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

Anonymous Author(s)

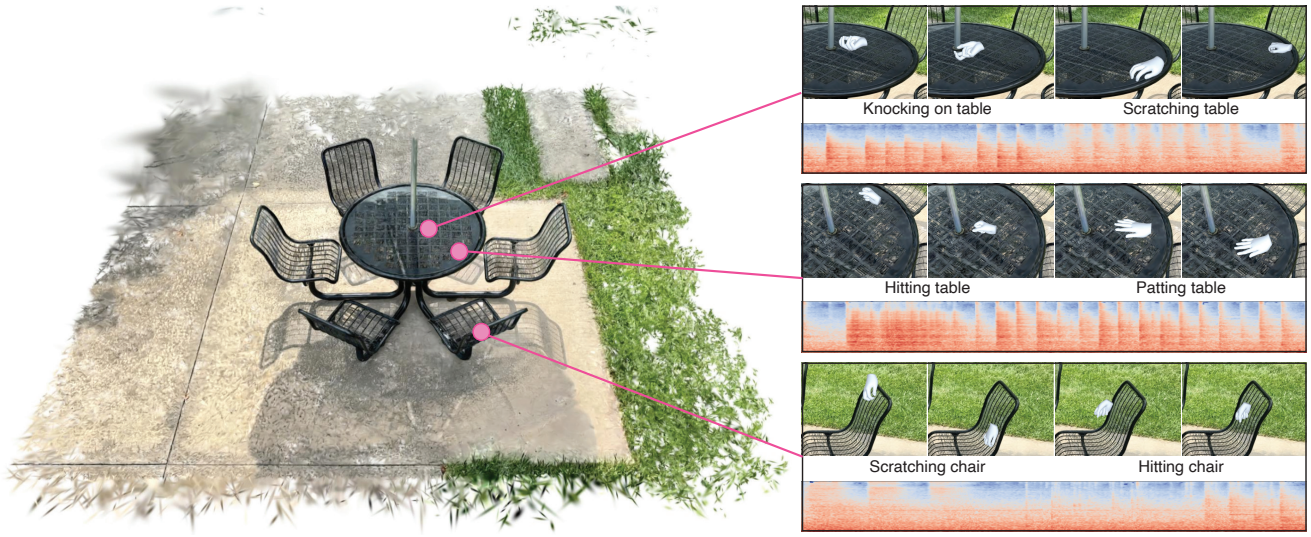

Figure 1. **What sound does this object make when you strike it with your hand?** We capture 3D scene reconstructions that can be used to simulate the sound that would result from a given hand motion. A human captures a 3D scene using Gaussian Splatting [20], then manipulates objects in the scene with their 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 with an arbitrary 3D hand action and the model will estimate the resulting sound. Here we show several captured hand and audio pairs (with representative video frames). **Please refer to our supplementary material for video and audio results.**

Abstract

We study the problem of making 3D scene reconstructions interactive by asking the following question: can we predict the sounds of human hands interacting with a 3D reconstruction of scene? We focus on human hands since they are versatile in their actions (e.g., tapping, scratching, patting) and very important to simulate human-like avatars in virtual reality applications. To predict the sound of hands, we train a video and hand-conditioned rectified flow model on a novel dataset of 3D-aligned hand-scene interactions with synchronized audio. Evaluation through psychophysical studies shows that our generated sounds are frequently indistinguishable from real sounds, outperforming baselines lacking hand pose or visual scene information. Through quantitative evaluations, we show that the generated sounds accurately convey material properties and actions. We will release our code and dataset to support further development in interactive 3D reconstructions.

1. Introduction

Today's 3D reconstruction systems [20, 31, 36, 38] produce models that, while visually impressive, largely represent scenes as rigid collections of surfaces and volumes. Missing from these representations is an ability to convey *physical interaction*, such as what would happen if we struck an object with our hands. Modeling the results of actions like these is a core challenge in a variety of applications, ranging from AR/VR to robotics.

An emerging line of work has captured different aspects of physical interaction, particularly by adding action-conditioned dynamics of the scene's objects and modeling physics, resulting in models that can open and close a microwave, operate scissors, or animate an object [18, 22, 23, 45, 48]. 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, 51].

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 hits [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 easily be modeled using trajectories of 3D hand models, which can easily be captured using ordinary video cameras, without special equipment [14, 37, 42].

Accurately simulating the sounds of hand motions can be a challenging task. In principle, one could exhaustively collect the sounds of actions for each scene and directly include sound encoding into the 3D reconstruction. To avoid such a tedious and time-consuming procedure, we propose leveraging the correlation between a material's visual appearance and the sound it generates upon interaction [10, 35, 51]. In contrast to vision-to-sound work, however, we are interested in generating the sound of user-specified *simulated* interactions (Fig. 1), whose visual appearance might not be sufficient for off-the-shelf video-to-sound models [4, 17, 28, 46, 52]. Therefore, we collect a new annotated dataset of real-world interaction videos paired with ground-truth sounds. From the videos, we produce "simulated" interactions by lifting hand poses to the same 3D space of a Gaussian Splatting reconstruction [20] of the scene (Fig. 1). This allows, by design, to remove body occlusions from the training data (Fig. 5). In addition, it enables 3D-consistent data augmentation by generating different views of the same interaction. The result is a novel dataset of sound-annotated dataset of 3D hand-scene interactions. We use such dataset to train a model based on rectified flow [27, 46] 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).

We evaluate our model in diverse indoor and outdoor real scenes. We design a psychophysical study to understand how often human subjects misclassify generated for real sounds. The results of the study indicate that the sounds generated by our approach are approximately 40% of the time misclassified for being real. In addition, our approach is significantly better than baselines, which do not use 3D hand poses or visual information about the scene. Finally, we show qualitative results indicating that generated sounds convey rich information about the scene's physical properties, *e.g.*, the materials present in the scene and a notion of the relative distance of objects from the camera's viewpoint.

Overall, the capabilities of our approach indicate a promising path forward to make 3D reconstruction more realistic, immersive, and interactive. Beyond computer graphics, our approach could have large implications in robotics applications by offering a simple way to create photo-realistic multimodal simulators. To make our results accessible, we will release code and data upon publication.

2. Related Work

Multimodal 3D scene reconstruction. A variety of recent works augment 3D reconstructions with other modalities. LERF [21] distills CLIP [39] features into a NeRF [31], which can be used in downstream tasks such as 3D visual grounding [50] and task-oriented grasping [40]. 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 [38] augments a NeRF with a displacement field, which adds temporal information to the NeRF. Recently, PhysGaussian [48] 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, 51]. 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 [52], 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. Frieren [46] 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, 49]. 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, 43] 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.4). Then, we explain the functional form that we use to instantiate F_ϕ (Sec. 3.2).

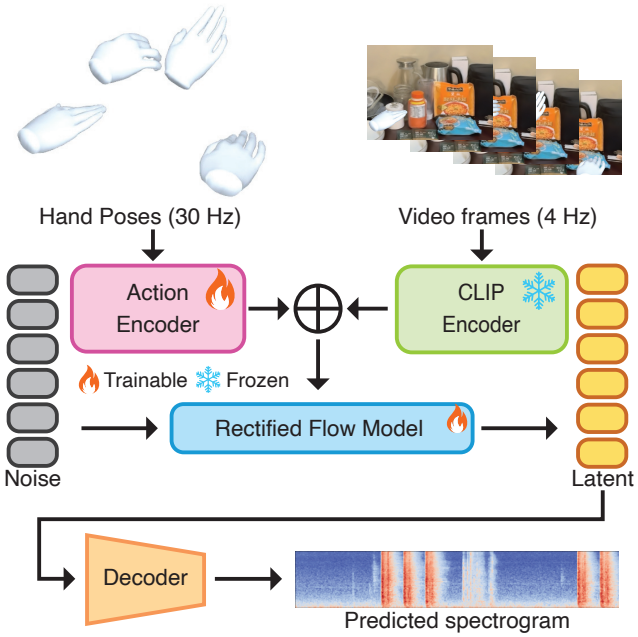

Figure 2. **Sound generation**. We train a rectified flow model [46] to generate a sound spectrogram from a sequence of 3D hand positions and video frames generated from a 3D reconstruction of a scene. The sound can subsequently be converted into a waveform using a vocoder.

3.1. A Dataset of 3D hand-scene interactions with synchronized audio

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 19 different scenes, including bedrooms, lobbies, outdoor trees, and musical instruments (see Fig. 5 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 [41].

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 [37] 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 reg-

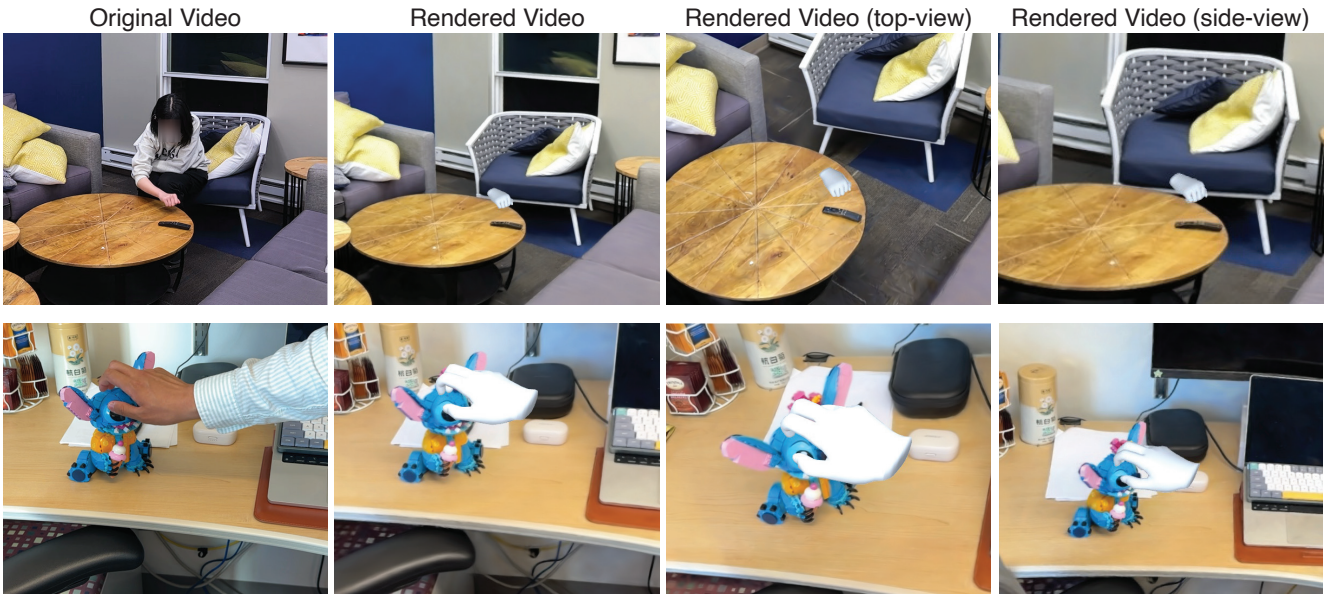

Figure 3. **Data capturing pipeline.** In the original video, a human collector interacts with the scene by performing various actions with their hands. We lift the annotator's hands to the same 3D space of the scene reconstruction. We render a video of the interaction by projecting 3D hands on multiple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

ister the camera on the tripod c to F_θ with COLMAP [41], 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 RGB view of F_θ at the camera position $T_c^{F_\theta}$ (Fig. 3). We label each simulated video $\mathbf{v}$ with the sound $\mathbf{s}$ from the original video of the human interacting with the scene.

We collect approximately 1,800 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 9.4 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. 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 [46]. 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 [39] 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. 5.

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. 6.2).

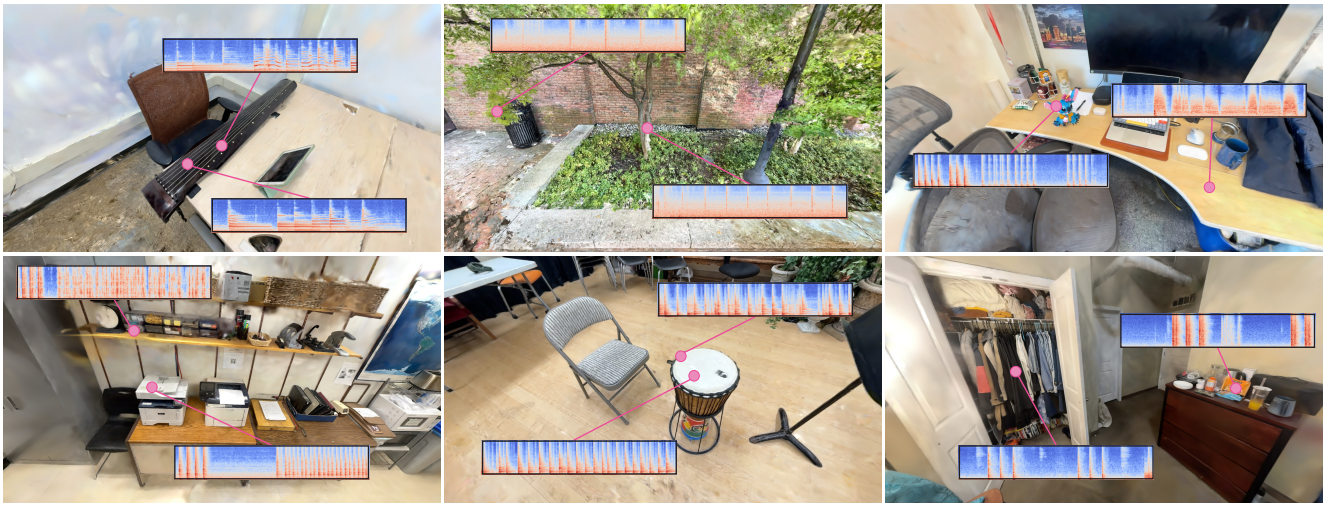

Figure 4. **Representative examples from the dataset.** Our dataset is collected in 19 scenes, including offices, outdoor trees, bedrooms, *etc*. We show six such scenes in the figure above, with examples of action-generated sounds. Our dataset covers a wide range of actions (hitting, scratching, patting, *etc*.) and interacted materials (wood, metal, plastic. *etc*.). In each scene, approximately 1,800 seconds of videos are collected, resulting in a total of 9.4 hours of interaction data.

4. A 3D Visual-Audio Dataset

4.1. Dataset Statistics

Our dataset is collected in 19 scenes including bedrooms, lobbies, outdoor trees, musical instruments, *etc*. We collect approximately 1,800 seconds of videos in each scene, which gives us a total of 9.4 hours of videos. The frame rate of the videos is set to 30 FPS. For each video, we synthesize three videos from different views (original view, top view, side view) using the pipeline described in Sec. **??**, resulting in 28.1 hours of videos in our final dataset.

Fig. **??** shows some representative samples from our dataset. Our dataset includes various hand actions (hit, scratch, squeeze, *etc*.) and materials (tables, plastic bags, trees) in the scene, resulting in distinct sounds of different videos. To the best of our knowledge, this is the first dataset the captures human actions along with its sounds that are spatially aligned with 3D scenes.

5. Implementation Details

We reconstruct the 3D scene using the Splatfacto method from Nerfstudio [44]. Approximately 1K images taken from various views are used for each scene. The gaussians are randomly initialized with scale regularization [48]. During training, we optimize the reconstruction with the Adam [24] optimizer for 20,000 steps on a single NVIDIA RTX 2080 Ti GPU.

5.1. Audio generation model training and inference

Our implementation of F_ϕ is based on Frieren [46] 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 $\mathbf{v}$. Note that, similarly to Frieren, we condition the model on the frames 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, 46], 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 22 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

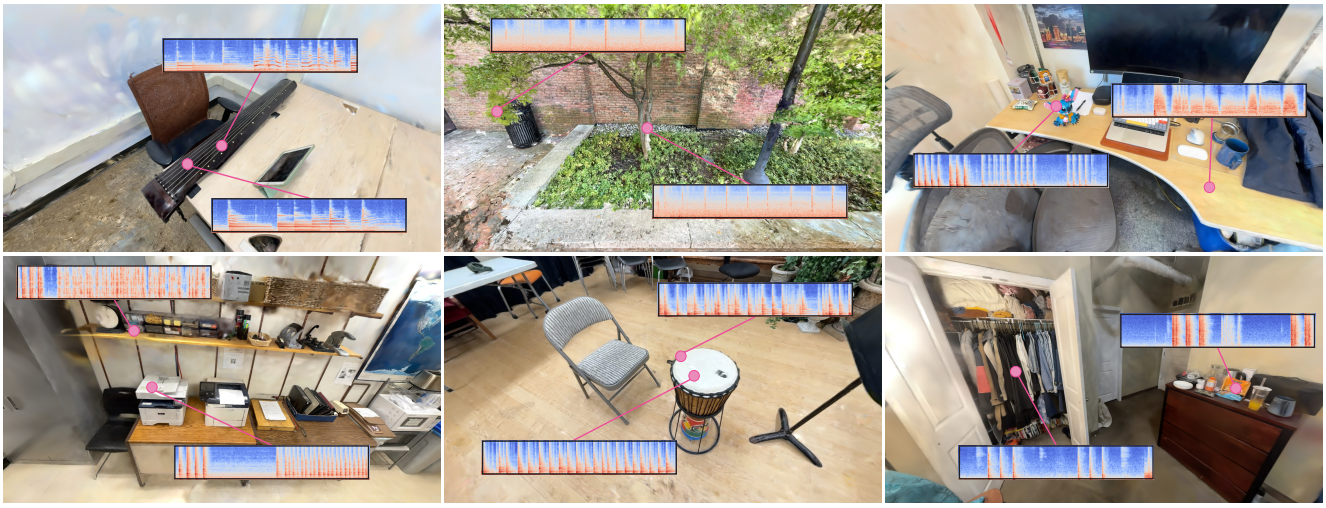

Figure 5. Dataset examples.

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. On the other hand, excluding multi-view data augmentation only slightly affects the performance.

Model variation	STFT ↓	Envelope ↓	FID ↓	IS ↑	CDPAM ($\times 10^{-4}$) ↓	CLAP-acc (%) ↑			Labeled *real* (%) ↑
						all	*action*	*material*	
Full	**0.65**	**0.75**	62.95	15.70	**2.45**	**32.36**	**52.42**	**54.55**	**39.84 ± 2.17**
w/o CLIP	0.94	0.92	**60.84**	16.83	3.65	25.87	47.38	46.32	34.96 ± 2.11
w/o hand pose	0.88	0.87	63.70	15.53	3.02	24.42	45.54	47.97	35.55 ± 2.12
w/o multi-view	0.72	0.78	63.75	**17.46**	2.78	31.59	51.55	53.97	**42.19 ± 2.18**

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 [46]. Finally, a pretrained vocoder [25] is used to transform the spectrogram into a waveform.

6. 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.

6.1. Experimental Setup

We use the following metrics to evaluate the quality of the synthetic 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 [46], 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-shelf CLAP model [47] 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 11 materials (e.g., wood, plastic). From these, we generate a set $\mathbb{P}$ of 77 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 a 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 utilizes linear models to classify materials.

Psychophysical study. We conduct a psychophysical user study to evaluate whether participants can distinguish between generated and real sounds. Sixty-four 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 1027 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.

6.2. 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 smallest impact, resulting in only 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 *action* accuracy, while removing hand pose features most affects *material* accuracy. This aligns with expectations: CLIP features primarily provide material information about the scene, while hand pose features are essential for encoding actions.

In Fig. 6, we present qualitative results for two interactions. Visual inspection reveals that the generated spectrograms contain less background noise than the ground-truth samples. Additionally, we find that removing hand pose features disrupts audio-video synchronization, as visual information alone is insufficient for accurately estimating precise hand motions. Similarly to quantitative results, models trained with and without multi-view augmentation produce relatively similar spectrograms.

6.2.1 Psychophysical study

We measure how often participants misclassify generated for ground-truth sound. Ideally, if the two sounds are completely indistinguishable from each other, we are to observe a miss-classification rate of 50%, which indicates that participants pick at random. The results of this analysis, averaged over all videos and participants, are shown in the rightmost column of Tab. 1. We find that both our full system generates high-quality sounds with a misclassification rate of approximately 40%.

Similarly to what was observed in auditory metrics, we did not find the difference between the full model and the one without multi-view augmentations to be significant ($p = 0.44$ with a two-sided t-test). However, both the full model and the one without multi-view augmentations are better than control baselines. Specifically, the model without multi-view augmentation significantly outperforms the baseline without CLIP embeddings ($p = 0.0175$) and the baseline without 3D hand poses ($p = 0.0293$).

We show results broken down by the properties of the material the hand interacts with in Fig. 7. Participants often misclassified synthetic sounds as real across all categories, though we hypothesize that the reasons varied by category.

For instance, sounds produced by hard materials tend to be sharp and resonant, making them more prototypical and predictable for the model. Conversely, soft surfaces often produce more varied sounds, which are more complex to predict both for our model and the participants.

It is worth noting that our dataset includes fewer interactions with soft than hard materials (342 vs. 690) and fewer with smooth than rough surfaces (273 vs. 759). This imbalance likely contributes to the higher standard error observed in these categories.

6.2.2 Learning about materials and distances from predicting sounds

Our previous experiments suggest that the predicted sounds reflect the scene's physical properties, such as its materials. Assuming that similar sounds imply similar materials, one could potentially use predicted sounds for unsupervised scene segmentation. In addition, assuming that the materials in the scene are not wildly different, the volume of generated sounds might serve as a proxy for estimating object distances from the camera. To test these hypotheses, we designed an interface that automatically selects plausible ac-

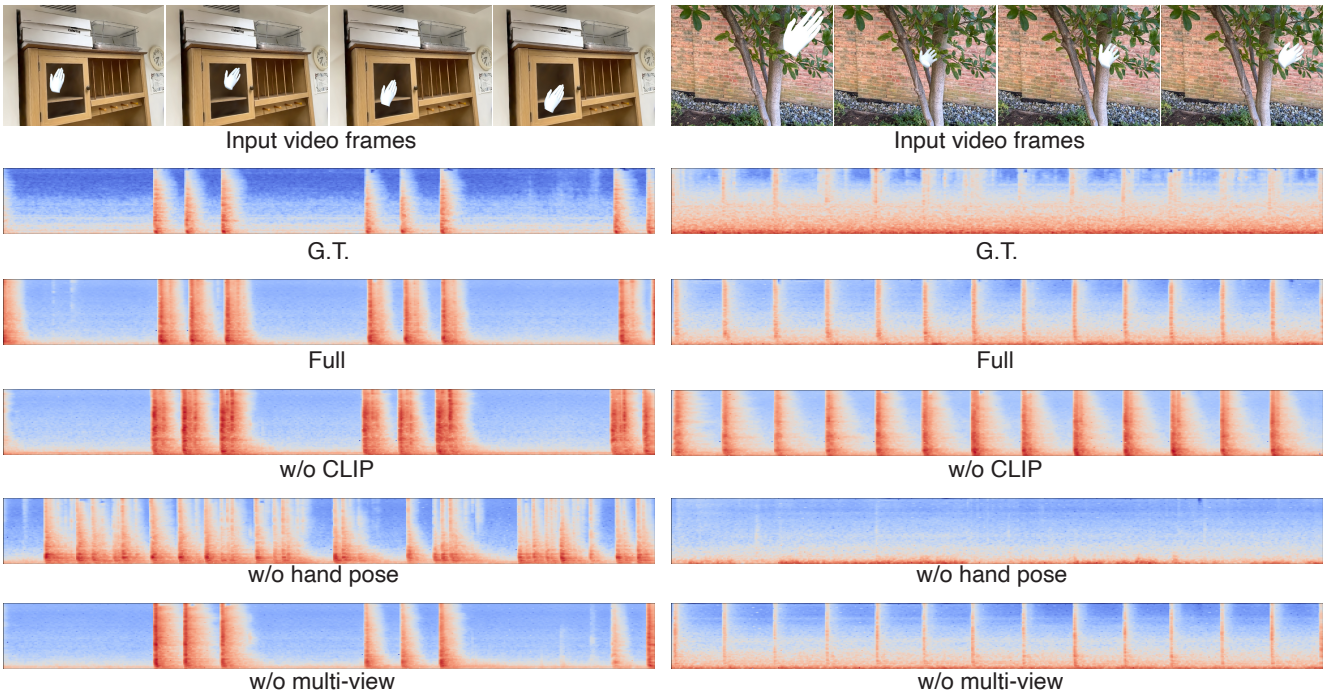

Figure 6. **Qualitative sound prediction results.** We show the spectrogram predictions from our full model and three ablations. We observed that our generated sounds contain less background noise when compared to the G.T. samples. We notice that removing CLIP features softens impact sounds while removing hand pose features results in poor audio-video synchronization. Similar to quantitative results, the model trained without multi-view augmentation performs similarly to the full model.

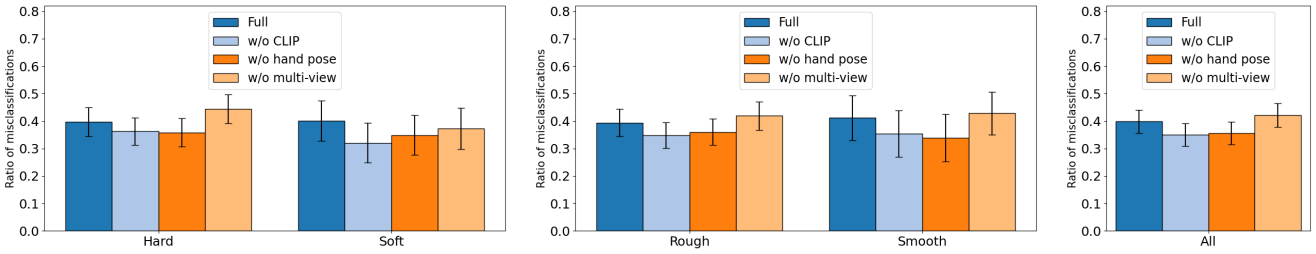

Figure 7. **Results of psychophysical study**. We show the ratio of humans being fooled by different variants of our model. We break down our results into three categories: softness, smoothness, and average over all samples. The error bars show 95% confidence intervals. We find that both our full model and the one trained without multi-view augmentation achieve a misclassification ratio of approximately 40%, indicating the high quality of the generated sounds. In addition, both models outperform baselines without visual or action information.

tions for interacting with various points in a scene's camera view. We start by choosing a sequence of 3D hand poses, $\mathbf{a}^*$ (e.g., patting a flat surface), from the training dataset and estimating the interaction surface normal N_p. Next, we divide the camera view into a 16×16 grid, projecting $\mathbf{a}^*$ onto each grid cell to generate an adapted action $\mathbf{a}_c$. This is done by estimating each cell's surface normal N_c and rotating $\mathbf{a}^*$ to match the angle between $\mathbf{a}_c$ and N_c with that of $\mathbf{a}^*$ and N_p. To produce the corresponding video $\mathbf{v}_c$, we rotate the camera frame so that the cell is centered in the view.

Finally, we pass $\mathbf{a}_c$ and $\mathbf{v}_c$ to F_ϕ to generate the associated sound $\mathbf{s}_c$ for each cell. This effectively simulates the sound that action $\mathbf{a}^*$ would produce if performed at each grid cell's position within the camera view.

We applied this procedure to an indoor scene (Fig. 8). To estimate pseudo-depth, we calculate the average sound volume in each cell as the mean of its squared spectrogram and rank the cells by volume, under the assumption that more distant objects will generate lower amplitude audio. For unsupervised segmentation, we encoded the generated sounds from each cell using a pre-trained CLAP model [47], applied K-means clustering ($K = 4$) to the encodings, and

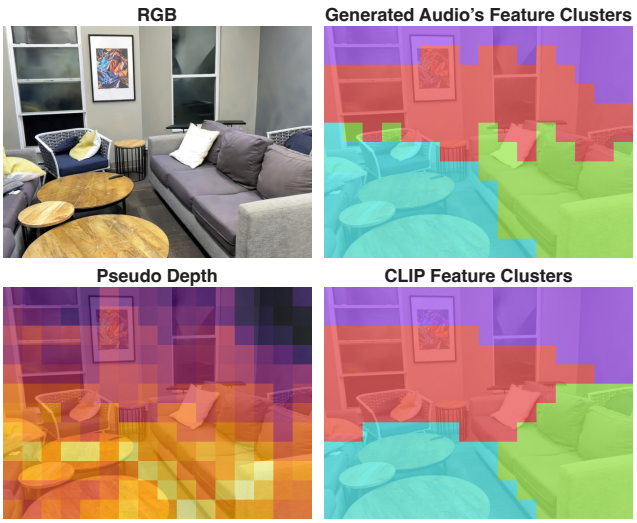

Figure 8. **Material and distance predictions from predicted sounds.** We split the RGB image into grid cells and query our designed interface with the center of each grid cell and a patting hand action, which gives us a predicted sound for each cell. We use the predicted sounds to estimate pseudo-depth by ranking the cells by sound volumes. We also obtain unsupervised segmentation by clustering the CLAP features of the predicted sounds, which achieves results comparable to those of the clusters obtained from CLIP embeddings of novel views of the scene centered in each grid's cell.

assigned a unique color to each cluster. For comparison, we performed the same clustering on CLIP embeddings from the first frame of each cell's video $\mathbf{v}_c$.

Qualitative results, shown in Fig. 8, show a correlation between volume ranking and depth and high similarity between audio and visual feature clusters. Additional qualitative experiments are included in the supplement.

7. Conclusion

Our work makes a step forward towards realistic and immersive 3D scene reconstructions, with promising potential for robotics and VR applications. We do so by predicting the sound of hands interacting with a scene. Both automated evaluations and psychophysical studies show that our synthetic sounds outperform baselines and are often indistinguishable from real sounds, while also providing insight into some physical properties of the scene. However, a key limitation is our approach is that we don't model object dynamics, which could significantly impact actions' sound as scene complexity grows. Addressing this limitation represents an exciting direction for future work.

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. 3

[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, 3

[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. 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, 3

[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, 3

[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. 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. 6

[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, 3, 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. 3

[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. 3

[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, 3, 7

[36] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1

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

[38] 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. 1, 2

[39] 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

[40] 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

[41] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[42] 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

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

[44] 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. 5

[45] 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

[46] 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

[47] 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. 6, 8

[48] 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, 5

[49] 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

[50] 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

[51] 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, 3

[52] 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, 3

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

Supplementary Material

Figure 9. **Example of human study.** We present each user with 32 pairs of videos, as the one above. One has real audio, and the other has synthetic audio generated by our model (in random order). The user is then asked to select the video that sounds more realistic. In the process, we ensured that users were exposed to an equal number of total 512 video pairs for each of the four models: full model, without CLIP, without handpose, and without multi-view.

A.1. Visualization Video

We invite the reader to watch our supplementary video. The video includes the data collection procedure with 3D rendering, a brief introduction of our model, user study, and the representative model prediction results. To judge the quality of the generated sound, we include 6 comparisons between the generated sound and the ground-truth recorded sound at the end of the video. Note that all the results are shown on a testing set unseen at train time.

A.2. Further Psychophysical Study Analysis

Both our full model and the one trained without multiview augmentation generate high-quality sounds with a misclassification rate of approximately 40%, generally outperforming baselines without visual or action information. An example view of the survey is shown in Fig. 9.

We conducted qualitative analysis to break down this result and categorize the interaction videos by hand motions and material properties with which the hand interacts. Synthetic sounds of the hand patting or beating a hard, unified surface (e.g. table, wooden shelf, white board, sofa armrest, tree trunk, etc.) have a higher misclassification rate than those of the hand rubbing or scratching cluttered small objects. We hypothesize that this is because large pieces of unified hard materials tend to make unified and prototypical sounds, which are easier and more straightforward for our models to learn and predict. In contrast, hand interactions with cluttered objects are more likely to cause irregular deformation or displacement due to different material properties, creating subtle and dynamic sound changes that are harder for our model to capture. Indeed, these cases have much lower misclassification rates.

We also noticed that the misclassification rate increases when the image reconstruction quality from the 3D scene decreases since it is difficult for humans to understand exactly what is being touched. Finally, while generally realistic, synthetic audio often sounds more "scratchy" and less sharp than real audio.

A.3. More examples on material and distance predictions from predicted sounds.

We perform a qualitative study to understand whether the generated sounds convey information about the physical properties of the material being touched. Fig. 10 shows further qualitative results of unsupervised segmentation and pseudo-depth estimation. These results show that pseudo-depth estimation using as a proxy the volume of the predicted sound is often not much aligned with the real depth of the scene, except if the visible materials are relatively similar. Such correlation between material and sound is evident in the unsupervised object segmentation results. Indeed, we occasionally found that the clusters of synthetic audio features convey sharper information about the materials than their respective CLIP embeddings. In Fig. 10, for instance, this happens with the white door in the second row and the closet in the third row. Overall, these experiments show that our approach conveys a deeper physical understanding of the scene.

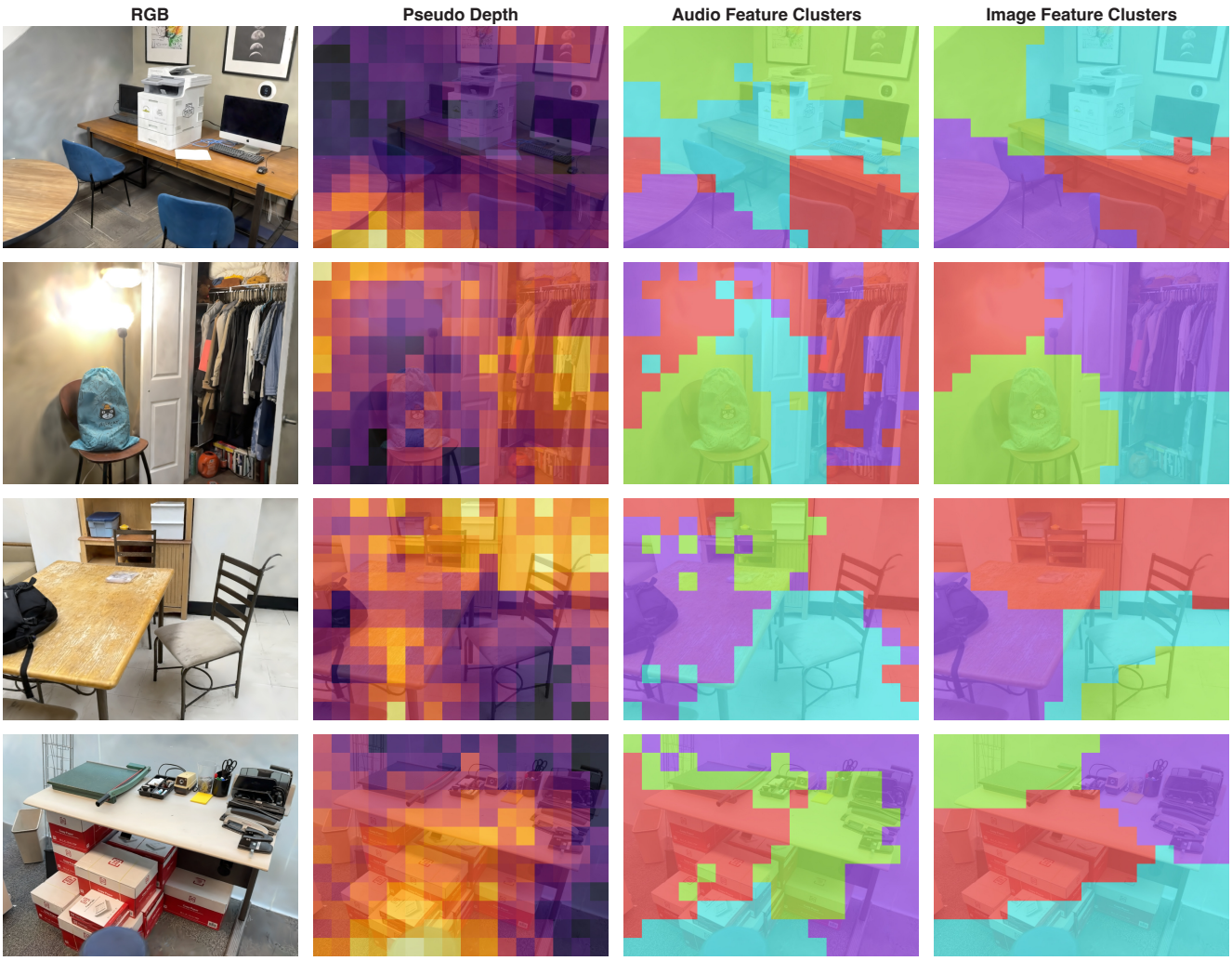

Figure 10. **Material and distance predictions from predicted sounds.** We split the RGB image into grid cells and query our designed interface with the center of each grid cell and a patting hand action, which gives us a predicted sound for each cell. We use the predicted sounds to estimate pseudo-depth by ranking the cells by sound volumes. We also obtain unsupervised segmentation by clustering the CLAP features of the predicted sounds, which achieves results comparable to those of the clusters obtained from CLIP embeddings of novel views of the scene centered in each grid's cell. We found pseudo-depth estimation to be generally low quality, especially if the materials in the scene sound very different. However, segmenting objects according to their sound qualitatively looks as good (and at times better), than the one obtained by clustering visual features.