Visual Acoustic Fields

Yuelei Li[1] Hyunjin Kim[1] Fangneng Zhan[2,3] Ri-Zhao Qiu[1] Mazeyu Ji[1] Xiaojun Shan[1]
Xueyan Zou[1] Paul Liang[3] Hanspeter Pfister[2] Xiaolong Wang[1]
[1]UC San Diego [2] Harvard University [3]MIT

Figure 1. Overview of Visual Acoustic Fields, a novel framework for integrating visual and auditory signals within a 3D scene. Our approach leverages 3D Gaussian Splatting (3DGS) to represent the scene and associates it with impact sounds. The framework enables two key tasks: vision-conditioned sound generation, where impact sounds are synthesized based on impact location, and sound localization, where the model identifies the source of a given sound within the 3D environment.

Abstract

Objects produce different sounds when hit, and humans can intuitively infer how an object might sound based on its appearance and material properties. Inspired by this intuition, we propose Visual Acoustic Fields, a framework that bridges hitting sounds and visual signals within a 3D space using 3D Gaussian Splatting (3DGS). Our approach features two key modules: sound generation and sound localization. The sound generation module leverages a conditional diffusion model, which takes multiscale features rendered from a feature-augmented 3DGS to generate realistic hitting sounds. Meanwhile, the sound localization module enables querying the 3D scene, represented by the feature-augmented 3DGS, to localize hitting positions based on the sound sources. To support this framework, we introduce a novel pipeline for collecting scene-level visual-sound sample pairs, achieving alignment between captured images, impact locations, and corresponding sounds. To the best of our knowledge, this is the first dataset to connect visual and acoustic signals in a 3D context. Extensive experiments on our dataset demonstrate the effectiveness of Visual Acoustic Fields in generating plausible impact sounds and accurately localizing impact sources. Our project page is at https://yuelei0428.github.io/projects/Visual-Acoustic-Fields/

1. Introduction

Our lives are filled with objects that produce distinct sounds. For example, striking a ceramic cup and a wooden table in a living room scene produces completely different sounds.

Studying the cross-modal connection between vision and impact sounds is important because it represents one of the most fundamental ways we learn about the physical world: infants explore their surroundings by simultaneously observing and interacting with objects, and studies have shown that this process helps them develop an intuitive understanding of physics [3, 20, 39]. Such an understanding can enhance applications that require reasoning about the physical properties of interactive objects, such as robotics [17, 18, 21], virtual reality [6, 42], and content creation [45, 46].

Existing cross-modal datasets that connect vision and sound typically pair a whole 2D image or video depicting a large scene with a soundtrack. These datasets encompass a diverse range of sounds, including impact sounds [13, 21, 32], speech [1, 7, 44], and background music [27, 50]. However, in all these datasets, the paired audio represents a holistic soundscape that describes the entire scene rather than isolating the specific source producing the sound. As a result, such datasets do not provide physical information about the precise sound source within the scene. Furthermore, they do not capture the relationship between auditory and visual signals in 3D.

In this work, we aim to collect spatially aligned visual-sound pairs in 3D space. This is a challenging task because we must determine the location of each local visual signal in the scene (e.g., a subpart of an object) within the entire scene, and this centimeter-level alignment of visual supervision and audio supervision is hard to achieve. Since no existing sound dataset is available for 3D scenes, we refer to visual-tactile datasets, which commonly involve a similar step of localizing signals within a scene. Early work relies on robotic arms capturing signals in controlled settings, where the gripper's location is determined via forward kinematics [4, 5, 24, 28]. [8] moves a step forward to enable signal localization at the scene level, but it requires a specialized device and a complex camera calibration step.

To solve the above problems, we propose an easy-to-use pipeline for localizing data at the scene level without requiring any new devices or calibration. Our approach relies solely on a smartphone to capture images and record sounds. First, we collect a set of multiview images of the scene. Next, we capture a set of labeled images, each corresponding to a location where an impact occurs, along with the associated sound. We use structure-from-motion [37, 38] to estimate the camera poses for all images. The multiview images are then used to reconstruct the scene using 3DGS [22], and the camera poses of the labeled images help determine the impact locations within the reconstructed scene.

With the collected dataset, we propose **Visual Acoustic Fields** to bridge visual and auditory signals in 3D scenes (see Fig. 1). Based on 3DGS [22], the Visual Acoustic Fields learn a set of 3D Gaussians augmented with AudioCLIP [15] features, which are supervised by the AudioCLIP embeddings of the collected multiview images. The AudioCLIP associates the visual and auditory signals, allowing to perform two tasks including (1) vision-conditioned sound generation and (2) sound localization. For sound generation from impact points, we infer the AudioCLIP feature of the impact point, which is further mapped to sound with an audio diffusion model. Specifically, we modify and finetune a pre-trained Stable Audio model [9] for the mapping to take advantage of its generalization ability. For sound localization, we adopt contrastive pretraining [15, 35] to finetune AudioCLIP on our collected dataset, which is used to encode the query sound. Then, a region or object can be localized by measuring the relevancy between the sound and visual encoding.

We conduct experiments on our collected dataset to evaluate our sound generation and localization models. Our experiments indicate that:

- The impact locations of collected visual-sound pairs can be localized in 3D space by learning a radiance field with synchronized camera poses.
- Predicted impact sounds in our Visual Acoustic Fields accurately align with the corresponding impact locations.
- Impact regions or objects can be precisely retrieved with sound using our Visual Acoustic Fields.

2. Related Work

2.1. Visual-Sound Datasets.

Most existing cross-modal datasets that connect vision and sound focus on providing general background soundtracks for given images (e.g., a cheerful soundtrack for images of a sunny outdoor scene, etc.) [14, 33, 36]. These general soundtrack datasets are easier to collect at scale from the Internet, whereas high-quality hitting sounds are not as readily available online. To date, only two datasets have specifically focused on hitting sounds: ObjectFolder [11–13] and The Greatest Hits Dataset [32]. ObjectFolder includes representations of individual neural objects in the form of implicit neural fields with simulated multisensory data. The Greatest Hits Dataset features videos of objects struck by a drumstick without the label for hitting position. Our dataset differs from these in two significant aspects: 1) it operates at the scene level rather than the object level, and 2) the collected visual-sound pairs are spatially registered within a 3D scene represented by 3DGS.

2.2. Predicting Sound from Visual Inputs

Predicting sound from visual inputs is a fundamental problem in cross-modal learning with applications in robotics, virtual reality, and content generation. Most methods focus on a holistic generation that yields a single soundtrack for the whole scene [30, 41, 43]. For instance, Owens et al. [32] introduced a self-supervised model for impact sound

generation, while Zhou et al. [49] learned direct mappings from video frames to sound representations. More recently, Sheffer and Adi [40] developed *im2wav*, which conditions sound synthesis on images but lacks spatial awareness.

Beyond scene-wide generation, recent approaches explore object-specific impact sound prediction. Gan et al. [10] introduce a system capable of synthesizing plausible music for silent video clips depicting people playing musical instruments. Multimodal and physics-informed approaches further improve accuracy by integrating visual, auditory, and tactile data. Su et al. [42] leveraged physics-driven simulation for realistic sound modeling, and Zhou et al. [48] incorporated material-aware conditioning into audio-visual segmentation. Gao et al. [11, 12] developed multisensory datasets, *ObjectFolder* and *ObjectFolder 2.0*, enabling cross-modal learning for object interactions. However, the above works fall short in sound generation in 2D space or virtual scenarios. In contrast, our method allows interaction with the real 3D environment by navigating in the learned Visual Acoustic Fields, enabling fine-grained impact sound synthesis at arbitrary 3D locations.

2.3. Sound Localization

Sound localization refers to the task of identifying the source of a sound within a scene. Many works leverage the natural synchronization between visual and auditory information to achieve localization. Arandjelovic and Zisserman [2] introduced a *self-supervised* approach to associate *spatial regions* in video frames with corresponding sounds, providing *pixel-level* audio localization. Similarly, Zhao et al. [47] proposed to localize sound sources by leveraging temporal coherence in video sequences, achieving *patch-level* localization. Besides localization, recent works have also explored *audio-visual segmentation (AVS)*. Zhou et al. [48] proposed a transformer-based model that applies *pixel-wise attention* to capture detailed audio-visual correspondences. To include explicit object-level alignment, Huang et al. [19] defines *audio queries* to explicitly associate sound features with individual objects and improves sound localization accuracy.

While most audio-visual learning methods operate on 2D video frames, their applicability remains limited in 3D. Jatavallabhula et al. [21] proposed ConceptFusion, which integrates audio and visual signals within a 3D representation, enabling object localization in 3D scenes. However, this method does not focus on hitting sounds; neither its sound localization model nor its dataset has been open-sourced. In contrast, we will open-source both our model and dataset.

3. Method

3.1. Data Collection Pipeline.

To train our Visual Acoustic Fields, we require multiview images annotated with both impact sounds and their corresponding locations. However, collecting such a dataset in the real world is a challenging task. While hitting sounds can be recorded using appropriate audio equipment, accurately identifying the impact locations in the captured multiview images remains a significant challenge. A straightforward solution is to use markers to label these locations; however, the presence of markers introduces artifacts into the images, which can interfere with model training. To address this issue, we propose a novel re-rendering strategy for data collection, as illustrated in Fig. 2.

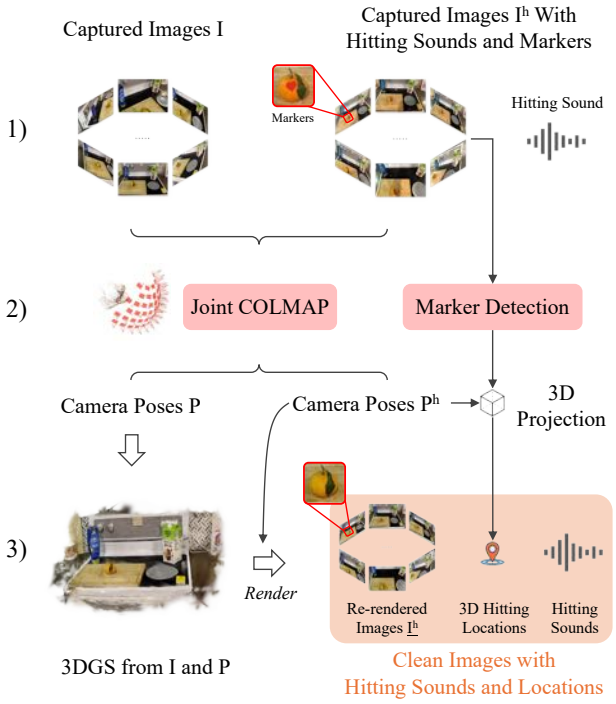

Figure 2. **Pipeline for data collection.** A novel re-rendering strategy is proposed to enable accurate annotation of impact sounds and their locations without introducing artifacts. 1) We capture two sets of multiview images including I of the scene and I^h marked with visible hitting markers and synchronized with corresponding hitting sounds. 2) Using *Structure-from-Motion (SfM)*, we jointly estimate camera poses of I and I^h, as denoted by P and P^h, respectively. The impact locations can be obtained by detecting the markers with OWL-v2 [31], which are further projected to 3D location with known camera poses P^h and depth map. 3) A 3DGS can be trained with multiview images I and camera poses P. The images with impact locations are re-rendered without markers from the 3DGS with camera poses P^h, yielding clean images $\underline{I^h}$ with paired hitting sounds and their hitting positions.

1) Image Capturing and Sound Collection. For each scene, we capture two sets of images: $I = \{i_n\}_{n=1}^{N}$ and $I^h = \{i_m^h\}_{m=1}^{M}$ with size $H \times W$. The set I consists of multiview images of the scene, collected by moving through the environment while recording a video to ensure dense

3D coverage. The set I^h comprises images where hitting is performed, with markers indicating the corresponding hitting positions. These markers can be small stickers, laser-projected patterns, or other convenient visual indicators.

For the markers in each image in I^h, we record a corresponding hitting sound. A metallic coffee stick is used to strike all the marker locations, ensuring consistency across all data. To remove background noise, we apply spectral gating, which estimates the noise profile using short-time Fourier transform (STFT), subtracts it from the signal, and reconstructs the denoised audio via inverse STFT. We then crop or pad the recording to 0.5 seconds to include the clean hitting sound. Since the force applied when hitting different locations may vary slightly during data collection, resulting in differences in sound amplitude, we apply Root Mean Square (RMS) normalization with a scale of 0.01 to standardize the inputs for our models.

2) Joint COLMAP and Hitting Localization To estimate camera poses, we use COLMAP [37, 38], obtaining pose sets $P = \{p_n\}_{n=1}^N$ for I and $P^h = \{p_m^h\}_{m=1}^M$ for I^h. However, because COLMAP assigns a random coordinate origin during each execution, separately estimating the poses of I and I^h results in inconsistent camera coordinate systems between P and P^h. To ensure alignment, we perform joint pose estimation, processing both sets of images together in a single COLMAP run. Notably, the markers in I^h are small enough to have a negligible impact on COLMAP's accuracy in practice.

To locate the collected sound within the 3DGS, we first apply an object detection network OWL-v2 [31] on the $\{i_n^h\}_{n=1}^N$ to obtain the pixel locations $\{(x_n^h, y_n^h)\}_{n=1}^N$ of the markers. With $\{(x_n^h, y_n^h)\}_{n=1}^N$ and $\{i_n^h\}_{n=1}^N$, we can follow the standard pinhole camera model to locate the hitting point in the camera coordinate frame:

$$(i_n, j_n, k_n) = \left(\frac{(x_n^h - c_x) \cdot Z_c}{f_x}, \frac{(y_n^h - c_y) \cdot Z_c}{f_y}, d_n \right) \quad (1)$$

where (x_n^h, y_n^h) is the detected marker position in image i_n^h, d_n is the depth value at (x_n^h, y_n^h) estimated by the 3DGS, f_x and f_y are the focal lengths in pixels, and (c_x, c_y) is the principal point of the camera.

3) Re-rendering. We reconstruct a clean 3D scene without markers from images I and poses P with 3DGS. Notably, camera poses P and P^h are in the same coordinate systems thanks to our joint camera pose estimation with COLMAP. Thus, to obtain a clean version (without markers) of I^h, we can query the 3DGS with camera poses P^h to re-render a new set of images without markers, which we denote as $\underline{I^h}$. Finally, we obtain a dataset with paired multiview images, hitting locations, and hitting sounds.

3.2. Predicting Sound from Visual Signals

With the collected dataset, we can train a model that can map visual inputs and hitting positions to the corresponding hitting sounds. However, accurately estimating the impact sounds requires identifying the specific object or region producing the impact, associating it with relevant acoustic features, and synthesizing realistic sound variations. To achieve this, we incorporate three key components: Segment Anything Model (SAM) for object segmentation, AudioCLIP for vision-audio feature alignment, and a pre-trained Stable Audio model for high-fidelity sound generation (see Fig. 3).

Object Segmentation with SAM. Some objects produce the same sound when hit at different locations, while many objects consist of multiple components that produce distinct sounds when struck. Thus, a key challenge in sound prediction is ensuring that the model focuses on the correct region. To address this, we leverage the Segment Anything Model (SAM) [26] to segment images at multiple levels, including subpart-level, part-level, and whole-object-level. Then, the hitting region can be localized by selecting the object segmentation (at multi-level) where the hitting position lies. This hierarchical segmentation allows our model to better capture and localize the hitting regions for various objects.

AudioCLIP for Vision-Audio Feature Alignment. Even with precise segmentation, predicting plausible sounds requires a meaningful feature representation that bridges the visual and auditory domains. Traditional feature extractors, such as CLIP [35], are optimized for image-text alignment but lack audio-specific embeddings. To overcome this, we employ AudioCLIP [15], which extends CLIP by incorporating audio representations alongside vision and text. AudioCLIP enables the extraction of semantically rich, multimodal embeddings that align visual textures with their corresponding sound characteristics. Moreover, AudioCLIP supports zero-shot generalization, allowing the model to fit diverse real-world scenes. In practice, we use AudioCLIP to extract the features of the multi-level object segmentations where hitting is performed. Building on this, we follow [34] to construct a feature-augmented 3DGS, enabling 3D view-consistent visual features at any location in the scene.

Pre-Trained Stable Audio for Sound Generation. We then use a model M to predict the sound x at a hitting location based on the corresponding multi-level features $\{f^s, f^p, f^w\}$ as $M(x|\{f^s, f^p, f^w\})$. However, generating high-quality, realistic impact sounds from a limited dataset is a major challenge, as training such a model from scratch would require an extensive dataset of diverse impact sounds. To address this, we fine-tune a pre-trained Stable Audio model [9], a state-of-the-art and text-conditioned audio generation model based on the diffusion transformer. This pre-trained Stable Audio offers several advantages, including enabling generalization to diverse objects by retaining prior knowledge about a wide range of impact sounds and achiev-

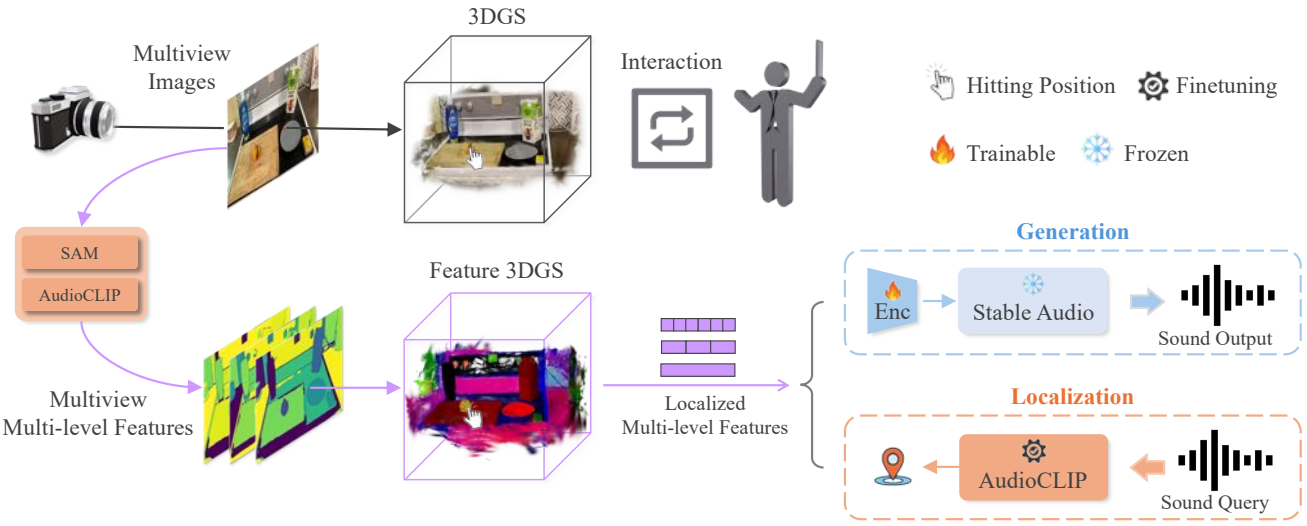

Figure 3. Overview of the Visual Acoustic Fields framework. The model consists of two main components: sound generation and sound localization. Given multiview images, a feature-augmented 3D Gaussian Splatting (feature 3DGS) representation is constructed. For sound generation, localized multi-level features queried from the feature 3DGS are used as conditions to fine-tune a pretrained Stable Audio diffusion model to synthesize impact sounds. For sound localization, a fine-tuned AudioCLIP encoder maps input audio queries to the feature 3DGS, allowing the model to localize the corresponding impact location by computing feature similarity. Trainable, frozen, and fine-tuned components are indicated in the diagram.

ing training efficiency by fine-tuning only a small part of the model. Specifically, we replace its original conditioning mechanism with a multi-level visual feature conditioner and shorten the generated sample length to match the duration of a hitting sound. Only the visual feature conditioner is trained during fine-tuning, while the transformer weights remain frozen to preserve the model's generalization ability.

Inference Pipeline. During inference, for any queried impact location, we extract multi-scale SAM segmentation features and AudioCLIP embeddings from the feature 3DGS rendering. These embeddings are then passed as conditions to the Stable Audio model, which synthesizes the corresponding impact sound.

3.3. Localizing Sound in 3D

To better understand the relationship between vision and sound, we also introduce the task of sound localization: given a hitting sound, the goal is to predict the location in the scene that produced it (see Fig. 3). To achieve this, we first train AudioCLIP [15], a cross-modal visual-sound encoder, using self-supervised contrastive learning on our training dataset, drawing inspiration from the text-image contrastive pre-training commonly used in image generation [35]. We then leverage this visual-audio encoder and SAM [26] to encode multiview images of each scene and extract visual features. These features are subsequently used to train the feature-augmented 3DGS [34] for each scene. After we have the feature-augmented 3DGS, we can query it with sound to localize the sound's origin.

Inference Pipeline. Given a hitting sound and a viewpoint, we first render the feature-augmented 3DGS from the specified viewpoint to obtain the rendered visual embedding $\phi_i \in \mathbb{R}^{(H\times W)\times d}$, where d is the embedding dimension. We then encode the hitting sound using the fine-tuned AudioCLIP to extract the audio features $\phi_s \in \mathbb{R}^{1\times d}$. Next, we compute the relevance score using a dot product operation, $\phi_i \cdot \phi_s \in \mathbb{R}^{(H\times W)\times 1}$. The region with a higher relevance score will be predicted with greater confidence as the localization result.

4. Dataset Statistics

We collected data from 15 different scenes, including a bedroom, kitchen, bathroom, office, library, tabletop, various corners in a teaching building, etc. The sound sources in our dataset include materials such as wood, ceramic, plastic, metal, LCDs, etc. Each scene contains between 100 and 200 data points, depending on the diversity of sound sources present. In total, our dataset comprises approximately 2,000 visual-sound data pairs. Fig. 4 provides examples of some of the collected scenes, and we include all 15 scene images in our supplementary materials.

5. Experiments

We conduct all experiments on our collected datasets. In all experimental settings, we adopt a 4:1 train-test split ratio.

Figure 4. **Example scenes in our dataset.** Our dataset consists of 15 diverse environments, including indoor and outdoor settings such as a bedroom, kitchen, bathroom, office, library, coffee corner, and garden. Each scene contains various materials (e.g., wood, metal, plastic, ceramic) and impact locations, yielding a rich collection of visual-audio pairs for training and evaluation.

Dataset	Scenario	Source	3D
ObjectFolder [11]	Object	Synthetic	✓
ObjectFolder 2.0 [12]	Object	Synthetic	✓
ObjectFolder Real [13]	Object	Robot	✓
Visually Indicated Sound [32]	Scene	Human	×
AudioSet [14]	Scene	Internet	×
Ours	Scene	Human	✓

Table 1. We compare our dataset with existing hitting sound datasets in terms of scenario (object-level vs. scene-level), data source (synthesized by simulators, collected by robot, or human-collected), and whether the dataset contains 3D spatial information. Unlike prior datasets focusing on object-level interactions or synthetic environments, our dataset captures real-world, human-collected impact sounds at the scene level with full 3D spatial alignment.

5.1. Hitting Sound Generation.

To evaluate the quality of our hitting sound generation, we compare our model with im2wav [41], the only open-sourced image-to-audio generation model available retrained on our dataset. Additionally, we investigate the impact of using object-level features versus local texture features around the hitting positions as conditioning inputs for sound inference.

In Tab. 2, im2wav [41] refers to the model conditioned on object segments obtained from SAM [26], while im2wav (local) uses 100×100 rendered images centered at the hitting locations as the input. Ours represents our approach, which conditions the model on multi-level object features extracted from SAM [26] and CLIP [15]. In contrast, Ours (local) conditions the model on local features extracted from 100×100 rendered images centered at the hitting position using ResNet-34 [16]. The metric we use includes Frechet Audio Distance (FAD) [25], Kullback-Leibler Divergence (KL), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR). Similarly to the Frechet Distance in image generation, FAD is widely used in audio generation to measure the distance between the generated and real distributions. KL is computed at the paired sample level, then summed and averaged to obtain the final result. Our method achieves lower FAD and KL scores compared to others, indicating that our proposed model more effectively captures the distributions and characteristics of the hitting sound data. At a low level, our method also attains higher SSIM and PSNR scores, demonstrating that our generated results more faithfully resemble their ground truth counterparts.

Besides computing sound metrics, we also evaluate the quality of our generated sounds through a survey, as we believe human perception is always the *gold standard* for evaluation. In our survey, we randomly select 50 hitting location images from our test set, ensuring a diverse range of hitting materials. For each location, we provide the ground truth hitting sound we collected, the sound generated by our method, and the sound generated by im2wav [41]. Participants are then asked to choose the sound they believe best matches the hitting sound at the given location. We collected 30 responses through a Google Form, and the results are shown in Tab. 3. Among the total $30 \times 50 = 1500$ survey data points, 42.93% of real sounds and 41.93% of our generated sounds were selected as the ones that best matched the

Method	FAD ↓	SSIM ↑	PSNR ↑	KL ↓
im2wav (local)	1.50	0.67	14.65	0.52
im2wav [41]	1.68	0.67	15.03	0.48
Ours (local)	0.66	0.77	18.68	0.43
Ours	0.35	0.82	20.83	0.38

Table 2. **Sound Generation Results.** We compare our method with im2wav [41] and several variants with various evaluation metrics. Our approach achieves the best results across all metrics, demonstrating its ability to generate impact sounds that closely match real-world recordings.

hitting location, which indicates that our generated sounds are almost indistinguishable from real sounds.

	Ground-truth	Ours	Im2wav [41]
Best Match	42.93 %	41.93 %	15.13%

Table 3. **Sound Generation User Study.** We conducted a user study to evaluate the perceptual quality of generated impact sounds. Participants were presented with images of impact locations along with three sound samples: the ground-truth recorded sound, our generated sound, and the sound generated by im2wav [41]. They were asked to select the sound that best matched the given impact location. The results show that our method produces nearly indistinguishable sounds from real impact sounds.

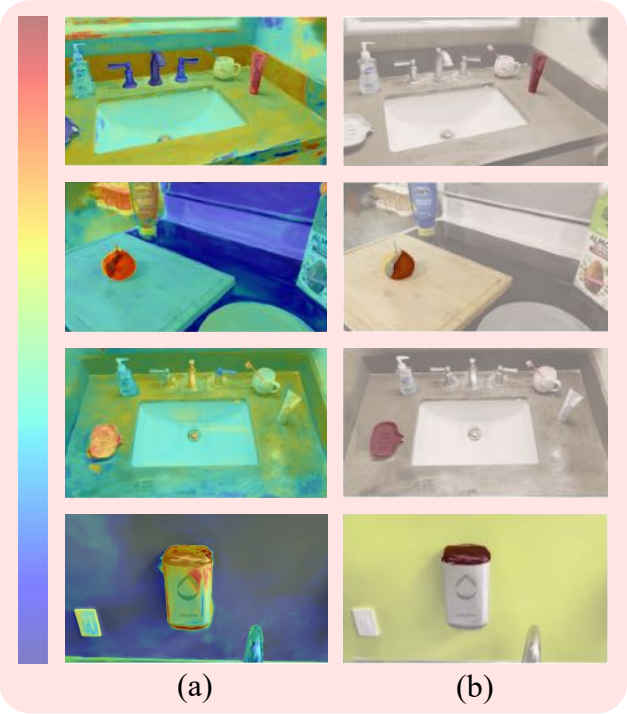

Figure 5. Visualization of sound localization results. Given an input hitting sound, our model predicts the most relevant impact location within the 3D scene. (a) The heatmap represents the localization confidence scores, where brighter regions indicate higher confidence for the predicted sound source. (b) The highlighted region denotes the final localized impact objects (or parts).

5.2. Hitting Sound Localization.

To quantitatively evaluate our method, we first use SAM [26] (default scale) to segment the rendered scene from 3DGS. Next, we encode each segmented image along with the given hitting audio using our fine-tuned AudioCLIP [15]. Finally, we compute the relevance score between the query audio and each segmented part, and the part with the highest relevance score is chosen as the localization result.

We compare our method with AudioCLIP without fine-tuning, which utilizes the AudioCLIP model trained on the large-scale AudioSet [14], and with Random, which selects a segment in the scene at random as the localization result. We manually label each result as correct or incorrect, as a single sound can sometimes correspond to multiple segments identified by SAM. For example, in the bathroom scene shown in Fig. 4, the metal faucet is segmented into three parts (left, middle, and right) by SAM. However, since all three segments produce the same sound, we consider the result correct if a faucet sound is mapped to any of these segments.

As shown in Table 4, we measure Acc(1) and Acc(3). Acc(1) represents the accuracy rate when the segment with the highest relevance score is the correct result, while Acc(3) indicates the accuracy rate when the correct location is covered by one of the top three segments with the highest relevance scores. Our method significantly outperforms the baselines in both top-1 and top-3 localization accuracy rates. This suggests that existing large-scale sound datasets, such as AudioSet [14], lack a sufficient number of clean and high quality hitting sound samples to train a contrastive learning model and effectively perform localization tasks, despite the prevalence of hitting sounds in everyday life. With our collected dataset and proposed pipeline, we can effectively localize sound in a 3D scene.

In Fig. 5, we present a visualization of the sound localization results. The left column shows the heatmap generated when using the hitting sound to query rendered features at every pixel. The right column highlights the segmented region selected by our method. These results demonstrate that our approach effectively identifies impact locations with high accuracy, even in complex environments.

In Fig. 6, we present visualizations of our generated sound, the baseline, and the groundtruth sound. The Mel spectrogram reveals that our generated sounds exhibit similar frequency content and temporal evolution compared with

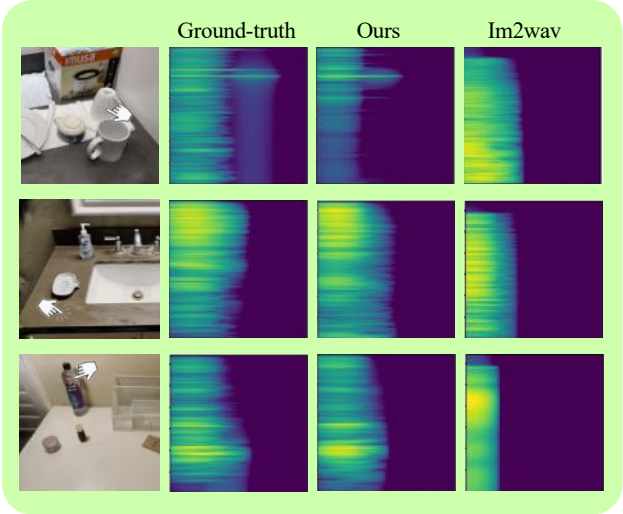

Figure 6. Mel Spectrogram of generated sounds. We compare the audio generated by our method with ground-truth recordings and results from im2wav [40]. The visualized spectrograms of sounds show that our approach produces impact sounds that closely resemble real recordings.

	AudioClIP [15]	FT-AudioClIP (ours)	Random
Acc(1)	19.0%	74.4%	10.2%
Acc(3)	35.9%	85.5%	26.8%

Table 4. **Sound Localization Accuracy.** We compare our method (FT-AudioCLIP) with the baseline AudioCLIP [15] (without fine-tuning) and the random selection. **Acc(1)** denotes the accuracy when the top-1 predicted location is correct, while **Acc(3)** represents the accuracy when the correct location is within the top-3 predictions. Our fine-tuned model significantly outperforms both baselines.

the groundtruth sound. This aligns with our survey results, which indicate that our generated sound is nearly indistinguishable from the ground truth sound.

5.3. Implementation Details

RGB and Feature Field. We follow the official 3DGS implementation [23] to render RGB images. For feature extraction, we use the Langsplat codebase [34] to obtain multilevel SAM features and reconstruct the feature fields. During data collection, we capture a video for each scene and uniformly sample approximately 300 frames from it as the training set for scene reconstruction. We train both the RGB fields and feature fields on a single NVIDIA GeForce RTX 3090 GPU.

Sound Generation and Localization. For sound generation, our implementation builds upon Stable Audio Open [9], using its released checkpoint as the starting point for training. We modify its conditioning encoder to process visual features by incorporating MLPs. During training, we optimize the conditioning encoder using the AdamW optimizer with a base learning rate of $5e^{-5}$ and a batch size of 8 on a single NVIDIA A100 GPU. For inference, we apply classifier-free guidance with a scale of 6 and perform 250 sampling steps. To evaluate sound quality, we use the codebase from [29]. For sound localization, we train AudioClIP [15] on our dataset using an SGD optimizer with a learning rate of $5e^{-5}$ on a single NVIDIA GeForce RTX 3090 GPU. We initialize the image encoder with pretrained CLIP [35] weights and freeze it during training. Only the weights of audio encoder part are updated.

Baseline. For the baseline, we use the open-sourced implementation of im2wav [41]. Since, by default, the model cannot handle very short sequences, such as our hitting sounds, we pad them to 10 seconds for training and inference, then crop them afterward for metric calculation. When training im2wav, we use a batch size of 16 for the VQVAE and a batch size of 8 for the upsampling and low-level models. All baseline training is done on a single NVIDIA GeForce RTX 3090 GPU.

6. Conclusion, Limitation, and Future Work

In this paper, we introduce Visual Acoustic Fields, a novel framework that integrates visual and acoustic signals within a 3D scene. Our approach leverages 3D Gaussian Splatting to establish a spatially consistent 3D scene representation for interactive sound generation. By incorporating Audio-CLIP features, our model enables both vision-conditioned sound generation and sound localization. To support this research, we propose a re-rendering strategy for dataset collection, providing sound annotations with accurate impact locations. Extensive experiments on our newly curated dataset demonstrate the effectiveness of our approach in generating plausible hitting sounds and accurately localizing sound sources in 3D space.

Limitations and Future Works. Despite the promising results, our approach has several limitations. First, our dataset, while diverse, remains limited to static scenes and a variety of materials. Expanding the dataset to include more environments, such as including dynamic scenes and more outdoor settings, and objects with richer acoustic properties would enable better generalization of our model. Another limitation of our Visual Acoustic Fields model is its agnosticism of the spatial location of the listener. Specifically, our model does not account for the perceived sound changes based on the listener's position. To overcome this limitation, future research could incorporate physics-based sound propagation models that account for distance attenuation, occlusion, and echo.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 3

[3] Renée Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development*, pages 47–83, 2002. 2

[4] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *arXiv preprint arXiv:1710.05512*, 2017. 2

[5] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018. 2

[6] Changan Chen, Puyuan Peng, Ami Baid, Zihui Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. Action2sound: Ambient-aware generation of action sounds from egocentric videos. In *European Conference on Computer Vision*, pages 277–295. Springer, 2024. 2

[7] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017. 2

[8] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields, 2024. 2

[9] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024. 2, 4, 8

[10] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020. 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. In *Conference on Robot Learning*, 2021. 2, 3, 6

[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. 3, 6

[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] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2, 6, 7

[15] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 4, 5, 6, 7, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 6

[17] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26406–26416, 2024. 2

[18] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Audio visual language maps for robot navigation. In *International Symposium on Experimental Robotics*, pages 105–117. Springer, 2023. 2

[19] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 3

[20] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Learning visual groups from co-occurrences in space and time, 2015. 2

[21] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023. 2, 3

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 8

[24] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. *Robotics: Science and Systems (RSS)*, 2023. 2

[25] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019. 6

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 4, 5, 6, 7

[27] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 2

[28] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10609–10618, 2019. 2

[29] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2871–2883, 2024. 8

[30] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023. 2

[31] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 3, 4

[32] 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*, pages 2405–2413, 2016. 2, 6

[33] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 2

[34] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting, 2024. 4, 5, 8

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 4, 5, 8

[36] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, 2014. 2

[37] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4

[38] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 4

[39] Laura Schulz. The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, 16(7):382–389, 2012. 2

[40] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3, 8

[41] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation, 2023. 2, 6, 7, 8

[42] Kun Su, Kaizhi Qian, Eli Shlizerman, Antonio Torralba, and Chuang Gan. Physics-driven diffusion models for impact sound synthesis from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9749–9759, 2023. 2, 3

[43] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models, 2023. 2

[44] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 2

[45] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161, 2024. 2

[46] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. 2

[47] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 3

[48] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 3

[49] 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. 3

[50] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision*, pages 182–199. Springer, 2022. 2