the sound that it generates when it is physically manipulated [10, 35, 50]. In contrast to vision-to-sound work, however, we are interested in generating the sound of user-specified *simulated* interactions, without the need for an input video (Fig. 1). To do this, we collect a dataset of 3D hand-scene interactions paired with sounds. We first record a video where a person interacts with objects using their hands. We then estimate hand pose and register it to the same space as a 3D scene reconstruction, obtained using Gaussian Splatting [20] (Fig. 1). This allows us to remove body occlusions from the training data (Fig. 4) and to obtain 3D-consistent data augmentation by generating different views of the same interaction. We use this data to train a model based on rectified flow [27, 45] that, from a sequence of 3D hand poses and visual content from the scene, can generate the sound resulting from the hand's motion (Fig. 1).

To help study this problem, we collect a dataset containing 24 indoor and outdoor 3D scenes and 9.1 hours of physical interactions. Through our experiments on this dataset, we find that our model successfully generates sounds that convey hand motion, such as by capturing the timing of contact. These experiments also suggest that the generated sounds convey material properties of objects in the scene.

2. Related Work

Multimodal 3D scene reconstruction. A variety of recent works augment 3D reconstructions with other modalities. LERF [21] distills CLIP [38] features into a NeRF [31], which can be used in downstream tasks such as 3D visual grounding [49] and task-oriented grasping [39]. ObjectFolder [11–13] constructs multimodal representations for objects. However, they only consider small object-level reconstructions of rigid objects that can be captured with a special apparatus (e.g., a turntable) and are limited to simple impact sound. In contrast, our goal is to produce scene-level reconstructions and to support complex actions represented by hand motions. Tactile-augmented radiance fields [8] register sparse tactile signals into the 3D space, allowing one to query how a given 3D location would feel if touched. We consider sound instead of touch, and crucially we do not treat sound as an intrinsic property of a surface (like they do with touch). Instead, it is a function of the action that is applied to the scene, which is specified via a 3D hand trajectory.

Material properties in 3D scene reconstruction. Another line of works focuses on integrating dynamics into 3D scene representations. Early work [7] used modal models to simulate deformation. D-NeRF [37] augments a NeRF with a displacement field, which adds temporal information to the NeRF. Recently, PhysGaussian [47] uses explicit 3D Gaussian Splatting [20] to model the dynamic behaviors, and VR-GS [18] further develops a dynamics-aware interactive Gaussian Splatting representation. Like these works, we model how a scene will react to a physical interaction. However, we focus on hand-based actions and predict sound rather than visual deformation. Sound prediction provides a complementary way to analyze physical properties, especially in cases where visual deformation is not available (such as for hard surfaces).

Video-to-audio generation. There have been many approaches for synthesizing audio from visual or language inputs. Early work predicted simple speech from vision [32]. Our approach is closely related to work that generates sound as a way to understand material properties [10, 35, 50]. Early work in this area predicted sound from videos of a drumstick striking objects [35]. In contrast, our input is a 3D trajectory of a hand, allowing us to query the model with user-specified actions at test time (without need for a video input), we trained with many samples within a single scene, and we use 3D constraints, such as to obtain a clear view of the action and materials. Later work used more powerful generative models for conditional audio generation, such as autoregressive models [51], GANs [4], and VQ-GANs [17]. Recent work uses diffusion models. Diff-Foley [28] represents the video using a joint audio-visual embedding [1, 34] from the video and generates a sound using latent diffusion.

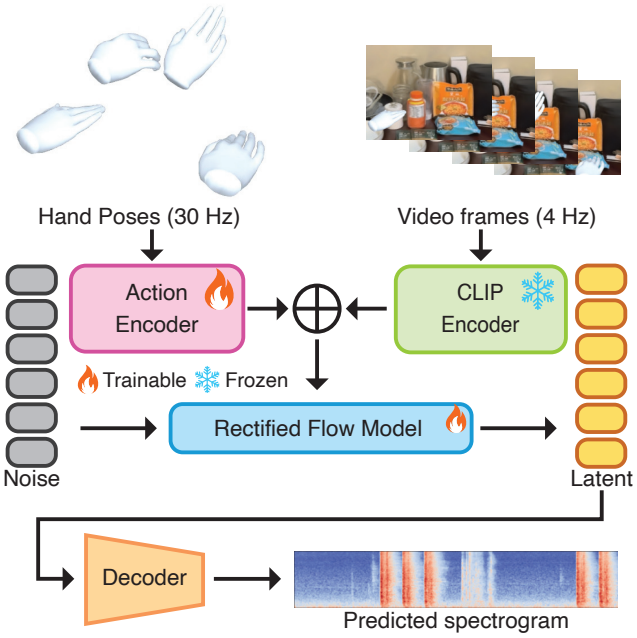


Figure 2. **Sound generation**. We train a rectified flow model [45] to generate a sound spectrogram from a sequence of 3D hand positions and video frames generated from a 3D reconstruction of a scene. The sound can subsequently be converted into a waveform using a vocoder.

Frieren [45] uses rectified flow matching [27] for better generation quality and efficiency. Our audio generation module is based on the Frieren's rectified flow matching, but we use conditional information from a sequence of 3D hand poses and visual content extracted from a Gaussian splatting representation, instead of predicting sound from a video.

3D audio reconstruction. A recent line of work has generated sound from 3D body pose [15, 48]. In contrast, we model the combination of the action and the real-world objects that it is physically interacting with, rather than the body itself. Work on acoustic reconstruction [2, 5, 9, 26, 29, 42] models how a sound propagates through a 3D scene, given the position of a sound and a listener. This line of work is complementary to ours: we model the generated sound in a scene, rather than the interaction between the listener and the sound.

3. Method

We aim to obtain a multimodal 3D reconstruction of a scene that allows us to predict the sound of actions. To do so, we combine a visual neural field $F_\theta : (\mathbf{x}, \mathbf{r}) \mapsto (\mathbf{c}, \mathbf{d})$ that maps a 3D point $\mathbf{x}$ and viewing direction $\mathbf{r}$ to its corresponding RGB color $\mathbf{c}$ and depth $\mathbf{d}$ with an action-conditioned audio estimator $F_\phi : (\mathbf{v}, \mathbf{a}) \mapsto \mathbf{s}$, which generates sound $\mathbf{s}$ given the video $\mathbf{v}$ and the action $\mathbf{a}$. This action specifies the trajectory of a hand that physically interacts with the scene. We focus on human hands since they are capable of many motions (e.g., tapping, scratching, patting); they are crucial within virtual world applications; and can be easily captured in 3D without special equipment.

In the rest of this section, we explain how to generate a large and diverse dataset to train F_ϕ (Sec.3.1). Then, we explain the functional form that we use to instantiate F_ϕ (Sec. 3.2).

3.1. Dataset

Training a generalizable F_ϕ requires a diverse dataset of synchronized interaction videos $\mathbf{v}$, actions $\mathbf{a}$, and resulting sound $\mathbf{s}$. We collect this dataset in 24 different scenes, including bedrooms, lobbies, trees, snow, and musical instruments (see Fig. 4 for some dataset samples). For each scene, we first generate a 3D reconstruction F_θ using Gaussian Splatting [20]. Specifically, a human collector scans the scene by recording multiple views, whose poses are estimated using the structure of motion [40].

After scanning, we collect videos of humans interacting with different regions of the scene (Fig. 3). During such interactions, the data collector performs a variety of actions with their hands, *e.g.*, squeezing, hitting, or scratching, on some of the objects present in the scene, *e.g.*, tables, plastic bags, or trees. We use this procedure to generate a set of videos with various impact sounds. Note that during each interaction, we keep the camera location fixed by mounting the recording device to a tripod.

We use HaMeR [36] for 3D hand detection in such interaction videos. Specifically, we define the sequence of N 3D hand keypoints for both hands as $\mathbf{a} \in \mathbb{R}^{2N \times 21 \times 3}$. If one hand is not visible, we pad its detections with zeros. We register the camera on the tripod c to F_θ with COLMAP [40], obtaining its global position $T_c^{F_\theta}$. Then, we use $\mathbf{a}$ and F_θ to generate a simulated interaction video $\mathbf{v}$. Specifically, we project the sequence of 3D hands $\mathbf{a}$ on an global RGB view of F_θ at the camera position $T_c^{F_\theta}$ (Fig. 3). We also re-center the camera position to each hand in $\mathbf{a}$ to obtain a sequence of local RGB views, which contains the local details of the regions being interacted with. The simulated video $\mathbf{v}$ represents the combination of both global RGB views $\mathbf{v}_g$ with hands and local RGB views $\mathbf{v}_l$. We label each $\mathbf{v}$ with the sound $\mathbf{s}$ from the original video of the human interacting with the scene.

We collect approximately 1,400 seconds of videos in each scene, with a frame rate of 30Hz. We pre-process these videos to generate $\mathbf{a}$, $\mathbf{v}$, and $\mathbf{s}$ as explained above. This pre-processing results in a dataset of approximately 9.1 hours of simulated interactions. We additionally use the relative position of the camera to the scene $T_c^{F_\theta}$ to project $\mathbf{a}$ from the local camera frame to the global frame of F_θ. This allows us to synthesize two novel views of the simulated interactions from slightly different viewpoints, *i.e.*, top view, side view.

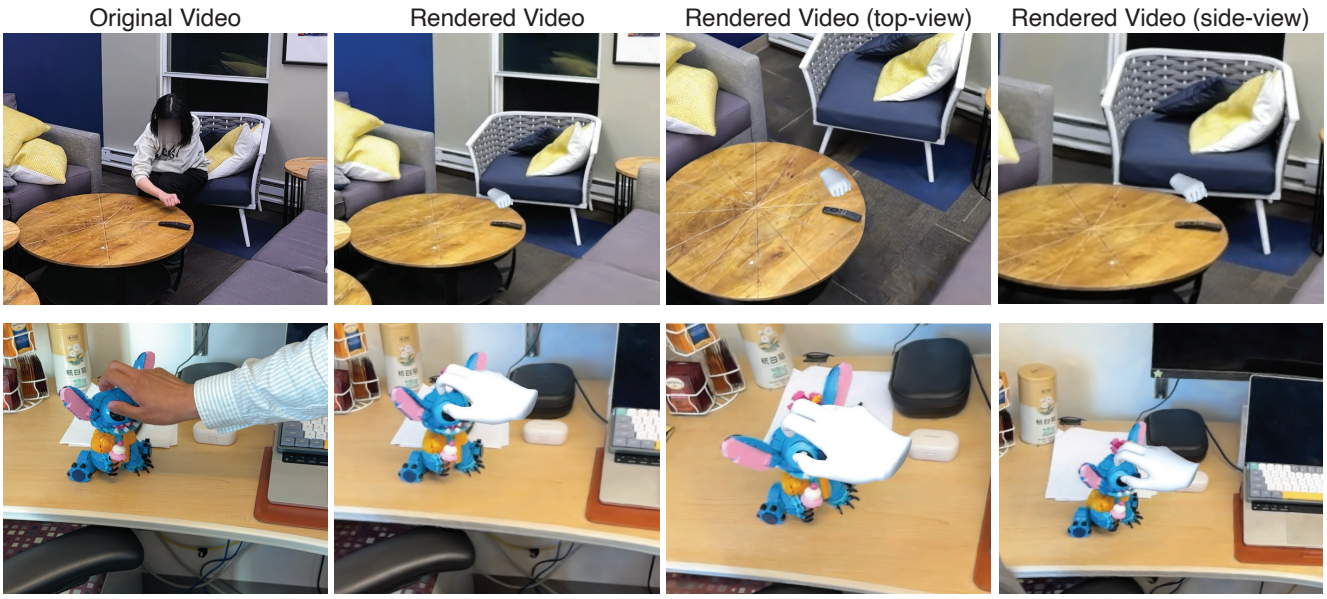


Figure 3. **Data capturing pipeline.** In the original video, a human collector interacts with the scene by performing various actions with their hands. We lift the annotator's hands to the same 3D space of the scene reconstruction. We render a video of the interaction by projecting 3D hands on multiple viewpoints of the scene. All rendered videos are synchronized with the sounds made by the hand actions.

Fig. 3 shows some representative samples for this process. To the best of our knowledge, this is the first dataset to capture human actions along with their sounds that are spatially aligned in 3D scenes.

3.2. Generating action-conditioned sound

We represent F_ϕ as a generative model $p_\phi(\mathbf{s} \mid \mathbf{v}, \mathbf{a})$ where $\mathbf{s}$ is the sound generated by $\mathbf{a}$ in the video $\mathbf{v}$. Similarly to previous work, we represent $\mathbf{s}$ as a mel-spectrogram, transforming audio synthesis into image generation.

We instantiate $p_\phi(\mathbf{s} \mid \mathbf{v}, \mathbf{a})$ as a rectified-flow matching generative model [27]. Our model is built upon the video-to-sound Frieren model [45]. Similarly to Frieren, we compress $\mathbf{s}$ to a latent vector with a pre-trained autoencoder, and train a generative model in latent space. However, we empirically found the Frieren model to fail to generate high-quality sound from our videos, even when fine-tuned on our dataset. This is because our videos contain simulated interactions, which lack the low-level details and consistency of real videos, e.g., the motion and deformation of objects. Therefore, we introduce two key modifications to Frieren: (i) we encode $\mathbf{v}$ with CLIP [38] instead of CAVP [28] since we found CLIP to have better spatial consistency and material understanding; and (ii) we explicitly condition the model on 3D action $\mathbf{a}$, which forces the model to focus on the low-level details of the hand motion. We empirically found these two modifications to be crucial for performance, as we demonstrate in the experimental section. A visualization of the schematics of our model can be found in Fig. 2. Further implementation details can be found in Sec. 4.

We train F_ϕ from scratch on our dataset. After training, we can generate the sound of previously unseen interactions $\hat{\mathbf{a}}$ in the scene F_θ by first selecting a camera viewpoint $T_c^{F_\theta}$ and then rendering a video of the interaction $\hat{\mathbf{v}}$. We then use our model to predict the interaction's sound $\hat{\mathbf{s}}$ by passing $\hat{\mathbf{a}}$ and $\hat{\mathbf{v}}$ to our generative model. We use the ability to generate sound for new actions in the scene to design an interactive interface for F_θ (Sec. 5.2).

4. Implementation Details

We reconstruct the 3D scene using the Splatfacto method from Nerfstudio [43]. Approximately 1K images taken from various views are used for each scene. The gaussians are randomly initialized with scale regularization [47]. During training, we optimize the reconstruction with the Adam [24] optimizer for 20,000 steps on a single NVIDIA RTX 2080 Ti GPU.

4.1. Audio generation model training and inference

Our implementation of F_ϕ is based on Frieren [45] but differs on the conditioning module to better suit our task. First, we use CLIP features instead of CAVP features for encoding the simulated interaction video $\mathbf{v}$. Specifically, we pass the global video $\mathbf{v}_g$ and local video $\mathbf{v}_l$ separately into the CLIP model and obtain two features, which are then concatenated into the input feature of our model. Note that, similarly to Frieren, we condition the model on the frames

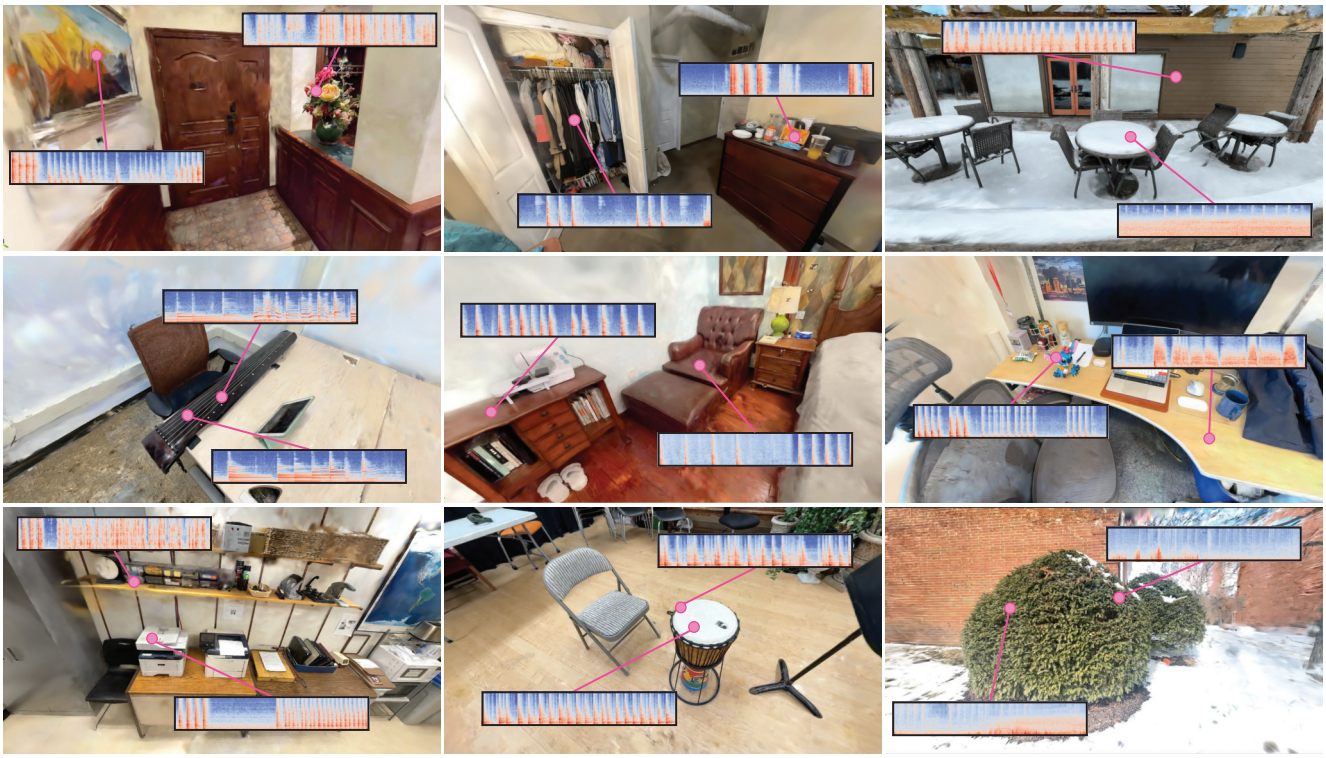


Figure 4. **Representative examples from the dataset.** Our dataset is collected in 24 scenes, including offices, outdoor trees, bedrooms, *etc*. We show six such scenes in the figure above, with examples of action-generated sounds. Our dataset covers a wide range of actions (hitting, scratching, patting, *etc*.) and interacted materials (wood, metal, plastic. *etc*.). In each scene, approximately 1,400 seconds of videos are collected, resulting in a total of 9.1 hours of interaction data.

from the video down-sampled at 4Hz. We also find that the visual features extracted from downsampled videos are insufficient to capture fine-grained hand motions present in our data. Therefore, we additionally condition the model on the action $\mathbf{a}$, which includes the trajectory of 3D hand poses. Being sampled at 30Hz, $\mathbf{a}$ gives the model a higher resolution view of the action. We encode $\mathbf{a}$ to the same dimension of the frame embeddings via a linear layer, and normalize it to a unit vector. Finally, we upsample the frames and actions embeddings to the same temporal frequency of the sound spectrogram, *i.e*., 31.25 Hz, using nearest neighbor upsampling. We then obtain the final conditioning vector by summing the two embeddings elementwise. This conditioning vector is concatenated to the input noise and passed to the vector field estimator to generate the latent spectrogram representation of the sound.

Following previous works [28, 45], we divide our dataset into non-overlapping chunks of eight seconds duration. The video's audio is downsampled to 16kHz and transformed into mel-spectrograms with 80 bins and a hop size of 256. We use 10% of the collected videos as the test set, 10% as validation, and the remaining as the training set. We use the knowledge of each video's camera pose $T_c^{F_\theta}$ to ensure that none of the camera views in the test set overlap with the ones in the training and validation set.

We then train the model for 40 epochs with a batch size of 128 using the Adam [24] optimizer. We initialize the learning rate to 10^{-5}, do a warmup to 4×10^{-4} over 1000 steps, and finally linearly decrease it to 3.4×10^{-4} over 22 epochs. We train on a single NVIDIA L40s.

At inference time, the model performs 26 sampling steps with a 4.5 guidance scale. The generated latent is then decoded into a mel-spectrogram with a pre-trained decoder [45]. Finally, a pretrained vocoder [25] is used to transform the spectrogram into a waveform.

5. Experiments

We design our experiments to answer the following questions: (1) Can F_ϕ generate synthetic sounds that are almost indistinguishable from real ones? (2) How important is conditioning on $\mathbf{v}$ and $\mathbf{a}$? (3) Do the predicted sounds convey physical properties of the scene, *e.g*., its material and their position relative to the camera? We answer these questions with qualitative and quantitative experiments.

Table 1. **Ablation study.** Since CLIP features and hand poses respectively provide material information and precise sound synchronization, removing either of them from conditioning will result in a significant drop in the overall performance. In particular, removing CLIP features and hand poses results in the greatest drop in the CLAP *material* accuracy and *action* accuracy, respectively. Excluding synthetic-view data augmentation affects the performance generally.

Model variation	STFT ↓	Envelope ↓	FID ↓	IS ↑	CDPAM ($\times 10^{-4}$) ↓	CLAP-acc (%) ↑			Labeled *real* (%) ↑
						all	*action*	*material*	
RegNet	0.62	0.77	63.84	5.73	3.38	1.08	42.55	3.52	-
Frieren	0.74	0.81	56.66	16.76	3.71	23.94	41.73	42.55	43.79 ± 2.64
Ours	**0.50**	**0.66**	59.02	**17.82**	**3.32**	**28.09**	**50.50**	**45.62**	47.18 ± 2.66
- w/o CLIP	0.68	0.77	**58.07**	17.10	3.86	18.25	43.90	31.80	41.24 ± 2.62
- w/o hand pose	0.69	0.77	58.92	16.76	3.77	20.96	38.21	39.11	43.50 ± 2.64
- w/o synthetic-view	0.62	0.73	58.99	17.42	3.66	24.12	47.61	40.56	43.22 ± 2.64

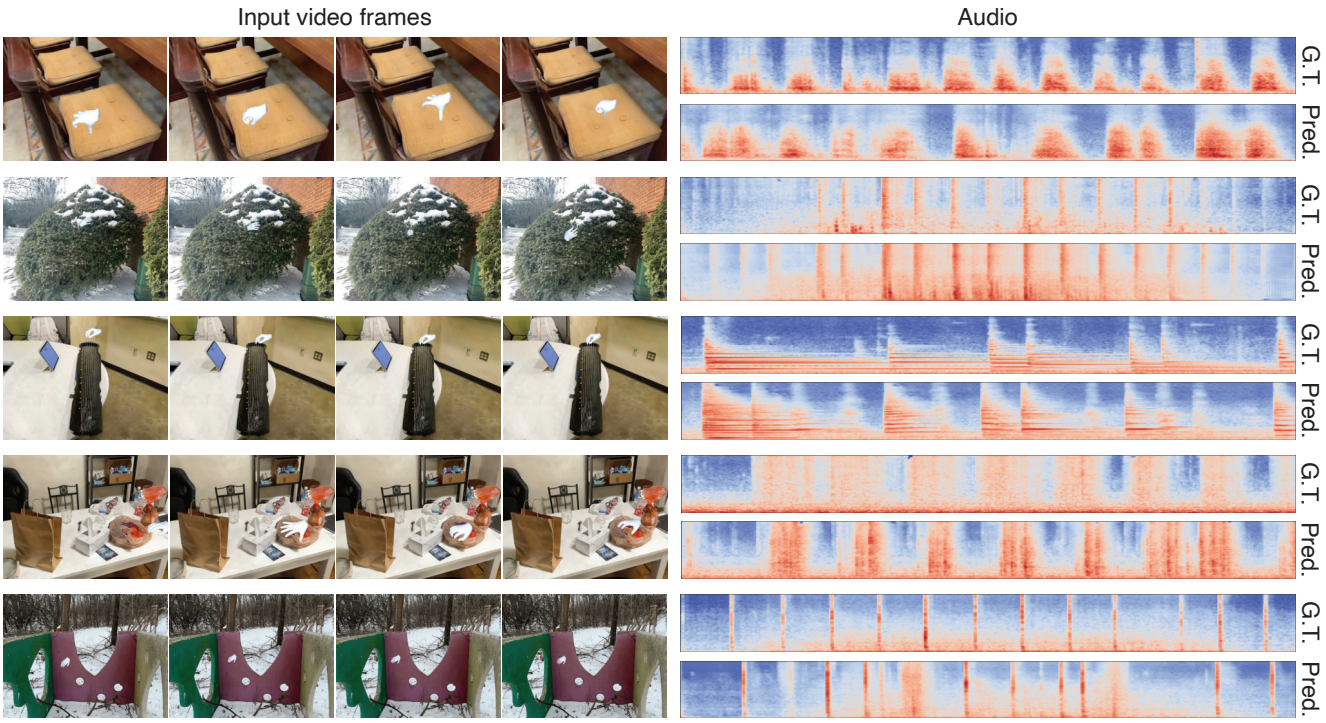


Figure 5. **Qualitative results.** We show the generation results on five interactions. Generally, the predictions match the ground-truth in both motion synchronization and material properties. Note that when the hand is less visible in the video or the motion is ambiguous (*e.g.*, the last row), our model will generate less-synchronized audio spectrograms.

5.1. Experimental Setup

We use the following metrics to evaluate the quality of the sounds generated by F_ϕ and compare it to a set of baselines.

Raw Audio Similarity. As custom in previous work [13], we measure the L2 distance between ground-truth and predicted audio signals in both the spectrogram (*STFT*) and waveform (*Envelope*) space. This metric primarily assesses the model's capability to capture low-level sound features.

Latent Space Similarity. We encode both ground-truth and generated sounds to a latent representation and measure their distance in this space. Specifically, we adopt the CDPAM [30] metric to measure distances in the latent space, which uses a pretrained model to quantify perceptual audio similarity. Additionally, following previous work [45], we compute the Frechet Inception Distance (FID) and Inception Score (IS) using the pretrained mel-ception encoder model from SpecVQGAN [16].

CLAP accuracy. To assess the model's effectiveness in generating sounds that accurately represent the actions and material properties in a scene, we introduce a new metric: *CLAP accuracy*. This metric evaluates whether an off-the-

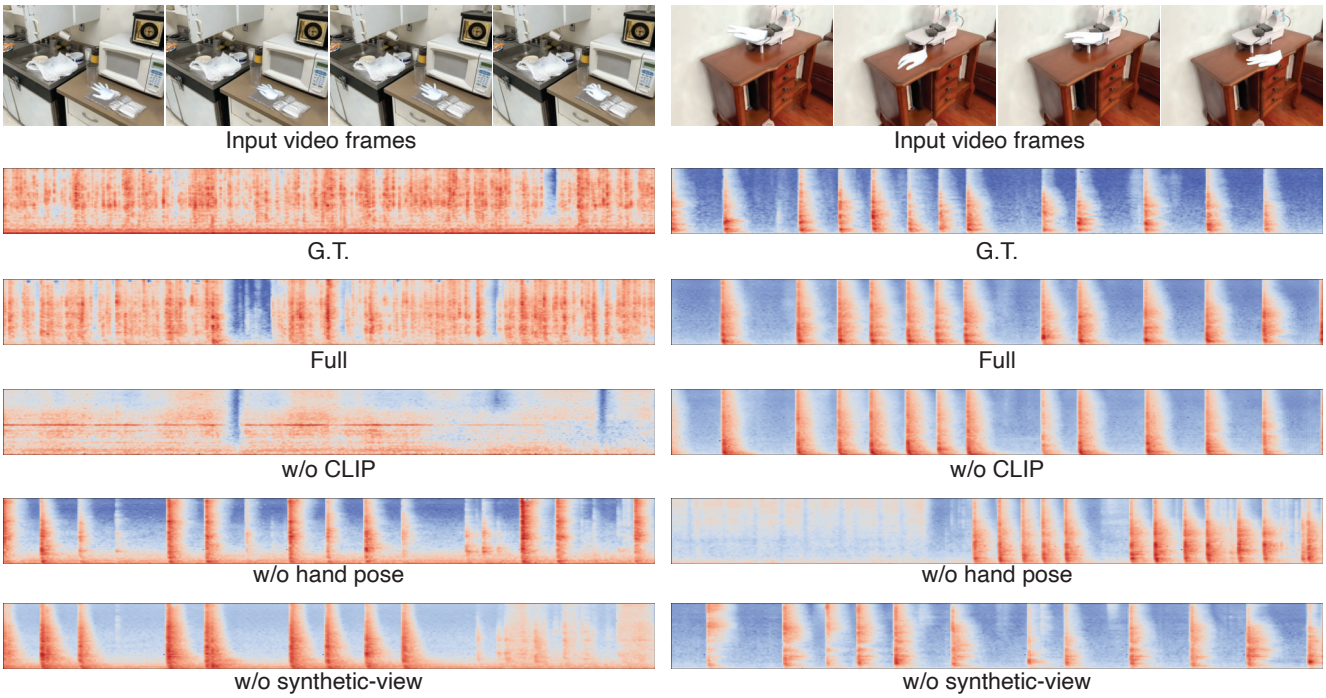


Figure 6. **Ablation study results.** We show the spectrogram predictions from our full model and three ablations. We notice that removing CLIP features softens impact sounds while removing hand pose features results in poor audio-video synchronization. Similar to quantitative results, the model trained without synthetic-view augmentation performs worse in both aspects.

shelf CLAP model [46] assigns the same zero-shot label to both the ground truth and synthetic sounds. Specifically, we define an action set $\mathbb{A}$ comprising 7 hand actions (e.g., knocking, scratching) and a material set $\mathbb{M}$ with 13 materials (e.g., wood, plastic). From these, we generate a set $\mathbb{P}$ of 91 action-material pairs by taking the Cartesian product of $\mathbb{A}$ and $\mathbb{M}$. For each pair in $\mathbb{P}$, we format the CLAP model's text prompt as: "This is a sound of hand {action} {material}," with {action} and {material} drawn from the pairs in $\mathbb{P}$. We then record the number of instances where the ground truth and generated sounds are assigned the same label (*CLAP-acc, All*). For a more fine-grained analysis, we additionally report the frequency of action label matches (*CLAP-acc, Action*) and material label matches (*CLAP-acc, Material*). This metric is inspired by prior work in sound generation [35], which similarly uses linear models to classify materials.

Real-or-fake study. We conduct a real-or-fake user study to evaluate whether participants can distinguish between generated and real sounds. Fifty-nine participants participated in this study. Each participant viewed 32 pairs of 8-second interaction videos $\mathbf{v}$ with each pair comprising one video with ground-truth sound and one with generated sound. These pairs were sampled from a set of 1107 video pairs, with sounds generated either by our full model or one of its ablations, selected with equal probability. Following prior work [35], we use a two-alternative forced-choice (2AFC) test, where participants select the video they believe has the most realistic sound in each pair. All videos in the study are from the test set.

5.2. Results

We begin by analyzing the differences in generated sounds produced by our full model and its ablations using quantitative distance metrics. The evaluation results, shown in Table 1, indicate that while all features of our model contribute to the generation quality, some are more essential than others. Notably, removing conditioning on either the CLIP embeddings of the video or the 3D hand poses leads to a significant drop in performance. In contrast, excluding multi-view data augmentation during training has the smaller impact, resulting in relatively minor changes in both raw audio and latent distance metrics. For metrics based on a pretrained melception model (FID and IS), all methods perform similarly. We hypothesize that this is due to our data differing significantly from VGGSound [3], the dataset on which the melception model was originally trained.

Interestingly, we observe that removing CLIP features results in the greatest drop in CLAP *material* accuracy, while removing hand pose features most affects *action* accuracy. This aligns with expectations: CLIP features primarily provide material information about the scene, while

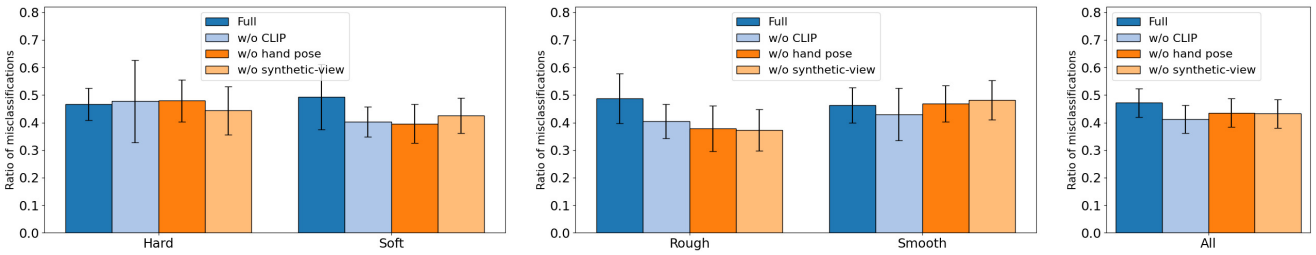


Figure 7. **Results of real-or-fake study**. We show the ratio of humans being fooled by different variants of our model. We break down our results into three categories: softness, smoothness, and average over all samples. The error bars show 95% confidence intervals. We find that our full model achieves a misclassification rate of approximately 47%, indicating the high quality of the generated sounds. In addition, our model generally outperforms baselines without visual or action information.

hand pose features are essential for encoding actions.

In Fig. 5, we show qualitative results of our full model for five interactions. Visual inspection reveals that our model generates sounds that generally align with the ground-truth in both synchronization and material properties. We further show the qualitative results of ablation study in Fig. 6. We find that removing hand pose features disrupts audio-video synchronization, as visual information alone is insufficient for accurately estimating precise hand motions. Removing CLIP features, on the other hand, makes the model unable to synthesize the sound with correct material properties. Removing synthetic view results in general performance drop in both aspects.

Real-or-fake study. We measure how often participants mistake our generated sounds for ground-truth sound. We collect the data for this study on Amazon Mechanical Turk, obtaining answers from 59 participants.

Ideally, if the two sounds are completely indistinguishable from each other, we would observe a misclassification rate of 50%, which indicates that participants pick at random.[1] The results of this analysis, averaged over all videos and participants, are shown in Tab. 1. We find that our full system generates high-quality sounds with a misclassification rate of approximately 47%.

We present results broken down by the material properties of the objects the hand interacts with in Fig. 7. Consistent with our quantitative findings, our approach outperforms all baselines on average. The improvement is especially notable for rough surfaces and soft materials, while differences are less pronounced for other categories.

Our study also suggested qualitatively interesting patterns in how users distinguish real from fake sounds. Notably, background noise in real recordings may sometimes be perceived as being artificial, whereas our model's clearer outputs are often judged as more realistic. Users also sometimes may have been unfamiliar with the typical sounds of certain materials — particularly those rarely encountered, such as snow - - which can lead to inconsistent judgments. Additionally, inaccuracies in hand tracking and irrelevant movements during data collection can make it unclear whether the hand is interacting with the object or simply moving through space. This ambiguity might be mitigated by modeling object deformations resulting from contact.

[1]However, this is not an upper bound on performance, since subjects may sometimes prefer unrealistic sounds.

6. Conclusion

We see our work as being a step toward creating realistic and immersive 3D scene reconstructions, with potential applications in robotics and AR/VR. We do so by predicting the sound of hands interacting with a scene. Both automated evaluations and real-or-fake evaluations that our synthetic sounds outperform baselines and are often indistinguishable from real sounds. They also may convey material properties and subtle actions.

Limitations. One key limitation of our approach is that assumes that the objects in the scene do not move or deform when manipulated. In practice, this assumption is often violated, especially when manipulating small objects. Another limitation comes from the errors of the 3D hand detection model, which might result in inaccurate hand motions in our dataset. This can be improved with future hand detection models.

Acknowledgements. We thank Jeongsoo Park, Ayush Shrivastava, Daniel Geng, Ziyang Chen, Zihao Wei, Zixuan Pan, Chao Feng, Xuanchen Lu, Ang Cao and the reviewers for the valuable discussion and feedback. We appreciate the generous help in data collection of Yueqi Ren, Qifan Wu, Shuangdi Zhang, Xingxian Li and Yuchen Huang from Qingyun Chinese Music Ensemble at UM. We thank all the friends who participated in the real-or-fake study. This work was supported by an NSF CAREER Award #2339071, a Sony Research Award, and the DARPA TIAMAT program.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 2

[2] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[3] 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

[4] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29:8292–8302, 2020. 2

[5] Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21886–21896, 2024. 3

[6] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Fredo Durand, and William T Freeman. The visual microphone: Passive recovery of sound from video. 2014. 2

[7] Abe Davis, Justin G Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM Transactions on Graphics (TOG)*, 34(6):1–7, 2015. 1, 2

[8] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26529–26539, 2024. 2

[9] Yilun Du, Katie Collins, Josh Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. *Advances in Neural Information Processing Systems*, 2021. 3

[10] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436, 2023. 2

[11] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. *arXiv preprint arXiv:2109.07991*, 2021. 2

[12] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022.

[13] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286, 2023. 2, 6

[14] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2

[15] Chao Huan, Dejan Markovic, Chenliang Xu, and Alexander Richard. Modeling and driving human body soundfields through acoustic primitives. *arXiv preprint arXiv:2407.13083*, 2024. 3

[16] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021. 6

[17] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021. 2

[18] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, and Chenfanfu Jiang. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. *arXiv preprint arXiv:2401.16663*, 2024. 1, 2

[19] Mark Kac. Can one hear the shape of a drum? *The american mathematical monthly*, 73(4P2):1–23, 1966. 2

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2, 3

[21] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2

[22] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. In *8th Annual Conference on Robot Learning*, 2024. 1

[23] Chung Min* Kim, Mingxuan* Wu, Justin* Kerr, Matthew Tancik, Ken Goldberg, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[24] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5

[25] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022. 5

[26] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *arXiv preprint arXiv:2302.02088*, 2023. 3

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2, 3, 4

[28] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4, 5

[29] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *Advances in Neural Information Processing Systems*, 2022. 3

[30] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *ICASSP 2021, To Appear*, 2021. 6

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[32] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, 2011. 2

[33] Rolf Nordahl, Luca Turchet, and Stefania Serafin. Sound synthesis and evaluation of interactive footsteps and environmental sounds rendering for virtual reality applications. *IEEE transactions on visualization and computer graphics*, 2011. 2

[34] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 2

[35] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 2, 7

[36] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2, 3

[37] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[38] 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*, pages 8748–8763. PMLR, 2021. 2, 4

[39] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023. 2

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[41] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[42] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 2022. 3

[43] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 4

[44] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024. 1

[45] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*, 2024. 2, 3, 4, 5, 6

[46] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023. 7

[47] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 1, 2, 4

[48] Xudong Xu, Dejan Markovic, Jacob Sandakly, Todd Keebler, Steven Krenn, and Alexander Richard. Sounding bodies: modeling 3d spatial sound of humans using body pose and audio. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[49] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024. 2

[50] Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Josh Tenenbaum, and Bill Freeman. Shape and material from sound. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[51] 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. 2