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

VibraVerse: A Large-Scale Geometry-Acoustics Alignment Dataset for Physically-Consistent Multimodal Learning

Bo Pang* Chenxi Xu* Jierui Ren Guoping Wang Sheng Li†
Peking University, Beijing, China
bo98@stu.pku.edu.cn, xuchenxi@pku.edu.cn, jerry@stu.pku.edu.cn, {wgp, lisheng@}@pku.edu.cn

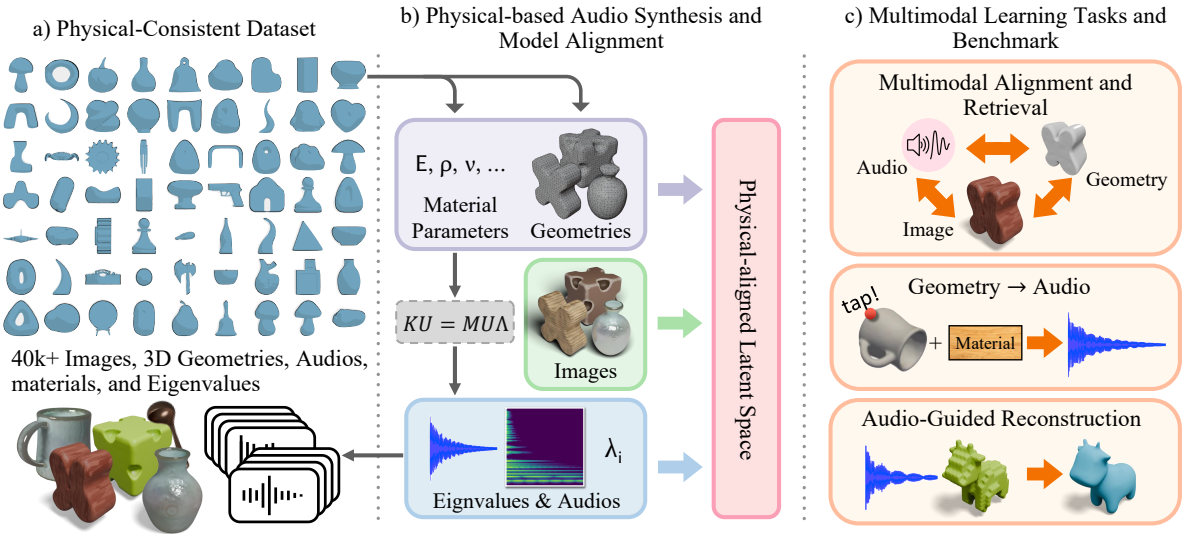


Figure 1. Overview of our framework for physically-consistent geometry–acoustics learning. (a) We build a large-scale physically-consistent dataset comprising over 40K objects, each annotated with images, 3D geometries, materials, eigenvalues, and physically synthesized audios. All data are generated under unified physical parameters to ensure geometry–material–acoustics consistency. (b) Using finite-element modal analysis, we derive eigenfrequencies and modal sounds, aligning each object's geometry and material with its intrinsic acoustic response in a shared physics-grounded latent space. (c) This physically-consistent dataset serves as the foundation for multimodal learning and reasoning, enabling **cross-modal alignment and retrieval**, **geometry-to-audio synthesis**, and **audio-guided 3D reconstruction**. The dataset and inference tasks establish a benchmark for physically-grounded multimodal understanding and sound-driven 3D reasoning, as a bridge enabling physically interpretable multimodal understanding of the physical world.

Abstract

Understanding the physical world requires perceptual models grounded in physical laws rather than mere statistical correlations. However, existing multimodal learning frameworks, focused on vision and language, lack physical consistency and overlook the intrinsic causal relationships among an object's geometry, material, vibration modes, and the sounds it produces. We introduce VibraVerse, a large-scale geometry–acoustics alignment dataset that explicitly bridges the causal chain from 3D geometry → physical attributes → modal parameters → acoustic signals. Each 3D model has explicit physical properties (density, Young's modulus, Poisson's ratio) and volumetric geometry, from which modal eigenfrequencies and eigenvectors are computed for impact sound synthesis under controlled excitations. To establish this coherence, we introduce CLASP, a contrastive learning framework for cross-modal alignment that preserves the causal correspondence between an object's physical structure and its acoustic response. This framework enforces physically consistent alignment across modalities, ensuring that every sample is coherent, traceable to the governing equations, and embedded within a unified representation space spanning shape,

*Equal contribution. Authors are ordered alphabetically by last name.
†Sheng Li is the corresponding author.

image, and sound. Built upon VibraVerse, we define a suite of benchmark tasks for geometry-to-sound prediction, sound-guided shape reconstruction, and cross-modal representation learning. Extensive validations on these tasks demonstrate that models trained on VibraVerse exhibit superior accuracy, interpretability, and generalization across modalities. These results establish VibraVerse as a benchmark for physically consistent and causally interpretable multimodal learning, providing a foundation for sound-guided embodied perception and a deeper understanding of the physical world. The dataset is available at https://huggingface.co/datasets/technetium66/VibraVerse.

1. Introduction

Sound and geometry are inherently linked through the laws of physics: when an object vibrates or is struck, its shape and material properties jointly determine how it resonates and emits sound. In essence, sound is the temporal and spectral projection of an object's geometry and physical constitution. Humans naturally leverage this relation, and we can often infer an object's material or thickness simply from the way it sounds. However, such auditory-based reasoning remains largely unexplored.

Early works in computer graphics and computational acoustics [2, 38] showed that an object's eigenfrequencies and eigenmodes can be derived from its geometry and material properties via finite-element or modal analysis. However, these physically based methods function only as forward models for sound synthesis, unable to address the inverse problem of inferring geometry from sound.

In computer vision and multimodal learning, several recent works have attempted to connect auditory and visual cues. For instance, *SoundSpaces* [4], and *MultiFoley* [10], explored linking sound with visual scenes or coarse 3D reconstruction from videos. Yet, these datasets are based on real-world recordings, which suffer from uncontrolled excitation, environmental noise, and unknown material properties, thereby lacking physical consistency between geometry and sound. Recently, DiffSound [22] introduced a differentiable modal sound rendering framework that enables inverse inference of geometry and material from sound under a fully physics-based pipeline.

From a data perspective, existing large-scale datasets focus primarily on semantics rather than physical causality. Audio–visual datasets such as *AudioSet* [19], *VGGSound* [6], *Fair-Play* [15], and *SoundSpaces* [4] capture environmental or human-generated sounds but omit object-level geometry and material information. Conversely, 3D geometry datasets such as *ShapeNet* [3], *ModelNet* [41], and *Objaverse* [14] contain rich shape diversity but lack any acoustic or modal annotations related to physical attributes.

While the ObjectFolder series [16–18] advanced audio related multisensory learning, it remains limited in capturing physically grounded geometry–acoustics relationships. **First**, the coupling among geometry, material, and sound is implicit rather than causal: auditory signals are produced from event-based surface interactions and generally represent empirical correlations between modalities but lacking explicit modeling of the physical mechanisms that link shape and material to acoustic behavior. **Second**, current datasets omit physically parameterized representations, including volumetric geometry, modal spectra, eigenfrequencies, and material properties necessary for describing intrinsic physical characteristics and enabling cross-domain generalization beyond purely data-driven associations. **Finally**, there is no systematic benchmark for assessing geometry–acoustics consistency or for evaluating physically grounded reasoning tasks, such as sound-guided shape reconstruction and geometry-to-sound synthesis, that require causal, physically interpretable understanding.

The absence of explicit physical grounding constrains representation learning, generalization, and interpretability. Without physical parameters, models capture only statistical correlations rather than the causal mechanisms governing real-world object behavior, leading to poor generalization across variations in shape, material, and volumetric structure. Incorporating physical features derived from modal analysis enables physically interpretable multimodal learning, improved cross-domain generalization, and quantitative evaluation of geometry–acoustics consistency.

Motivation and Objectives: To bridge these gaps, we aim to construct a **physically consistent geometry–acoustics alignment dataset** that explicitly encodes the causal chain: Geometry (3D shape and volumetric data) $\rightarrow$ Physical properties (E, ρ, ν) $\rightarrow$ Modal parameters (eigenvalue, eigenvector) $\rightarrow$ Sound signals. Each 3D model has explicit material parameters and is subjected to modal analysis to compute its vibration modes. The resulting eigenfrequencies and mode shapes are then used to synthesize the corresponding impact sound under controlled excitation. This process guarantees that every sample in the dataset is physically coherent, where geometry, material, and sound are tightly coupled through physical laws and fully traceable to simulation parameters.

Building upon this foundation, we introduce a large-scale dataset for physically consistent multimodal learning, enabling AI systems to jointly reason across 3D geometry, 2D image, material, and impact sound. We further design a suite of novel cross-modal and inverse reasoning tasks, including geometry-to-sound synthesis, sound-guided shape reconstruction, material identification, and tri-modal retrieval, many of which have not been feasible with previous datasets. Together, these components provide a platform for evaluating physics-aware multimodal learning

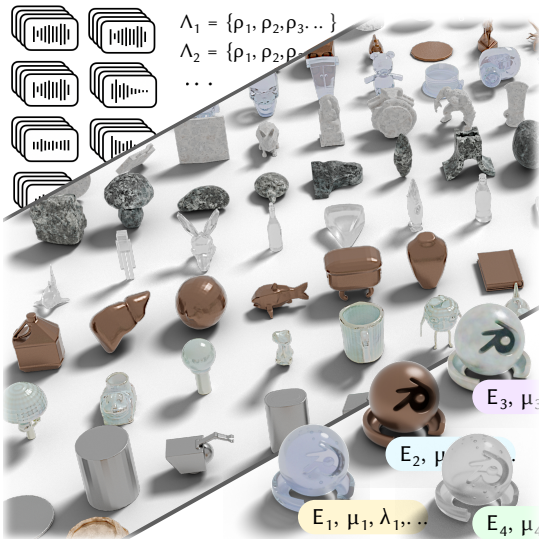


Figure 2. The VibraVerse dataset comprises a diverse collection of objects spanning a wide range of physical materials (bottom). Each object is defined by its physical parameters, which are utilized to synthesize corresponding eigenfrequencies, eigenmodes, and modal sounds (top). This process establishes a physically grounded correspondence linking object geometry, material properties, and acoustic signatures.

and establish a benchmark for physically grounded perception and reasoning beyond purely semantic or visual alignment (see Fig.1).

Our key contributions include:

1. **A large physically-consistent geometry–acoustics dataset.** Each object is associated with complete 3D geometry (both surface and volumetric data), physical attributes, modal spectra, and synthesized sound, forming a traceable physical chain. All samples have been verified for physical plausibility and simulation consistency.
2. **Multimodal tasks and benchmark.** We establish a suite of tasks, including *geometry-to-sound prediction*, *shape reconstruction*, *cross-modal retrieval*, and *material classification*, along with evaluation protocols and a physically aligned contrastive learning framework that unifies geometry–sound representations. Together, these form a benchmark for consistent evaluation of physically grounded multimodal reasoning.

2. Related Works

2.1. 3D Datasets on Object-level Geometry

With the rapid development of 3D vision tasks, numerous datasets and benchmarks have been proposed.Early datasets such as ModelNet [41] and ShapeNet [3] provided large-scale collections of 3D models for object classification and segmentation tasks. Subsequently, datasets like Thingi10K [45] and ABC [23] introduced more diverse and complex 3D shapes, enabling advancements in shape analysis and reconstruction. Recent web-scale datasets such as Objaverse [14], Objaverse-XL [13], Animal3D [42], and OmniObject3D [40] further expand category coverage, spanning man-made, organic, and articulated objects.

2.2. Acoustic-Related Datasets

Several sound-related datasets have been developed to support research in audio recognition and generation. AudioSet [19] serves as a foundational dataset for audio classification tasks, recording 632 categories of human-labeled sound events. Zhang et al. [43] introduced a synthetic dataset of object shape, material and corresponding sound, enabling the study of sound generation from visual inputs. Gao et al. [16, 17] proposed ObjectFolder, an object-centric dataset of around 1,000 everyday objects with high-quality 3D models and sound simulations, together with touching events. Clarke et al. [12] presented RealImpact, which records 150k knock sound from 50 real-world objects in highly controlled acoustic environments.

2.3. Multimodal Learning involving Physical Attributes

Recent multimodal approaches increasingly incorporate *physical attributes* such as material and contact dynamics, where impact acoustics provide cues complementary to visual appearance [12]. At the object level, Object-centric datasets such as ObjectFolder [16, 17] capture multimodal data including geometry, rendering, and contact-induced sounds. At the scene level, SoundSpaces [4, 5] simulates room impulse responses aligned with 3D indoor environments. The physical realism has been further improved by recent efforts in real-scene acoustic field measurements [9] Other works like Neural Acoustic Fields and audio-visual neural radiance field models [27, 29] jointly encode geometry and sound propagation. However, these scene-level and acoustic-level approaches generally lack object-specific modal alignment.

3. VibraVerse Dataset

In this section, we detail the construction of the VibraVerse dataset, a dataset including more than 40,000 3D objects and their corresponding physical properties, modal parameters, and synthesized sounds. A visual overview of the dataset is shown in Fig. 2. We first provide an overview of modal analysis and sound synthesis. As compact descriptors of an object's global physical behavior and material attributes, modal representations guide the data generation process and underpin our pursuit of physically consistent multimodal learning. Then, we formally define the formulation of our dataset and its components. Then, we detail the steps to create the 3D geometry, assign material properties, and perform acoustic simulation to generate physically

consistent sound signals.

3.1. Formulation of Modal Analysis and Sound Synthesis

Modal Analysis. The synthesis of physically plausible impact sounds from 3D objects is fundamentally based on analyzing their inherent vibrational properties. Given a 3D object represented by a volumetric mesh, its physical behavior is governed by its geometry and material properties, specifically density (ρ), Young's modulus (E), and Poisson's ratio (ν). Finite Element Method (FEM) [33] discretizes the continuous object, allowing the construction of global mass (M) and stiffness (K) matrices, which encapsulate the inertial and elastic properties of the object, respectively. Assuming small deformations and linear elastic material, the undamped free vibration of the discretized object is governed by [33]:

$$M\ddot{x} + Kx = 0, \tag{1}$$

where x is the displacement vector of the mesh nodes, and $\ddot{x}$ represents nodal accelerations. By assuming a set of harmonic solutions of the form $x_j = u_j e^{i\omega_j t}$, where u_j is the mode shape and ω_j is the $j-$th angular natural frequency, it transforms into a generalized eigenvalue problem as:

$$KU = MU\Lambda. \tag{2}$$

Here, Λ is a diagonal matrix of the object's eigenvalues $\lambda_1, \ldots, \lambda_n$, where $\lambda_j = \omega_j^2$, and $U = [u_1, \ldots, u_n]$ contains mode shapes (eigenvectors) of the vibrations. Solving this problem yields a set of eigenvalues λ_j, which relate to the natural frequencies of the object's vibration ($f_j = \frac{\sqrt{\lambda_j}}{2\pi}$), and their corresponding mode shapes (eigenvectors) u_j.

Sound Synthesis. Once the eigenvalues λ_i and eigenvectors u_i are computed, The nodal displacement $x(t)$ can be represented as a linear combination of its mode shapes and reduced modal coordinates $q_i(t)$ [1]:

$$x(t) = \sum_i u_i q_i(t). \tag{3}$$

For the full damped equation of motion [11]:

$$M\ddot{x} + C\dot{x} + Kx = f(t), \tag{4}$$

where $C = \alpha M + \beta K$ is the Rayleigh damping matrix, we can decouple Eq. (4) for the i-th mode using $q_i(t)$:

$$\ddot{q}_i(t) + (\alpha + \beta\lambda)\dot{q}_i(t) + \lambda_i^2 q_i(t) = F_i(t) \tag{5}$$

where $F_i(t) = u_i^T f(t)$ is the modal force of $i-$th mode reduced from nodal force $f(t)$, and ζ_i is the modal damping ratio. For an impulse excitation $F_i(t)$ at $t = 0$, the solution to each mode is a damped sinusoid. The resulting audio signal is their superposition:

$$S(t) = \sum_i A_i e^{-\sigma_i t} \sin(\omega_{d,i} t) \tag{6}$$

where A_i is the amplitude of mode i, $\omega_{d,i}$ is the damped natural frequency, and σ_i is the decay rate of mode i.

3.2. Dataset Components and Formulation

Formally, each sample in the VibraVerse dataset consists of the following components:

- **3D Geometry**: Watertight 3D models defined by triangle surface meshes and their corresponding tetrahedral volumetric discretizations.
- **Visual Representation**: A single-view image rendered from a fixed viewpoint, using predefined materials.
- **Material Properties**: Physical attributes such as density (ρ), Young's modulus (E), and Poisson's ratio (ν) that define the object's material characteristics.
- **Modal Parameters**: Eigenfrequencies and modal shapes obtained through modal analysis, which describe the object's vibrational behaviors.
- **Audios**: Sample audio signals generated based on the modal parameters and the Rayleigh damping coefficient (alpha, beta).
- **Metadata**: Additional information such as object category, size, source, and other relevant attributes.

3.3. VibraVerse Dataset Creation Pipeline

3.3.1. Geometry Creation, Processing

We source 3D geometries from a combination of publicly available datasets and procedural generative techniques.

Geometry Sources Part I: Objaverse. We curated an open-source dataset from Objaverse [14] by applying a rigorous filtering process based on Objaverse++ [28] annotations. The selection criteria required models to be single, non-scene, non-transparent, non-humanoid objects with a quality score ≥ 2 and a file size $< 5MB$.

Geometry Sources Part II: Generated. We generated a synthetic dataset of 40,000 models using a two-stage pipeline. First, Flux Dev [24] synthesized 2D images from prompts generated by Google Gemini [37]. Second, Hunyuan3D 2.0 [44] reconstructed 3D geometries from these images. This process yielded 2,000 instances for each of 20 distinct categories (see Appendix for details).

Geometry Processing. The acquired models often contain simulation-inhibiting defects (e.g., holes, non-manifold edges). Our preprocessing pipeline first normalizes all models (translation to origin, scaling to [-1, 1]). We then

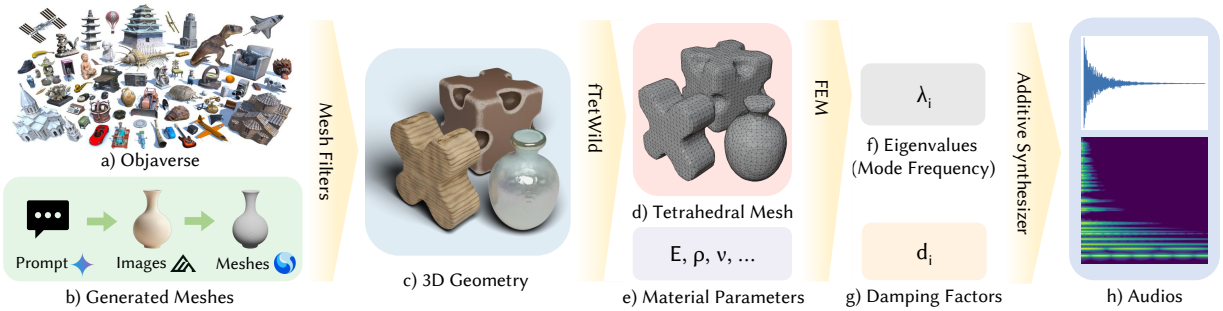


Figure 3. Pipeline for generating our VibraVerse dataset. Meshes from Objaverse and text-to-3D generation are filtered and then tetrahedralized, assigned material parameters, and analyzed via finite-element modal analysis to obtain eigenvalues and damping factors. An additive synthesizer then produces corresponding modal sounds, forming physically consistent geometry–acoustics pairs.

perform a voxel remesh to generate high-quality watertight manifolds from the (potentially non-watertight) inputs. Finally, the mesh is made simulation-friendly using the method from GeGnn [30].

3.3.2. Physical Validity Filtering

To ensure both the physical validity and representational diversity of the dataset, we implement a rigorous multi-stage geometry filtering and validation pipeline prior to modal analysis and acoustic synthesis. This process enforces structural coherence, topological simplicity, and physical plausibility, thereby ensuring that all retained models serve as stable and meaningful samples for geometry-acoustics learning. Specifically:

Topological Connectivity Filtering. We first eliminate geometries that contain multiple disconnected components or floating elements, which violate the assumption of a single physically coherent body required for modal analysis. Connectivity is assessed using graph-based component labeling and surface adjacency detection. Only single-connected meshes are preserved to guarantee well-defined boundary conditions for vibration mode computation.

Topological and Geometric Complexity Control. To prevent numerical instability and the overrepresentation of degenerate structures, we exclude non-manifold geometries, and models exceeding a specified topological genus threshold are removed. This ensures the dataset remains computationally tractable and physically interpretable.

Physical Plausibility and Modal Validity Screening. Finally, each candidate undergoes a physical sanity check to validate its suitability for finite-element modal analysis. We discard meshes that fail to satisfy minimum thickness constraints (to avoid thin-shell structures prone to non-linear vibrations) or that yield numerically unstable or non-physical eigenvalue spectra, such as negative eigenvalues, to preserve the integrity of the geometry–physics correspondence.

Through this hierarchical filtering pipeline, the resulting dataset achieves a high level of geometric fidelity, topological soundness, and physical realism, providing a reliable foundation for downstream physically-consistent learning tasks. This also helps maintain a high-quality dataset that is diverse and plentiful. After the above filtering steps, we obtain a final set of approximately 46,000 high-quality 3D geometries for inclusion in the VibraVerse dataset, 10,000 from Objaverse++ and 36,000 from our generation pipeline.

3.4. Material Properties

The material properties of each object are crucial for determining its vibrational characteristics and the resulting sound. We assign material properties based on a predefined set of common materials, such as wood, metal, plastic, and glass. Each material is characterized by its density ρ, Young's modulus E, and Poisson's ratio ν. Specifically, we define a material library with the following materials and their corresponding properties (See Appendix for details). The dataset contains 10 material categories: wood, plastic, ceramic, glass, steel, copper, aluminum, concrete, stone, and polycarbonate. To determine the material properties for each object, we employed two distinct strategies. For models originating from the Objaverse dataset, we leveraged a Vision-Language Model (VLM) to classify their rendered visual appearance into a set of predefined material categories. For our procedurally generated models, we programmatically assigned a plausible material category based on the object's semantic class.

3.5. Sound Synthesis

For each 3D object with its geometry and material properties defined, we compute its natural frequencies, which are subsequently used to synthesize corresponding impact sounds. The process is as follows: first, we convert the 3D geometry into an explicit volumetric tetrahedral mesh using the method in fTetWild [20]. Subsequently, following the Modal Analysis method detailed in Section 3.1, we solve the 64 smallest eigenvalues λ_i, which correspond to the squared natural frequencies of the first 64 vibrational

modes $\omega_i = \sqrt{\lambda_i}$. This decomposition is performed using the ARPACK library [25].

Finally, to synthesize the impulse audio waveform, we apply a unit impulse excitation $\delta(t)$ to each mode:

$$F_i(t) = \delta(t), i = 1, \ldots, 64 \tag{7}$$

Substituting this into Eq. (5) allows us to solve for the amplitude A_i, damped frequency ω_i and damping coefficient d_i of each mode's time-dependent vibration signal:

$$S_i(t) = A_i e^{-d_i t} sin(2\pi\omega_i t). \tag{8}$$

Following Eq. (8), we sample a 1-second signal at a sample rate of 32,000 Hz for each of the 64 modes. The resulting signals $S(t)$ are then summed to produce the object's corresponding impact sound.

4. Benchmark Tasks and Validation

We designed several benchmark tasks to validate VibraVerse for physically-consistent multimodal learning. These tasks evaluate the cross-modal mappings between 3D geometry, materials, and acoustics, covering applications like conditional generation, reconstruction, and retrieval. We detail the experimental setup, methods, evaluation protocols, and quantitative/qualitative results for each task to demonstrate the dataset's effectiveness.

Unless otherwise specified, the following experiments are all based on our full dataset, with a training/testing split of 90%/10%. We provide the technical detail of each task and experimental settings in the supplementary material.

4.1. Geometry→Sound: Data-Driven Synthesis

Given the geometry and material parameters of a 3D object, FEM-based modal analysis shows the process of synthesizing its impact sound (Sec. 3.1). However, it requires generalized eigenvalue decomposition on large matrices, which is computationally expensive and slow. To evaluate the effectiveness of the VibraVerse dataset for data-driven sound synthesis, we design a learning-based task in which a neural network takes an object's 3D geometry and material parameters (density, Young's modulus, and Poisson's ratio) as input and directly predicts its natural frequencies.

We use an OCNN-based [39] shape encoder to extract geometric features from the input 3D shapes, and a multilayer perceptron (MLP) is used to encode the material parameters, then both features are concatenated and fed into a Sinusoidal Representation Network (SIREN [36]), which predicted the first 64 modal frequencies. We minimize the mean squared error (MSE) between the predicted scaled frequencies and their corresponding ground-truth values.

Specifically, we compare to the following methods:

- **NeuralSound [21]:** Retraining NeuralSound on our dataset even improves its vibration-solver performance beyond the original report.
- **FEM [33]**: We perform FEM-based modal analysis using two eigenvalue solvers:ARPACK and LOBPCG.

We evaluate the quality and efficiency of all methods using two metrics. The Frequency error is defined as the Mean Squared Error (MSE) between predicted and ground-truth frequencies on scaled Mel spectrograms. The Time cost is measured as the total time taken to process the test set, which contains approximately 4,600 meshes. The detailed statistics are shown in Tab. 1.

Table 1. Comparison of different methods. Note that ARPACK and LOBPCG, being traditional FEM-based solvers, are generally treated as ground truth. Our dataset enables superior performance.

Method	Freq. error ↓	Time cost (s) ↓
FEM (ARPACK)	$< 10^{-7}$	15374
FEM (LOBPCG)	$< 10^{-7}$	14112
NeuralSound	3.50×10^{-3}	441
Ours	6.06×10^{-4}	**353**

4.2. Sound-Guided Shape Reconstruction

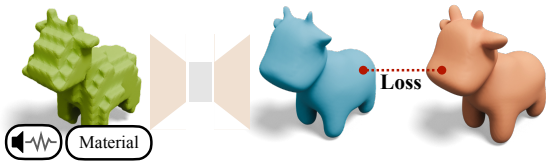


Figure 4. Sound-Guided Shape Reconstruction. Given a voxel initial shape, the audio eigenvalues, and material properties, we reconstruct the 3D geometry in just one forward pass.

Inferring a complete 3D geometry from impact sound is inherently an ill-posed problem: acoustic responses encode only a subset of an object's modal characteristics and are highly sensitive to local structural variations, allowing distinct geometries to produce nearly indistinguishable sound patterns. To date, the only method capable of performing geometry-from-sound inversion is DiffSound [22], which introduces a differentiable, physics-based modal sound rendering framework to infer geometry under a coarse voxel constraint. Motivated by this challenge, we designate sound-guided shape reconstruction as one of the core benchmark tasks in VibraVerse, enabling the systematic evaluation of learnable and generalizable geometry inference from sound.

The overall pipeline of this reconstruction is illustrated in Fig. 4. Same as [22], we take sound eigenvalues, as well as a coarse voxel grid and material parameters as input, and reconstruct the detailed 3D geometry. We apply the VAE structure and training methodology of Step1X-3D [26]. We use the VAE to encode the voxel grid, which is then con-

catenated with the encoded audio features and material embeddings as conditions. The combined features are then fed into a decoder network to reconstruct the final 3D geometry. We compared the performance of DiffSound and our method on this task. We randomly sampled 100 test meshes from the test set, evaluated the accuracy using Intersection over Union (IoU) and Chamfer Distance (CD), and measured efficiency using inference time on test meshes. Quantitative results are reported below, and qualitative examples are shown in Fig. 5. The results show that our VibraVerse dataset can facilitate a data-driven approach to directly reconstruct 3D geometry from audio, achieving simultaneous improvements in both accuracy and efficiency.

Method	**IoU** ↑	**CD** ↓	**Time(s)** ↓
Initial Voxel	0.837	4.97×10^{-3}	\
DiffSound	0.856	4.45×10^{-3}	46594
Ours	**0.871**	$\mathbf{3.32 \times 10^{-3}}$	**175** (×0.004)

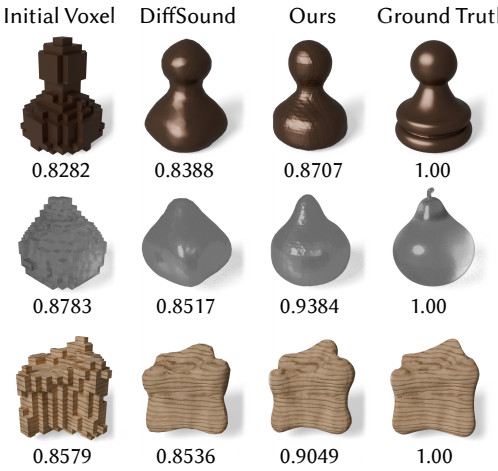


Figure 5. Results of audio-guided reconstruction. From left to right are initial shapes, DiffSound results, our results, and the ground truth. The IoU metric is shown below each shape.

4.3. Cross-Modal Retrieval

Our VibraVerse dataset provides a unified platform to explore the mutual correspondence between shape, vision, and sound. To demonstrate the effectiveness of our dataset in terms of cross-modal tasks, inspired by CLIP [31], we design a contrastive learning [7, 8] framework for cross-modal retrieval between 3D shapes, 2D images, and sounds, named *Contrastive Learning of Audio, Shape, and Physical-Properties* (CLASP), as in Fig. 6. We use different encoders to extract features from each modality, and train the model using a contrastive loss to align the embeddings of matching pairs while pushing apart non-matching pairs, with the

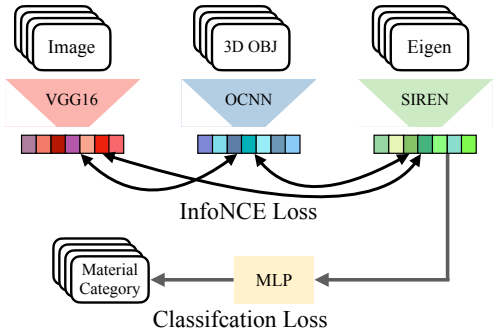


Figure 6. Architecture of the CLASP model. Three encoders are used to extract features from each modality, and a contrastive learning mechanism is employed to align the features.

InfoNCE loss:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{9}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the embeddings of a matching pair from different modalities, $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, τ is a temperature hyperparameter set to 0.07, and N is the total number of samples in the batch.

The model consists of a SIREN-based [36] sound encoder, an OCNN-based [39] 3D shape encoder, and a VGG-based [35] image encoder. To retrieve the most relevant item from a library, we compute the cosine similarity between the query embedding and all candidate embeddings, selecting the those with highest similarity scores.

We consider cross-modal retrieval tasks across three modalities: audio (or its eigenvalues), 3D geometry, and 2D images. The tasks focus on the bidirectional retrieval between sound and 3D shapes, and between sound and 2D images. The visual result of cross-modal retrieval is shown in Fig. 7. Quantitative results are presented in Sec. 4.3, where we report the Recall@K (R@K) metrics for each retrieval task across different subsets of our VibraVerse dataset. A more detailed comparison between our dataset and Objectfolder [14] is provided in the supplementary material.

4.4. Sound→Material: Material Classification

Material classification is defined as predicting an object's material category solely from its acoustic properties. We add a classification head in CLASP, as the bottom part of Fig. 6, which takes in the embedding extracted from encoder and predicts the material category. The prediction accuracies are reported below, which suggest that our dataset effectively supports material classification tasks.

	Objaverse	Generated	All
Accuracy ↑	51.03%	89.54%	80.33%

Some previous works [12, 16–18] have also explored

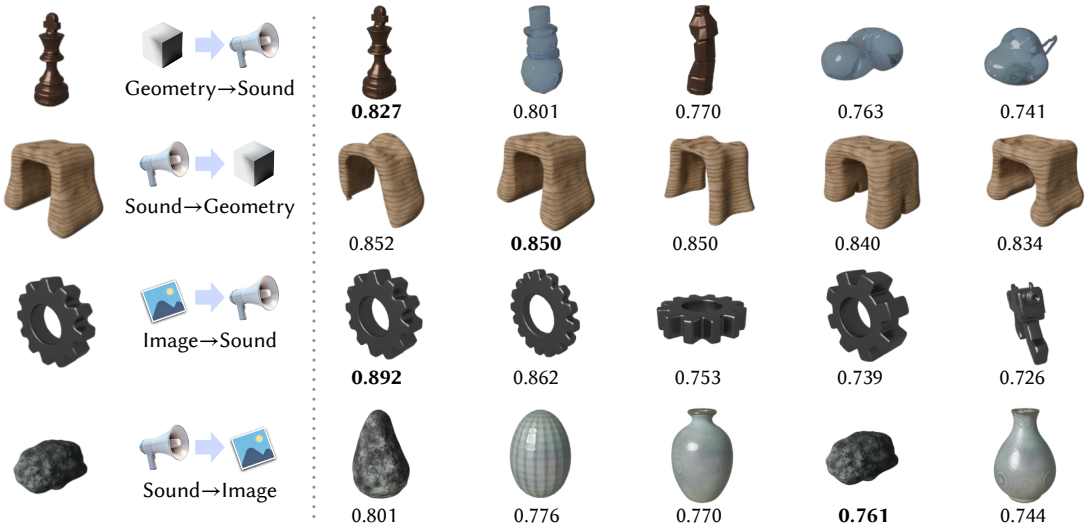


Figure 7. Cross-modal retrieval results. Given a query from one modality (left), we retrieve the most relevant items from other modalities (right). The numbers are the cosine similarity, while the bold number indicates the ground truth.

	Overall (#Sample = 4672)			Objaverse (#Sample = 1117)			Generated (#Sample = 3555)		
Task	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
Geometry → Sound	0.409	0.766	0.865	0.288	0.654	0.787	0.484	0.848	0.929
Sound → Geometry	0.417	0.768	0.866	0.312	0.662	0.797	0.488	0.854	0.936
Image → Sound	0.287	0.610	0.727	0.207	0.475	0.604	0.334	0.688	0.807
Sound → Image	0.308	0.617	0.732	0.224	0.474	0.610	0.359	0.702	0.818
Geometry → Image	0.509	0.837	0.909	0.474	0.815	0.892	0.548	0.873	0.930
Image → Geometry	0.499	0.835	0.914	0.471	0.806	0.895	0.532	0.869	0.936

Table 2. Cross-modal retrieval performance (R@1, R@5, R@10) of our datasets. Retrieval between geometry and sound achieves higher accuracy compared to retrieval between image and sound, likely due to the more direct correlation between geometry and sound. Our result has strong performance on all retrieval tasks, demonstrating the effectiveness of our VibraVerse dataset in facilitating cross-modal learning.

this task, but with different settings. For a detailed comparison, please refer to our supplementary material.

5. Conclusion

Our VibraVerse dataset and benchmark suite provide a large-scale, physics-grounded foundation in which geometry, material, and sound are explicitly coupled through physically consistent simulation. By integrating principles from computational acoustics into multimodal learning, it enables models to infer geometric and material properties from auditory cues, marking a step toward physically interpretable auditory intelligence.

We reinterpret modal analysis as a language of physical behavior that reveals how objects vibrate, store energy, and express their intrinsic properties. As the dynamic fingerprint of an object, it provides the causal foundation linking geometry, material, and sound within a unified representation framework. Building on this perspective, our work advances sound-guided 3D perception, physics-consistent multimodal reasoning, and embodied physical understanding, with promising potential for sim-to-real transfer.

Nevertheless, our approach has limitations. All data in VibraVerse are fully synthetic and generated under idealized simulation conditions, which may not fully capture the noise and variability of real-world acoustic measurements. Moreover, the benchmark has not yet been validated against real recordings or experimentally measured modal properties, leaving the degree of sim-to-real generalization to be explored in our future work.

Beyond serving as a benchmark for multimodal reasoning, our dataset also holds potential for advancing physics-informed neural networks (PINNs) [32] and neural-based physical simulation [34], offering a pathway to unify data-driven learning with physically grounded modeling. These need further validation and will be our future work.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No. 2023YFF0905103). This work was also partly supported by the Shenzhen Science and Technology Program (Grant No. KJZD20230923114114028) and the Shandong Key Research and Development Program (Grant No. 2024CXGC010801).

References

[1] Jernej Barbic. Siggraph 2012 course notes fem simulation of 3d deformable solids: A practitioner's guide to theory, discretization and model reduction. part 2: Model reduction (version: August 4, 2012). 4

[2] John N. Chadwick, Sam Parker, and Doug L. James. Harmonic shells: A practical nonlinear sound model for near-rigid thin shells. *ACM Transactions on Graphics (TOG)*, 28 (5):119:1–119:10, 2009. 2

[3] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 3

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European conference on computer vision*, pages 17–36. Springer, 2020. 2, 3

[5] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022. 3

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 7

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 7

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

[10] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18770–18781, 2025. 2

[11] Indrajit Chowdhury and Shambhu P Dasgupta. Computation of rayleigh damping coefficients for large systems. *The Electronic Journal of Geotechnical Engineering*, 8(0):1–11, 2003. 4

[12] Samuel Clarke, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L James, and Jiajun Wu. Realimpact: A dataset of impact sound fields for real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2023. 3, 7

[13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 3

[14] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 2, 3, 4, 7

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

[16] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021. 2, 3, 7

[17] 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 *CVPR*, 2022. 3

[18] 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 (CVPR)*, pages 17276–17286, 2023. 2, 7

[19] Jort F Gemmeke, Daniel PW 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. IEEE, 2017. 2, 3

[20] Yixin Hu, Teseo Schneider, Bolun Wang, Denis Zorin, and Daniele Panozzo. Fast tetrahedral meshing in the wild. *ACM Transactions on Graphics (ToG)*, 39(4):117–1, 2020. 5

[21] Xutong Jin, Sheng Li, Guoping Wang, and Dinesh Manocha. Neuralsound: learning-based modal sound synthesis with acoustic transfer. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 6

[22] Xutong Jin, Chenxi Xu, Ruohan Gao, Jiajun Wu, Guoping Wang, and Sheng Li. Diffsound: Differentiable modal sound

rendering and inverse rendering for diverse inference tasks. In *SIGGRAPH '24: ACM SIGGRAPH Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 6

[23] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. ABC: A big CAD model dataset for geometric deep learning. In *CVPR*, 2019. 3

[24] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 4

[25] Richard B Lehoucq, Danny C Sorensen, and Chao Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, 1998. 6

[26] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 6

[27] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36:37472–37490, 2023. 3

[28] Chendi Lin, Heshan Liu, Qunshu Lin, Zachary Bright, Shitao Tang, Yihui He, Minghao Liu, Ling Zhu, and Cindy Le. Objaverse++: Curated 3d object dataset with quality annotations. *arXiv preprint arXiv:2504.07334*, 2025. 4

[29] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022. 3

[30] Bo Pang, Zhongtian Zheng, Guoping Wang, and Peng-Shuai Wang. Learning the geodesic embedding with graph neural networks. *ACM Trans. Graph. (SIGGRAPH ASIA)*, 42(6), 2023. 5

[31] 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. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[32] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. 8

[33] Junuthula Narasimha Reddy. An introduction to the finite element method. *New York*, 27(14), 1993. 4, 6

[34] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020. 8

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[36] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 6, 7

[37] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4

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

[39] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-CNN: Octree-based convolutional neural networks for 3D shape analysis. *ACM Trans. Graph. (SIGGRAPH)*, 36(4), 2017. 6, 7

[40] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 3

[41] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 2, 3

[42] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023. 3

[43] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H McDermott, Joshua B Tenenbaum, and William T Freeman. Generative modeling of audible shapes for object perception. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1251–1260, 2017. 3

[44] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 4

[45] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10,000 3d-printing models. *arXiv preprint arXiv:1605.04797*, 2016. 3