Objects as Audio-Visual Modal Sound Fields

Zisen Shao*, Zihao Wei*, Derong Jin, and Ruohan Gao

University of Maryland, College Park, USA
{zisen, zihaowei, djin77, rhgao}@umd.edu

Abstract. While modern 3D reconstruction excels at modeling object geometry and appearance, it largely ignores the rich acoustic cues revealed through physical interaction. Object impact sounds convey material, stiffness, and structural properties that complement vision, yet existing impact sound modeling approaches either rely on expensive physics-based simulation or require large datasets to generalize in a purely data-driven manner. We introduce Audio-Visual Modal Sound Field (AV-MSF), a novel object-level acoustic representation reconstructed from multi-view images and only a few impact sound recordings. AV-MSF builds on 3D Gaussian Splatting integrated with dense 3D visual feature to provide a strong geometry-aware prior, and represents the impact sound field using compact, physically meaningful modal parameters, enabling robust few-shot reconstruction. Experiments on two real-world datasets show that AV-MSF achieves state-of-the-art impact sound rendering, outperforming both physics-based and data-driven baselines. Furthermore, we demonstrate downstream applications enabled by our representation, including contact localization and object sound editing.

Keywords: Impact Sound · Audio-Visual · Multisensory Objects

1 Introduction

Our everyday lives are filled with a wide variety of physical objects. From a quick glimpse, we can often infer an object's shape, appearance, and semantic function. Yet a full understanding of an object goes beyond passive visual observation: it often requires active physical interaction, which generates vibrations and sound. For example, a transparent wine goblet may look nearly identical whether it is made of glass or plastic. A light tap, however, makes the difference immediately clear: glass produces a bright, ringing sound, while plastic sounds duller and more heavily damped. In daily life, we rely on such impact sounds to judge material properties, stiffness, and thickness—physical attributes that are often hard to infer reliably from appearance alone.

Despite the importance of modeling objects' acoustics, most existing approaches to representing or virtualizing objects remain overwhelmingly vision-centric. Modern 3D reconstruction pipelines can recover detailed geometry and appearance from images, enabling downstream applications such as novel view

* Equal contribution.

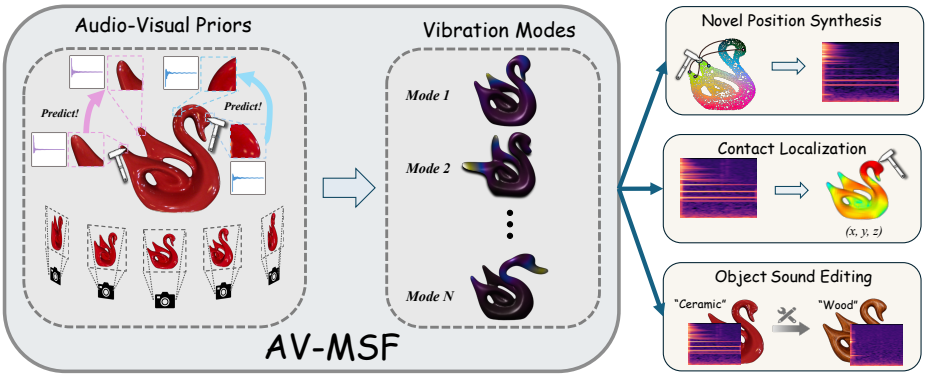

Fig. 1: ***Left:*** We reconstruct the Audio-Visual Modal Sound Field (AV-MSF) from multi-view RGB observations and only a few impact recordings, leveraging the insight that visually similar object regions exhibit similar vibration patterns. ***Right:*** AV-MSF enables diverse applications, including novel-position impact sound synthesis, contact localization, and object sound editing.

synthesis [20, 28], controllable relighting [10, 45, 46, 48], and text-driven 3D editing [34, 38, 44]. Extending these representations to jointly model photorealistic appearance *and* physically consistent impact sounds would unlock new capabilities, including more faithful and immersive contact feedback in VR, impact sound simulation for training multimodal embodied agents, and scalable creation of realistic multisensory object assets.

Existing approaches to modeling object impact sounds broadly fall into two categories: *physics-based* and *data-driven*. In engineering and graphics, physics-based rigid-body impact sound synthesis is often performed via linear modal analysis [5, 7, 29, 35]. While physically interpretable, these simulations are often computationally expensive and can be inaccurate due to coarse or uncertain material parameters. Recent differentiable inverse-rendering pipelines seek to recover these physical parameters from real recordings via gradient-based optimization [3, 19]; however, they either involve computationally heavy, error-prone optimization or adopt simplified modal parameterizations. In contrast, data-driven generative methods learn to synthesize impact sounds conditioned on multimodal cues [24, 40, 42]. Although they can produce diverse and perceptually plausible results, they are typically less physically grounded and often require substantial training data to generalize across objects and contact conditions.

To bridge this gap, our approach builds upon two key observations. First, impact sounds vary predictably across an object and are strongly correlated with local visual and geometric cues—how an object region looks often indicates how it will sound when struck. For example, as shown in Fig. 1, tapping the two symmetric wings of a swan sculpture produces nearly identical vibrations, and the impact sound produced at the swan's neck can be inferred from the sound at neighboring locations with similar curvature. These visual priors can potentially

help the model generalize even from only a few impact recordings. Second, an object's impact sound field can be distilled into a compact set of physically meaningful modal parameters. In particular, the model vibration frequencies are global properties determined by an object's shape and material, whereas each mode's excitation gain is position-dependent, capturing how strongly that mode is activated at different contact positions. This representation is both physically grounded and more sample-efficient to learn.

Building on these insights, we introduce the *Audio-Visual Modal Sound Field (AV-MSF)*, an object-level representation that reconstructs an object's modal impact-sound field from multi-view RGB observations and only a few impact recordings. AV-MSF first builds a 3D visual representation via 3D Gaussian Splatting (3DGS), and then lifts pre-trained visual features into a dense 3D feature field, providing a geometry-aware visual prior. Conditioned on this visual prior and sparse impact recordings, AV-MSF reconstructs an object modal sound field parameterized by physically meaningful vibration modes, with additional residual components capturing unmodeled environmental effects. Together, AV-MSF offers an end-to-end framework that learns efficiently from real-world audio-visual data, while remaining physically interpretable.

Experiments on two real-world datasets, ObjectFolder Real [12] and RealImpact [2], show that our approach significantly outperforms both physics-based and data-driven baselines. In the 20% training-data setting, we obtain a 2× improvement over both prior physics-based methods and data-driven baselines pretrained on large-scale simulated data. Moreover, our audio-visual, physics-based representation unlocks new downstream applications as illustrated in Fig. 1. Because the representation includes position-dependent parameters that capture how an object responds differently across contact locations, AV-MSF can infer the impact position from a previously unseen recording. Finally, by disentangling material-related acoustic parameters, our representation supports intuitive sound editing: we introduce an object sound editing pipeline that modifies an object's acoustic properties using Audio Score Distillation (ASD) [36].

Our key contributions are: (1) We present Audio-Visual Modal Sound Field (AV-MSF), a physics-based representation that reconstructs an object's modal sound field from multi-view RGB observations and a few impact sound recordings; (2) We show that AV-MSF substantially improves impact sound synthesis at novel contact locations over prior methods on two real-world datasets; (3) We demonstrate new downstream applications enabled by AV-MSF, including contact localization and object sound editing.

2 Related Work

Virtualizing 3D Objects. Object digitization has long been a central problem in computer vision. Scanning-based methods recover high-fidelity meshes via laser scanning and range sensing [6, 22]. In parallel, image-based reconstruction methods recover 3D representation from multi-view observations. While earlier systems [14,37] often produced coarse geometries, recent advances like NeRF [28]

and 3D Gaussian Splatting (3DGS) [20] enable photorealistic reconstruction. These representations have powered a wide range of downstream tasks, including novel-view synthesis [1, 20, 28], relighting [10, 45, 46, 48], geometry recovery [15, 17, 43], and controllable editing [34, 38, 44]. Despite the progress, most existing pipelines remain largely appearance-centric; in contrast, we aim to incorporate position-dependent impact sound into the object digitization process.

Multisensory Object Modeling. Beyond visual appearance, a growing body of work begins to model additional object-centric sensory modalities, particularly audio and touch. Earlier efforts focused on capturing multisensory object properties to build interactive physical assets [32]. More recently, audio and touch have been combined with vision to enable tasks such as 3D reconstruction [39, 41], cross-modal generation [9, 47], contact localization [21, 26], and robotic manipulation [16, 23]. Supporting these multisensory tasks, several new datasets have also been introduced. The ObjectFolder series [11, 13] builds datasets of implicitly represented neural objects that encode visual, acoustic, and tactile observations, whereas VibraVerse [33] explicitly models acoustics through modal analysis aligned with object geometry. However, both heavily rely on simulation and therefore suffer from sim-to-real gaps. Complementing these directions, ObjectFolder Real [12], RealImpact [2], and X-Capture [4] take important steps toward real-world multisensory capture, but their controlled acquisition setups limit scalability. Together, these limitations highlight the need for efficient reconstruction pipelines that can scale the creation of realistic multisensory assets.

Physics-based Impact Sound Synthesis. Traditional physics-based methods [5, 7, 29, 35] synthesize impact sounds by modeling the vibration modes of rigid objects. This is typically achieved through modal analysis, where sound is generated by solving the wave equation given a volumetric mesh and material properties such as Young's modulus, Poisson's ratio, and density. More recent differentiable frameworks [3,19] extend this paradigm toward inverse rendering of physically meaningful acoustic parameters from recorded audio. DiffSound [19] estimates physical and shape attributes through a differentiable pipeline. While physically interpretable, it is computationally expensive and susceptible to error accumulation due to its long optimization chain involving high-order finite element methods (FEM). In contrast, DiffImpact [3] improves efficiency by parameterizing impact sounds with modal quantities such as frequency, damping, and gain instead of explicitly optimizing material properties. Building on these works, we aim to reconstruct a physically interpretable modal sound field in a more efficient and practical regime. By incorporating visual priors, our method avoids costly FEM-based forward simulation over volumetric meshes while preserving the few-shot and physically grounded benefits of inverse approaches.

Data-driven Impact Sound Synthesis. Data-driven methods synthesize impact sounds by learning directly from real-world recordings with deep neural networks (often conditioned on multimodal cues). Video-to-sound approaches [8, 31, 40] can generate semantically plausible Foley effects, but do not model an explicit 3D acoustic representation. More recent works [25, 42] move toward 3D audio-visual modeling by embedding acoustics into 3D object or scene represen-

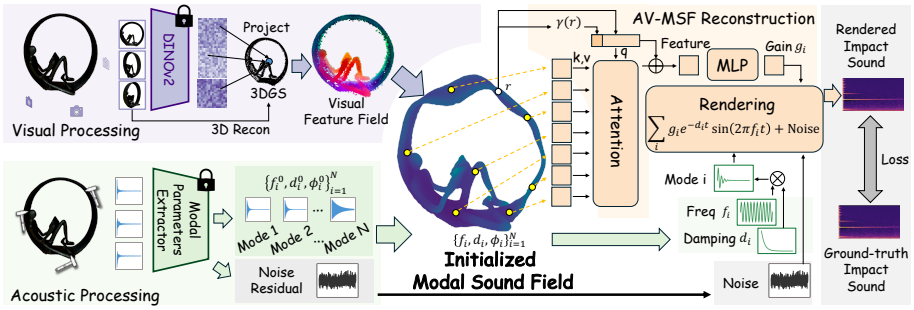

Fig. 2: Overview of AV-MSF. Given multi-view image observations and few-shot impact recordings, our framework reconstructs an audio-visual modal sound field. **(1) Visual Processing:** We extract dense features with a pre-trained vision encoder and lift them into a 3D Gaussian Splatting (3DGS) representation to form a geometry-aware visual feature field. **(2) Acoustic Processing:** We extract modal parameters from a few reference recordings to initialize optimization, and model unmodeled environmental noise with a residual component. **(3) AV-MSF Reconstruction:** We jointly optimize global modal frequencies and dampings, residual noise, and an implicit spatially varying neural gain field, guided by the visual feature field.

tations. However, these methods rely on fine-tuning pre-trained text-to-audio models, whose priors are often mismatched to transient, contact-driven impact sounds. Overall, purely data-driven pipelines tend to be task-specific and lack physical interpretability; as a result, they typically require large-scale training data—often difficult to obtain—to generalize reliably, which limits their applicability to diverse downstream applications.

3 Approach

We first review the background on modal sound synthesis in Sec. 3.1 and define our task and relevant notations in Sec. 3.2. Next, we describe the visual processing and acoustic processing components in Sec. 3.3 and Sec. 3.4, respectively. We then discuss our AV-MSF reconstruction pipeline in Sec. 3.5. Finally, we introduce two downstream applications enabled by our method in Sec. 3.6. Our method is illustrated in Fig. 2.

3.1 Background on Modal Sound Synthesis

We first briefly review *linear modal analysis* [18], the standard pipeline to synthesize rigid-body sounds from an object's physical properties and shape.

Elastic Vibration from Physical Parameters. Given an object with volumetric mesh and material properties, linear modal analysis models its small elastic vibration as:

$$\mathbf{M}\ddot{\mathbf{u}}(t) + \mathbf{D}\dot{\mathbf{u}}(t) + \mathbf{K}\mathbf{u}(t) = \mathbf{f}(t), \tag{1}$$

where $\mathbf{u}(t)$ denotes the nodal displacement, $\mathbf{M}$ and $\mathbf{K}$ are the mass and stiffness matrices, $\mathbf{D} = \alpha\mathbf{M} + \beta\mathbf{K}$ is the Rayleigh damping, and $\mathbf{f}(t)$ is the external force. Here, $\mathbf{M}$ depends on the object's shape and density ρ, while $\mathbf{K}$ depends on its shape and material parameters (Young's modulus E and Poisson's ratio ν).

Modal Decomposition and Sound Synthesis. By generalized eigenanalysis $\mathbf{KU} = \mathbf{MUS}$, Eq. (1) can be reformulated as:

$$\ddot{\mathbf{q}}(t) + (\alpha\mathbf{I} + \beta\mathbf{S})\dot{\mathbf{q}}(t) + \mathbf{S}\mathbf{q}(t) = \mathbf{U}^\top \mathbf{f}(t), \tag{2}$$

where $\mathbf{U}$ is the matrix of vibration modes, $\mathbf{S}$ is the diagonal matrix of modal eigenvalues, and $\mathbf{q}(t)$ satisfies $\mathbf{u}(t) = \mathbf{U}\mathbf{q}(t)$, with each entry describing the response of one mode over time. The resulting impact sound is modeled as the sum of all decaying modal vibrations:

$$s(t) = \sum_{i=1}^{N} g_i e^{-d_i t} \sin(2\pi f_i t), \tag{3}$$

where f_i, d_i, g_i are the frequency, damping, gain of the i-th mode, respectively.

Assumptions and Limitations. Notably, our approach relies on two main assumptions. First, we assume the modal frequency f_i and damping d_i are intrinsic, position-invariant properties of the object, while only the modal gain g_i varies with contact location. Second, the Rayleigh damping model of linear modal analysis is a numerical approximation of real physical behaviors [18], which do not hold for all real-world objects. As we observe in real datasets, the damping can vary across contact locations for some objects. We therefore make damping optionally learnable as a spatially varying parameter. For simplicity, however, we denote only the gain as a spatially varying parameter in our formulation below.

3.2 Problem Formulation

We consider the problem of reconstructing an object's Audio-Visual Modal Sound Field (AV-MSF) from visual observations and few-shot impact sound recordings. Specifically, for a single object, we are given a set of calibrated multi-view RGB images $\mathcal{I} = \{(I_i, \Pi_i)\}_{i=1}^{I}$, where I_i denotes the i-th image and Π_i denotes the known camera parameters, together with few-shot impact sound recordings $\mathcal{S} = \{(\mathbf{x}_j, s_j(t))\}_{j=1}^{S}$, where $\mathbf{x}_j \in \mathbb{R}^3$ is the 3D contact location on the object surface and $s_j(t)$ is the corresponding audio waveform.

Our goal is to learn an object-level representation that can render the impact sound at an arbitrary novel contact location $\mathbf{x}$ on the surface. Following Eq. (3), we model the impact sound generated at $\mathbf{x}$ as a combination of decaying modal vibrations (assuming the object is initially static) and a non-modal residual component:

$$s(\mathbf{x}, t) = \sum_{i=1}^{N} g_i(\mathbf{x})\, e^{-d_i t} \sin(2\pi f_i t) + \sum_{i=1}^{F} \epsilon(m_i, t), \tag{4}$$

where N is the number of modes, $\{f_i\}_{i=1}^{N}$ and $\{d_i\}_{i=1}^{N}$ are the modal frequencies and dampings, $g_i(\mathbf{x})$ is the excitation gain of the i-th mode at contact location $\mathbf{x}$, and $\epsilon(m_i, t)$ represents residual noise parameterized by a learnable frequency-domain magnitude $\{m_i\}_{i=1}^{F}$ over F frequency bins.

As noted in Sec. 3.1, $\{f_i\}_{i=1}^{N}$ and $\{d_i\}_{i=1}^{N}$ are object-intrinsic global parameters, while only the modal gains vary spatially across contact locations. We therefore represent the object's modal sound field as $\mathcal{F} = \Big(\{f_i, d_i\}_{i=1}^{N},\ \{m_i\}_{i=1}^{F},\ \mathcal{G}_\theta(\mathbf{x})\Big)$, where $\mathcal{G}_\theta(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_N(\mathbf{x})]^\top$ is an implicit spatial gain field that predicts modal gains for any queried surface point $\mathbf{x}$.

Overall, the reconstruction problem can thus be formulated as a mapping $\Phi : (\mathcal{I}, \mathcal{S}) \mapsto \mathcal{F}$, which estimates (i) a set of global modal parameters $\{f_i, d_i\}_{i=1}^{N}$ and noise residual magnitudes $\{m_i\}_{i=1}^{F}$, and (ii) an implicit neural gain field $\mathcal{G}_\theta(\mathbf{x})$ that predicts the modal gains for any queried surface location $\mathbf{x}$.

3.3 Multi-View Visual Feature Extraction and Refinement

To effectively guide acoustic prediction of impact sound, we construct a geometry-aware 3D feature field from multi-view RGB images. We first reconstruct a 3D Gaussian Splatting (3DGS) representation, yielding a dense point cloud of Gaussian centers $\mathcal{P} = \{\mathbf{o}_i\}_{i=1}^{O}$. We then extract 2D visual embeddings using pre-trained DINOv2 [30] and lift them into 3D, assigning each Gaussian spatial center $\mathbf{o}_i$ a feature vector $\mathbf{f}_i \in \mathbb{R}^D$.

However, due to viewpoint variations, these lifted features may fail to respect the object's intrinsic geometric symmetries. For example, points lying on the same rotational orbit should ideally share similar semantics. To address this issue and enforce geometric consistency, we automatically detect object symmetries, such as rotational or planar mirror symmetries, by evaluating the geometric alignment error of candidate transformations over the point cloud. Given the detected symmetries, we perform symmetry-aware feature alignment. Specifically, we aggregate and average the semantic features across their corresponding symmetric counterparts through orbit or reflection pooling, explicitly enforcing geometric consistency and producing final refined feature $\mathbf{f}_i^{\text{refined}}$ for each Gaussian. For simplicity, we use $\mathbf{f}_i$ in the following. See Supp. for details on the feature refinement procedures.

3.4 Modal Sound Parameters Extraction and Initialization

Directly optimizing $\mathcal{F}$ from few-shot recordings $\mathcal{S}$ is highly non-convex and prone to local minima. To stabilize training, we leverage the physical prior that the modal frequencies $\{f_i\}_{i=1}^{N}$ and dampings $\{d_i\}_{i=1}^{N}$ are intrinsic object properties, explicitly extract them from $\mathcal{S}$ for initialization.

Modal Parameters Extraction. For each recorded impact sound $s_j(t) \in \mathcal{S}$, we compute its Short-Time Fourier Transform (STFT) and apply a per-frequency inverse-STFT to decompose the audio into a set of narrow-band time-domain signals, $\{s_{j,n}(t)\}_{n=1}^{F}$, where F is the total number of frequency bins. We explicitly

filter out noise by discarding peaks that occur beyond a time threshold or are not significantly damped. For each candidate mode signal $s_{j,n}(t)$, we estimate its damping d_n and gain g_n via log-linear regression:

$$\log |\mathcal{H}\{s_{j,n}(t)\}| = \log(g_n) - d_n \cdot t, \tag{5}$$

where $\mathcal{H}$ denotes the Hilbert transform. To obtain the global object-level parameters, we identify the frequencies that consistently appear across recordings. For the i-th selected mode, the global damping coefficient d_i is computed as the average of the estimated dampings across all recordings. These globally extracted $\{f_i, d_i\}_{i=1}^{N}$ serve as the shared intrinsic parameters for the modal sound field.

Residual Noise Modeling. Real-world sound recordings often contain non-modal effects. To stabilize modal parameter optimization, we model a residual component to capture (1) low-frequency background noise and (2) other unpredictable non-modal factors, such as variations in contact force and microphone distance. We model the residual component as static filtered noise with learnable per-band magnitudes $\mathbf{m} = \{m_i\}_{i=1}^{F}$ following prior work [3]. We initialize $\{m_i\}_{i=1}^{F}$ from $\mathcal{S}$ by averaging the lowest-energy temporal segments within each frequency bin. At synthesis time, we draw white noise $\epsilon(t) \sim \mathcal{N}(0, 1)$ and shape it with a differentiable frequency filter:

$$\epsilon(\mathbf{m}, t) = \sum_{i=1}^{F} m_i \left(b_i * \epsilon(t)\right), \tag{6}$$

where b_i denotes the band-pass filter for the i-th frequency bin. The final waveform is obtained by adding this filtered residual to the modal reconstruction.

3.5 Reconstructing Audio-Visual Modal Sound Field

Up to this point, we have obtained a geometry-aware 3DGS representation and initialized with extracted acoustic parameters $\{f_i, d_i\}_{i=1}^{N}$ and $\{m_i\}_{i=1}^{F}$ (middle of Fig. 2). Our goal is to leverage the visual prior to optimize $\mathcal{F}$ through the differentiable modal sound synthesizer in Eq. (4).

For each object, we cluster the Gaussian centers $\mathcal{P} = \{\mathbf{o}_i\}_{i=1}^{O}$ into K spatial groups using 3D Euclidean distance, with $\mathbf{c}_k$ denoting the center of cluster k. Within each cluster, we average the symmetry-aligned features to form a single region-level descriptor:

$$\bar{\mathbf{f}}_k \;=\; \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{f}_i, \tag{7}$$

where $\mathcal{C}_k$ denotes the set of Gaussians assigned to cluster k. The resulting descriptors $\{\bar{\mathbf{f}}_k\}_{k=1}^{K}$ provide a compact set of visual priors.

Predicting Location-dependent Gains. Given a queried impact location $\mathbf{x}$, we first find its nearest Gaussian center $i^\star = \arg\min_i \|\mathbf{x} - \mathbf{x}_i\|_2$, and use its local visual feature $\mathbf{f}_{i^\star}$ as a fine-grained cue. We also encode geometric context by computing the relative offsets from the query point to all cluster centers.

Specifically, $r_k = \mathbf{x}_{i^\star} - \mathbf{c}_k$. and we stack the offsets to all K clusters into $\mathbf{r} = [r_1, \ldots, r_K]$.

We then predict the modal gains $\mathcal{G}_\theta(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_N(\mathbf{x})]^\top$ by conditioning on (i) the local feature $\mathbf{f}_{i^\star}$, (ii) the offset encoding $\mathrm{PE}(\mathbf{r})$, and (iii) global context aggregated from all K cluster descriptors via attention. Specifically, we follow NeRF [28] and use sin-cos function as our $\mathrm{PE}(\cdot)$ function. Concretely, we form a query vector

$$\mathbf{q} = W_q\,[\mathrm{PE}(\mathbf{r}); \mathbf{f}_{i^\star}], \tag{8}$$

and define keys/values from the region descriptors

$$\mathbf{k}_k = W_k \bar{\mathbf{f}}_k, \qquad \mathbf{v}_k = W_v \bar{\mathbf{f}}_k. \tag{9}$$

Dot-product attention produces a context feature

$$\alpha_k = \mathrm{softmax}_k\Big(\frac{\mathbf{k}_k^\top \mathbf{q}}{\sqrt{d}}\Big), \qquad \mathbf{z} = \sum_{k=1}^{K} \alpha_k\, \mathbf{v}_k. \tag{10}$$

Finally, the context feature is projected and concatenated with the local Gaussian feature to predict the modal gains:

$$\mathcal{G}_\theta(\mathbf{x}) = \mathrm{MLP}_\theta(\mathbf{z}; \mathbf{f}_{i^\star}). \tag{11}$$

Differentiable Synthesis and Optimization. For each observed impact recording $(\mathbf{x}_j, s_j(t)) \in \mathcal{S}$, we synthesize

$$\hat{s}_j(t) = \sum_{i=1}^{N} \mathcal{G}_\theta(\mathbf{x}_j)\, e^{-d_i t} \sin(2\pi f_i t) \;+\; \epsilon(\mathbf{m}, t), \tag{12}$$

where $\epsilon(\mathbf{m}, t)$ is the filtered-noise residual from Sec. 3.4. We optimize θ jointly with the global parameters $\{f_i, d_i\}_{i=1}^N$ and $\mathbf{m}$ by minimizing a reconstruction loss between $\hat{s}_j(t)$ and $s_j(t)$ over few-shot contacts in $\mathcal{S}$.

Training Objectives. We train each object using a two-stage optimization strategy. In the warm-up stage, we optimize $\mathcal{G}_\theta$ using the gain field extracted in Sec. 3.4. For each audio sample i, we predict the gain $\mathcal{G}_\theta(x_i)$ and minimize the mean-squared error (MSE) with respect to the extracted target gain $\mathbf{g}_i$:

$$\mathcal{L}_{\mathrm{warmup}} = \frac{1}{S} \sum_{i=1}^{S} \|\mathcal{G}_\theta(\mathbf{x}_i) - \mathbf{g}_i\|_2^2 . \tag{13}$$

This warm-up anchors the gain field in a well-conditioned optimization regime, reducing the risk of collapse and accelerating convergence.

In the training stage, we train all parameters end-to-end with a multi-scale STFT reconstruction loss. We apply it to both the full waveform and the early-time segment of length t, where the impact energy is most concentrated and the signal is less affected by environmental noise. The training loss is defined as:

$$\mathcal{L}_{\mathrm{spec}} = \sum_r \left(\|\mathrm{STFT}_r(s_i) - \mathrm{STFT}_r(\hat{s}_i)\|_1 + w_c \, \|\mathrm{STFT}_r(s_i[:t]) - \mathrm{STFT}_r(\hat{s}_i[:t])\|_1 \right), \tag{14}$$

where r indexes FFT resolutions and w_c is the weight for the early cut-off loss.

3.6 Applications

Our physics-based audio-visual modal sound field can potentially support a range of downstream tasks by leveraging the distinct physical roles of its components. We showcase two below: contact localization and object sound editing.

Contact Localization. In this task, we aim to infer the 3D impact location $\mathbf{x}^*$ from a novel sound recording $s_{new}(t)$. Relying on the object-intrinsic global parameters $\{f_i, d_i\}_{i=1}^N$ obtained during training, we first extract the corresponding mode-specific gains $\hat{\mathbf{g}} = [\hat{g}_1, \ldots, \hat{g}_N]^\top$ from $s_{new}(t)$. Because our learned spatial gain field $\mathcal{G}_\theta(\mathbf{x})$ uniquely characterizes the acoustic response at any point on the object, we can localize the impact by finding the surface coordinate $\mathbf{x}$ that minimizes the discrepancy between the predicted and extracted gains:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathcal{D}\left(\mathcal{G}_\theta(\mathbf{x}), \hat{\mathbf{g}}\right), \tag{15}$$

where $\mathcal{D}$ denotes a distance metric. In practice, we use cosine distance to ensure the localization is invariant to the absolute magnitude of the initial impact force.

Object Sound Editing. To enable semantic, text-driven editing of the impact sound, we adopt Audio-SDS [36]. Given a target text prompt p, we optimize the object-intrinsic modal parameters $\theta = \{f_i, d_i, g_i\}_{i=1}^N$ by rendering the modal audio $s(\theta, t)$ and distilling guidance from a frozen, pretrained text-to-audio diffusion model. A significant challenge arises from the fact that modal frequencies are sparse, in which independent optimization of $\{f_i\}_{i=1}^N$ often struggles to overcome the large spectral gap between contrastive materials (e.g., shifting from wood to metal). To facilitate convergence, we propose a hierarchical frequency parameterization: $f_i = \exp(\log \alpha + \log f_{i,\text{init}} + \Delta f_i)$. Here, α serves as a global pitch-scaling factor that captures the primary material-dependent shift, while Δf_i allows for per-mode residual refinement.

To circumvent optimization instabilities, we employ Decoder-SDS combined with multi-step denoising. Specifically, we map the rendered audio to the latent space, add noise, perform partial DDIM sampling steps conditioned on p, and decode the result back to the audio domain to form a target pseudo-ground-truth waveform $\hat{s}(t)$. The parameters θ are updated by minimizing a multi-resolution STFT loss between the rendered $s(\theta, t)$ and the distilled target $\hat{s}(t)$.

Physically Grounded Sound Editing. While Audio-SDS successfully modifies the acoustic parameters $\{f_i, d_i\}_{i=1}^N$ and reference gains $\{g_i\}_{i=1}^N$ to match a text prompt, we must ensure these edits remain consistent with linear modal analysis to preserve the object's acoustic physics. As detailed in Supp., altering Young's modulus, density, and Rayleigh damping coefficients scales the modal frequencies and decay rates but leaves the spatial mode shapes unchanged. Consequently, after editing material parameters, the modal gains update via a simple per-mode rescaling $a_k^{(2)}(x) = a_k^{(1)}(x)\frac{f_k^{(1)}}{f_k^{(2)}}$. Because the frequency-dependent factors cancel out, the gain ratio between any two impact locations remains invariant to such parameter editing. This physical property allows us to seamlessly preserve the learned spatial modal field and perform physically grounded sound editing by only updating the per-mode frequencies, decay rates, and gain scales.

4 Experiment

We now validate AV-MSF on novel position object impact sound rendering and two new downstream applications, and compare it against prior approaches and baseline methods.

4.1 Datasets

We evaluate our method on ObjectFolder Real [12] and RealImpact [2], two real-world multisensory datasets that include impact sound recordings. For both datasets, we normalize the impact sound recordings using the provided force profiles, zero-align the peak signals, and truncate them to 3 seconds to preserve the full damping behavior. Dataset-specific details are provided below.
ObjectFolder Real [12] contains multisensory measurements of 100 real-world objects across seven material types, including 3D meshes, videos, impact sounds, and tactile readings. Each object is associated with 30–50 impact sounds manually recorded across its surface using a muted, force-sensing impact hammer. By default, 20% of the impact recordings for each object are selected via farthest-point sampling for training, with the rest used for evaluation. We use the dataset for the main quantitative comparison of novel-position impact sound synthesis, and its subset for ablations and downstream evaluations.
RealImpact [2] contains 150,000 impact sound recordings from 50 everyday objects. Because it was designed to study spatial acoustics of impact sounds, each object has only 5 distinct impact locations, each struck by an automatically controlled, calibrated impact hammer mounted on a rotational gantry system, and recorded 600 times from different microphone positions. In our experiments, we use only the microphone closest to the object, resulting in 5 recordings per object. Due to its limited number of impact locations, we adopt a leave-one-out cross-validation protocol, iteratively evaluating on one impact location while training on the remaining ones.

4.2 Baselines and Evaluation Metrics

We compare to existing physics-based and data driven methods [19, 42] and multiple baselines on novel-position impact sound rendering:
White Noise: Random noise scaled to match the average loudness of the impact recordings used to train our method.
Random Impact: We use a randomly selected impact sound from another object as the predicted impact sound for the queried location and object.
KNN: For each query contact location, we retrieve the K nearest training impact sounds and average them to produce the rendered impact sound for the queried location. We use K=3 in all of our settings.
DiffSound [19] is a state-of-the-art physics-based method that estimates material parameters via inverse rendering. Position-specific impact sounds can be simulated from the object's vibration modes using the inferred parameters. We further add a learned noise residual from our method for fairness.

Table 1: Quantitative comparison on novel-position impact sound rendering. We compare all the baselines on two popular real-world impact sound datasets, ObjectFolder Real [12] and RealImpact [2]. Lower is better for all metrics.

	ObjectFolder Real				RealImpact			
Method	L1	L1 Log	ENV	CDPAM	L1	L1 Log	ENV	CDPAM
White Noise	3.774	6.859	0.305	1.35e-3	3.789	7.258	0.308	1.23e-3
Random Impact	0.035	1.367	0.019	2.47e-4	0.039	1.582	0.022	2.71e-4
KNN	0.014	**0.930**	0.014	1.53e-4	0.026	1.036	0.017	2.25e-4
DiffSound [19]	0.031	1.298	0.030	2.53e-4	0.029	1.533	0.024	2.39e-4
SonicGauss [42]	0.033	1.281	0.018	2.01e-4	0.039	1.580	0.025	2.43e-4
Ours	**0.013**	0.951	**0.014**	**1.35e-4**	**0.021**	**0.996**	**0.017**	**2.16e-4**

SonicGauss [42] is a state-of-the-art data-driven method that trains an impact sound diffusion model conditioned on contact position and encoded 3D Gaussian object features. We finetune each object from its Stage-2 checkpoint on ObjectFolder 2.0, with improved force and audio normalization.

Evaluation Metrics. We compare with baselines using following metrics: (1) *L1 Distance*, which computes the Euclidean distance between the ground-truth and predicted spectrograms; (2) *L1-log Distance*, which is the same as L1 Distance but measuring the distance between log-mel spectrograms for better perceptual alignment; (3) *Envelope Distance*, which computes the Euclidean distance between the amplitude envelopes of the ground-truth and the predicted waveforms; and (4) *CDPAM* [27], a perceptual audio similarity metric; (5) *Relative Mean Euclidean Distance (RMED)*, distance between the predicted and ground-truth coordinates normalized by the bounding-box diagonal for contact localization.

4.3 Novel Position Impact Sound Rendering

We report novel position impact sound rendering results in Tab. 1. AV-MSF substantially outperforms prior physics-based and data-driven baselines (DiffSound and SonicGauss). DiffSound underperforms because it heavily relies on inverse rendering to estimate material parameters from real recordings, which is often error-prone and leads to inaccurate rendered sound. SonicGauss is highly data-hungry: even with pretraining on ObjectFolder 2.0 [13], few-shot finetuning remains ineffective. Moreover, as a generative model, it tends to introduce hallucinated or distorted patterns, indicating difficulty in faithfully preserving the underlying physical properties. KNN is a strong baseline because many objects are symmetric (see object visualizations in Supp.), and nearby points often share similar spectra. Nonetheless, AV-MSF achieves better results on most metrics, particularly when local geometry or appearance changes sharply.

Case Study on Non-symmetric Objects. The evaluation datasets are mainly composed by symmetric objects (*e.g.*, bowls, plates), which make KNN a strong base-

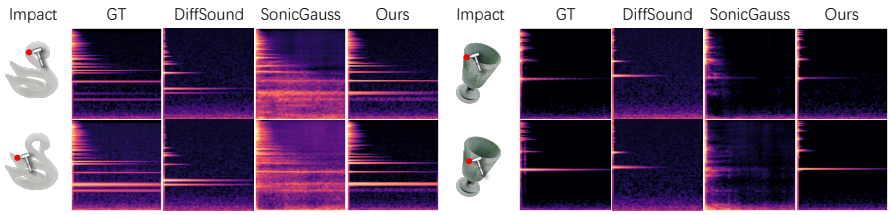

Fig. 3: Qualitative Spectrogram Comparisons. We visualize the rendered impact sounds from DiffSound [19], SonicGauss [42], and our method, alongside the ground truth. All spectrograms are generated on the settings of 20–20k Hz, the first second, scaled to $(-1, 1)$. See Supp. for more examples.

Table 2: Ablation Study Results.

Method	L1	L1 Log	ENV	CDPAM
w/o visual	0.028	1.148	0.015	4.20e-4
w/o init	0.045	1.077	0.026	2.97e-4
w/o align	0.019	1.002	0.016	2.62e-4
w/o residual	0.031	5.764	0.026	3.22e-4
ours	**0.019**	**0.927**	**0.015**	**2.00e-4**

Table 3: Case Study Results on Non-symmetric Objects.

Method	L1	L1 Log	ENV	CDPAM
KNN	0.0135	0.9724	0.0138	2.15e-4
Ours	0.0110	0.9259	0.0131	1.53e-4

line. To further demonstrate the strength of our method, we evaluate on a subset of less symmetric objects from ObjectFolder Real (see Supp. for more details). As shown in Tab. 3, our method consistently outperforms the KNN baseline on this non-symmetric subset, with clear improvements across all metrics.

Qualitative Comparison. Figure 3 presents qualitative comparisons on two objects, visualizing the 20–20k Hz spectrograms of the first second for two different impact locations per object. DiffSound produces clear and sharp structures, but the error in its physical parameter estimation can lead to incorrectly placed frequency bands or dampings. SonicGauss recovers the overall frequency content more roughly correctly, yet it can introduce hallucinated or distorted patterns due to its generative nature. In contrast, our method yields clearer, more faithful spectral structures while remaining strongly position-aware, producing distinct responses across impact locations that better match the ground truth.

Ablation Study. We conduct an ablation study to assess the contributions of four key components in our method. The results, summarized in Tab. 2, confirm the effectiveness of each component. (1) *Ablate on Appearance*, removing the DINO features, which provide local appearance, consistently degrades performance, with the largest drop on perceptual quality. We attribute this to DINO encoding fine-grained and semantically meaningful local details, such as curvature, edges, and texture. In contrast, 3D point coordinates mainly capture coarse geometry and therefore provide a weaker signal for accurate sound rendering. (2) *Ablate*

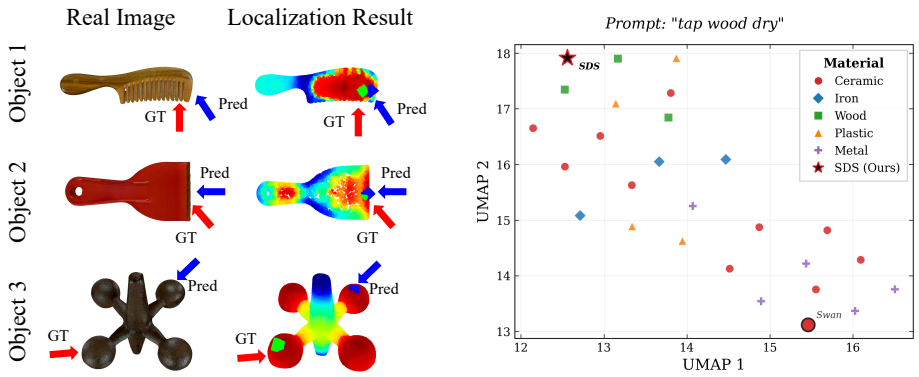

Fig. 4: Examples for Downstream Applications. Left: Contact localization heatmaps. Red arrows indicate novel impact locations, and blue arrows indicate predictions. Object 3 shows a failure case. **Right**: Visualization of frequency distributions of modal parameters before and after sound editing.

Table 4: Contact Localization.

Method	RMED ↓
DiffSound [19]	41.78%
Ours	**34.61%**

Table 5: Sound Editing.

Method	UMAP ↓
Generation	3.077
Audio-SDS	4.234
Ours	**2.753**

on Modal Parameter Initialization Warm-up, we further verify the importance of the modal parameter initialization, which yields a significant performance gain. This warm-up stage stabilizes optimization by bringing modal parameters into a physically reasonable region. Without it, training can collapse, as the model may become stuck in a poor local minimum. (3) *Ablate on Feature Alignment*, directly lifting DINO features into 3D can introduce misalignment with the geometry prior due to lighting and texture variations; for instance, symmetric regions may receive inconsistent features across two sides. Our feature-alignment module mitigates this issue by enforcing geometric consistency, improving performance across all metrics. (4) *Ablation on Residual Component*, removing the residual component yields the worst L1 Log performance, highlighting its importance for capturing low-frequency environmental noise. Moreover, removing it worsens the contact localization error from 38.4% to 43.8% on the ablation subset, suggesting that the gain field is otherwise sacrificed to fit non-modal effects.

4.4 Results on Downstream Applications

We demonstrate the utility of AV-MSF through two downstream applications, validating that it captures objects' spatial acoustic properties and supports impact-sound-driven tasks such as contact localization and object sound editing.

Contact Localization. We evaluate the ability of AV-MSF to infer the 3D impact position $\mathbf{x}^*$ from a novel impact sound recording. We use DiffSound [19] as our primary baseline, as it also employs a modal-based optimization approach. The mode-specific gains are extracted from the audio and matched against our learned spatial gain field $\mathcal{G}_\theta(\mathbf{x})$ using cosine distance. Notably, real-world contact localization from impact sounds is challenging due to the environmental noise, uniform material property, and non-modal effect introduced during data collection process. It is especially ambiguous for highly symmetric and small objects. Therefore, we report the experiment results on the same subset of non-symmetric objects used in Tab. 3. As shown in Tab. 4, our method substantially outperforms DiffSound, which struggles to localize contacts accurately when its estimated physical parameters are incorrect. We also show heatmap visualizations in Fig. 4. Object 1 and 2 exhibit accurate spatial distributions that align well with the ground-truth contact locations, while Object 3 illustrates a failure case caused by high symmetry. See more examples in Supp.

Object Sound Editing. We evaluate the effectiveness of our Audio-SDS framework for text-driven material editing, with an emphasis on the semantic alignment with the editing prompt. We compare our hierarchical optimization scheme against text-to-audio generation and Audio-SDS [36]. As reported in Tab. 5, our method achieves the lowest UMAP distance, significantly outperforming both the Generation baseline and Audio-SDS (see Supp. for the metric definition and setup). This demonstrates that our hierarchical frequency optimization scheme accurately shifts the acoustic profile to match the target material. In contrast, unconstrained Audio-SDS independently updates individual modal parameters, failing to achieve effective frequency editing across contrastive material pairs. Furthermore, in Fig. 4, we visualize an editing example—transforming a ceramic object into a contrasting wooden object—by extracting frequency distributions with our modal parameter extractor (Sec. 3.4) and embedding the impact sounds before and after editing into a low-dimensional space using UMAP. We further contextualize this result by overlaying it on the UMAP embedding of 26 objects spanning diverse material properties. We observe that our method can successfully edit the material type of the object. See Supp. for more examples.

5 Conclusion

We presented AV-MSF, an audio-visual modal sound representation for object impact sound synthesis. By leveraging visual cues from multi-view image observations and physical parameters estimated from a few impact sound recordings, AV-MSF predicts impact sounds significantly more accurately than prior methods. Beyond impact sound synthesis, our representation also enables new downstream applications involving object impact sounds, including contact localization and object sound editing. Notably, existing real-world object impact sound datasets contain only objects with uniform materials. As a result, we focused on modeling impact sounds under this setting, while leaving generalization to non-uniform materials as an interesting direction for future work.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
2. Clarke, S., Gao, R., Wang, M., Rau, M., Xu, J., Wang, J.H., James, D.L., Wu, J.: Realimpact: A dataset of impact sound fields for real objects. In: CVPR (2023)
3. Clarke, S., Heravi, N., Rau, M., Gao, R., Wu, J., James, D., Bohg, J.: Diffimpact: Differentiable rendering and identification of impact sounds. In: CoRL (2021)
4. Clarke, S., Wistreich, S., Ze, Y., Wu, J.: X-capture: An open-source portable device for multi-sensory learning. In: ICCV (2025)
5. Corbett, R., van den Doel, K., Lloyd, J.E., Heidrich, W.: Timbrefields: 3d interactive sound models for real-time audio. Presence (2007)
6. Curless, B.: From range scans to 3d models. SIGGRAPH (1999)
7. van den Doel, K., Kry, P.G., Pai, D.K.: Foleyautomatic: physically-based sound effects for interactive simulation and animation. In: SIGGRAPH (2001)
8. Dou, Y., Oh, W., Luo, Y., Loquercio, A., Owens, A.: Hearing hands: Generating sounds from physical interactions in 3d scenes. In: CVPR (2025)
9. Dou, Y., Yang, F., Liu, Y., Loquercio, A., Owens, A.: Tactile-augmented radiance fields. In: CVPR (2024)
10. Fan, J., Luan, F., Yang, J., Hasan, M., Wang, B.: Rng: Relightable neural gaussians. CVPR (2025)
11. Gao, R., Chang, Y.Y., Mall, S., Fei-Fei, L., Wu, J.: Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In: CoRL (2021)
12. Gao, R., Dou, Y., Li, H., Agarwal, T., Bohg, J., Li, Y., Fei-Fei, L., Wu, J.: The objectfolder benchmark: Multisensory learning with neural and real objects. In: CVPR (2023)
13. Gao, R., Si, Z., Chang, Y.Y., Clarke, S., Bohg, J., Fei-Fei, L., Yuan, W., Wu, J.: Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In: CVPR (2022)
14. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: ICCV (2007)
15. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In: CVPR (2024)
16. Higuera, C., Sharma, A., Fan, T., Bodduluri, C.K., Boots, B., Kaess, M., Lambeta, M., Wu, T., Liu, Z., Hogan, F.R., et al.: Tactile beyond pixels: Multisensory touch representations for robot manipulation. In: CoRL (2025)
17. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH (2024)
18. James, D.L., Langlois, T.R., Mehra, R., Zheng, C.: Physically based sound for computer animation and virtual environments. In: ACM SIGGRAPH 2016 Courses (2016)
19. Jin, X., Xu, C., Gao, R., Wu, J., Wang, G., Li, S.: Diffsound: Differentiable modal sound rendering and inverse rendering for diverse inference tasks. In: SIGGRAPH (2024)
20. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG (2023)
21. Lee, M., Yoo, U., Oh, J., Ichnowski, J., Kantor, G., Kroemer, O.: Sonicboom: Contact localization using array of microphones. RAL (2025)

22. Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., Shade, J., Fulk, D.: The digital michelangelo project: 3d scanning of large statues. In: SIGGRAPH (2000)
23. Li, H., Zhang, Y., Zhu, J., Wang, S., Lee, A.M., Xu, H., Adelson, E., Li, F.F., Gao, R., Wu, J.: See, hear, and feel: Smart sensory fusion for robotic manipulation. In: CoRL (2022)
24. Li, T., Huang, B., Zhuang, X., Jia, D., Chen, J., Wang, Y., Chen, Z., Anumanchipalli, G., Wang, Y.: Sounding that object: Interactive object-aware image to audio generation. In: ICML (2025)
25. Li, Y., Kim, H., Zhan, F., Qiu, R.Z., Ji, M., Shan, X., Zou, X., Liang, P., Pfister, H., Wang, X.: Visual acoustic fields (2025)
26. Luo, S., Mou, W., Althoefer, K., Liu, H.: Localizing the object contact through matching tactile features with visual map. In: ICRA (2015)
27. Manocha, P., Jin, Z., Zhang, R., Finkelstein, A.: Cdpam: Contrastive learning for perceptual audio similarity. In: ICASSP (2021)
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
29. O'Brien, J.F., Shen, C., Gatchalian, C.M.: Synthesizing sounds from rigid-body simulations. In: ACM SIGGRAPH/Eurographics Symposium on Computer Animation (2002)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
31. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
32. Pai, D.K., Doel, K.v.d., James, D.L., Lang, J., Lloyd, J.E., Richmond, J.L., Yau, S.H.: Scanning physical interaction behavior of 3d objects. In: SIGGRAPH (2001)
33. Pang, B., Xu, C., Ren, J., Wang, G., Li, S.: Vibraverse: A large-scale geometry-acoustics alignment dataset for physically-consistent multimodal learning (2025)
34. Qi, Z., Yang, Y., Zhang, M., Xing, L., Wu, X., Wu, T., Lin, D., Liu, X., Wang, J., Zhao, H.: Tailor3d: Customized 3d assets editing and generation with dual-side images (2024)
35. Ren, Z., Yeh, H., Lin, M.C.: Example-guided physically based modal sound synthesis. TOG (2013)
36. Richter-Powell, J., Torralba, A., Lorraine, J.: Score distillation sampling for audio: Source separation, synthesis, and beyond. arXiv preprint arXiv:2505.04621 (2025)
37. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
38. Sella, E., Fiebelman, G., Hedman, P., Averbuch-Elor, H.: Vox-e: Text-guided voxel editing of 3d objects. In: ICCV (2023)
39. Smith, E.J., Calandra, R., Romero, A., Gkioxari, G., Meger, D., Malik, J., Drozdzal, M.: 3D shape reconstruction from vision and touch. In: NeurIPS (2020)
40. Su, K., Qian, K., Shlizerman, E., Torralba, A., Gan, C.: Physics-driven diffusion models for impact sound synthesis from videos. In: CVPR (2023)
41. Suresh, S., Si, Z., Mangelson, J.G., Yuan, W., Kaess, M.: ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In: ICRA (2022)
42. Wang, C., Li, H., Luo, Y.: Sonicgauss: Position-aware physical sound synthesis for 3d gaussian representations. In: ACMMM (2025)

43. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: NeurIPS (2021)
44. Ye, J., Xie, S., Zhao, R., Wang, Z., Yan, H., Zu, W., Ma, L., Zhu, J.: NANO3d: A training-free approach for efficient 3d editing without masks. In: ICLR (2026)
45. Yu, H.X., Guo, M., Fathi, A., Chang, Y.Y., Chan, E.R., Gao, R., Funkhouser, T., Wu, J.: Learning object-centric neural scattering functions for free-viewpoint relighting and scene composition. TMLR (2023)
46. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: neural factorization of shape and reflectance under an unknown illumination. TOG (2021)
47. Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J.B., Freeman, W.T.: Shape and material from sound. In: NeurIPS (2017)
48. Zhao, X., Srinivasan, P.P., Verbin, D., Park, K., Martin-Brualla, R., Henzler, P.: Illuminerf: 3d relighting without inverse rendering. In: NeurIPS (2024)