2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Audio-2-Shape: 3D Generation from What You Hear

Xuran He[1] and Xian-Feng Han[1] and Shi-Jie Sun [2]

***Abstract*— Audio serves as an important bridge connecting humans to their surroundings, providing a unique modality for perceiving the world. For embodied AI systems, such as robots and autonomous vehicles, enabling them to understand the world through sound is a promising and significant research direction. In this paper, we explore the underexplored domain of audio-driven 3D shape generation and propose a novel architecture for audio-conditioned 3D shape synthesis. Specifically, our framework comprises three key modules: cross-modal alignment, a latent diffusion model for generation, and a 3D Gaussian Splatting (3DGS) based optimization module. We first align audio and 3D shape representations within a unified embedding space using a contrastive learning strategy, which conditions a latent diffusion model to generate an initial coarse 3D structure. Subsequently, we introduce a refinement stage utilizing 3D Gaussian Splatting to produce high-fidelity 3D shapes. Extensive qualitative and quantitative experiments validate the effectiveness of our proposed method, demonstrating its capability to generate semantically coherent 3D shapes from audio input.**

I. Introduction

Humans perceive, understand, and interact with the world through a combination of sensory modalities, including hearing, sight, touch, smell, and taste [1]. Among these, audio provides a richly structured modality that encodes temporal dynamics, semantic cues, and spatial properties of physical interactions, conveying vital information about actions, environmental context, and emotional tone [2], [3]. Recent advancements in artificial intelligence have led to the development of multimodal embodied agents [4] equipped with audio sensors, allowing them to process and interpret sound in ways that mimic human capabilities. This evolution not only enhances the understanding of auditory information but also opens new avenues for interpreting and synthesizing 3D structures from sound. Such capabilities are becoming increasingly essential for human-AI collaboration across various domains, including robotics [5], augmented reality [6], autonomous driving [7], and immersive 3D content creation [8]. Consequently, cross-modal generation has emerged as a compelling area of research [9], [10], [11].

Existing studies have primarily concentrated on vision-based 3D reconstruction [12], text-conditioned 3D generation [13], 2D image synthesis from sound [14], and audio-visual cross-modal learning [15], [16], [17], [18]. These approaches typically build upon well-established multimodal data pairs, employing techniques based on Generative Adversarial Networks (GANs) [19], [20] or diffusion models [21], [22], [23]. Despite substantial progress in these areas, the direct synthesis of 3D geometry and appearance from audio signals remains largely unexplored, and there is no prior work on audio-to-3D shape generation, in which acoustic cues drive the creation of geometrically plausible and texturally coherent 3D assets. This raises a critical question: Can AI agents "imagine" 3D structures solely from auditory observations, emulating human cognitive capabilities?

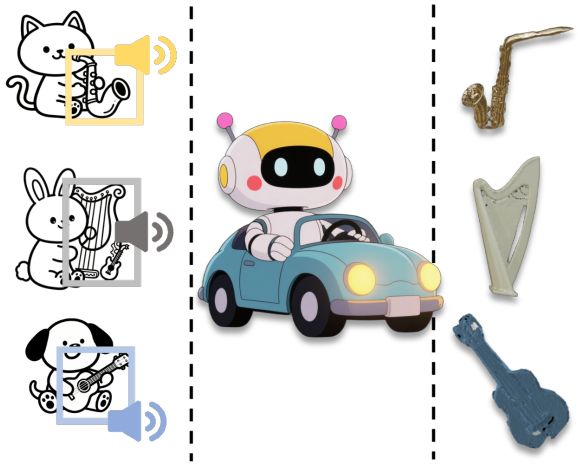

Fig. 1: Illustration of our audio-driven 3D generation framework.

Therefore, we aim to investigate the task of audio-conditioned 3D shape generation, with the goal of facilitating more intuitive and interactive AI systems that bridge the gap between human perception and machine understanding. However, there exist fundamental challenges, namely the semantic gap between audio signals and spatial geometric data, as well as the scarcity of high quality audio-point cloud paired datasets, hindering advancements in this field.

To alleviate these challenges, we introduce a novel architecture, termed Audio-2-Shape, for sound-driven 3D generation, leveraging the inherent contextual information of the audio signal to create 3D objects. The overall framework integrates three key components: cross-modal alignment, latent diffusion model based 3D shape generation, and 3D gaussian splatting based optimization. We first bridge the audio-geometric domain gap by distilling knowledge from pre-trained Wav2CLIP (audio-visual) [24] and CLIP (vision-

*This work was supported in part by National College Students' Innovative Entrepreneurial Training Plan Program(202510635030), in part by the Fundamental Research Funds for the Central Universities (SWU-KT24003)

[1]Xuran He is with Hanhong College, Xian-Feng Han is with College of Computer and Information Science, Southwest University, Chongqing, China. dinosaur@email.swu.edu.cn xianfenghan@swu.edu.cn

[2]Shi-Jie Sun is with School of Data Science and Artificial Intelligence, Chang'an University, Xi'an, China shijieSun@chd.edu.cn

Corresponding author: Xian-Feng Han

language) [9] into a 3D encoder, aligning point features, audio features, and visual features into a unified representation space. This alignment ensures that acoustic conditions correlate with geometric properties and semantic concepts. Then, we employ a latent diffusion architecture [25] with cross-attention mechanism to guide the generation of semantically consistent 3D shapes from audio. Finally, we implement a 3D Gaussian splatting optimization strategy to refine the generated 3D shapes with fine grained details.

Our contributions are summarized as follows:

- To the best of our knowledge, this is the first work concentrating on the audio-to-3D shape generation task.
- we propose a novel audio-conditioned 3D generation architecture that integrates cross modal alignment, latent diffusion model based generation, and 3DGS based optimization stages to achieve a coarse-to-fine shape generation.
- We build a multimodal dataset of aligned audio, image, and point cloud triplets to support our generation task. Quantitative and qualitative experiments validate the effectiveness of our proposed framework.

II. Related Work

Audio-Visual Generation. Recent advancements in audio-visual cross-modal generation have attracted increasing attention, driven by its applications in multimedia content creation, virtual reality, and embodied AI interaction. This field encompasses two primary directions: audio-to-visual generation [26] that translates acoustic signals to images [27] or videos [28], [29], [30], and visual-to-audio generation [31], which produces temporally coherent sounds from visual inputs. Both paradigms fundamentally rely on effective cross-modal alignment and advanced generative modeling. Where Generative Adversarial Networks (GANs) [32], Variational Autoencoders (VAEs) [33], and diffusion models have been widely adopted for cross-modal conditioning. Among these, diffusion models have recently emerged as an effective approach, due to their iterative refinement process [34], enabling stable training and high-fidelity generation in complex conditional settings.

In the audio-to-visual generation paradigm, current works focus on 2D tasks such as image reconstruction from sound [35]. For instance, SoundAdapter integrates a transformer-based architecture with multi-granularity supervision to enhance alignment and generalization [36], [14], demonstrating effectiveness in audio-conditioned image generation. In addition, other approaches have used cross-modal attention mechanisms [37] and contrastive learning within teacher-student frameworks [24] to better capture temporal and semantic relationships across modalities.

Unlike well-established image-text or audio-image corpora, the scarcity of large-scale audio-3D shape pairs makes the direct synthesis of 3D geometry and appearance from audio signals remain an open challenge. In this work, we investigate this task by introducing a novel approach to bridge the gap between the audio and 3D domains.

Point Cloud Generation. Point cloud generation has undergone a significant evolution, transitioning from traditional geometric modeling techniques to advanced deep learning approaches. This shift has enabled the creation of more realistic, diverse, and complex 3D structures, which are crucial for applications in virtual reality, robotics, and digital content creation.

Among the pioneering deep learning methods, Generative Adversarial Networks (GANs) have achieved notable success in synthesizing point clouds [38], [39]. In this framework, a generator learns to produce point clouds, while a discriminator evaluates their fidelity and realism against real-world samples. This competitive process guides the generator to approximate the true data distribution, yielding diverse and high-quality 3D outputs.

More recently, diffusion models have emerged as powerful alternatives in generative modeling, renowned for their training stability and superior sample fidelity. Initially demonstrating groundbreaking success in image synthesis, diffusion-based methodologies have been successfully extended to 3D point cloud generation[40], [41]. This is typically achieved by simulating a progressive forward noising process on point clouds, followed by an iterative denoising procedure to reconstruct pristine 3D shapes. Such a gradual generation mechanism contributes to better preservation of structural details, improved topological consistency, and smoother convergence during training.

Beyond direct generative models, concurrent advancements in 3D representation and rendering have also profoundly impacted the ability to create and manipulate 3D data, which can then be used to obtain point clouds. Neural Radiance Fields (NeRF), for instance, introduce a new paradigm by representing 3D scenes as continuous volumetric radiance fields, which allows neural networks to implicitly learn light propagation and reflection, enabling the synthesis of novel views with photorealistic quality [42], [43]. NeRF-based methods can be used to extract dense and geometrically accurate point clouds through various sampling or surface extraction techniques [44], [45].

Furthermore, Gaussian Splatting has recently gained significant attention as an efficient and differentiable approach for 3D scene representation and real-time rendering [46]. By modeling scenes as collections of anisotropic 3D Gaussians, this method allows optimized control over parameters such as position, scale, orientation, and appearance. Although primarily designed for high-fidelity rendering and scene reconstruction from captured data, its differentiable nature opens avenues for generative tasks where these Gaussian parameters can be directly learned or manipulated to effectively synthesize large-scale and detail-rich 3D representations, which can subsequently be rasterized into point clouds [47], [48].

III. Method

We focus on generating semantically meaningful 3D shapes that correspond to the semantic information conveyed by audio inputs. Formally, given an audio signal a from

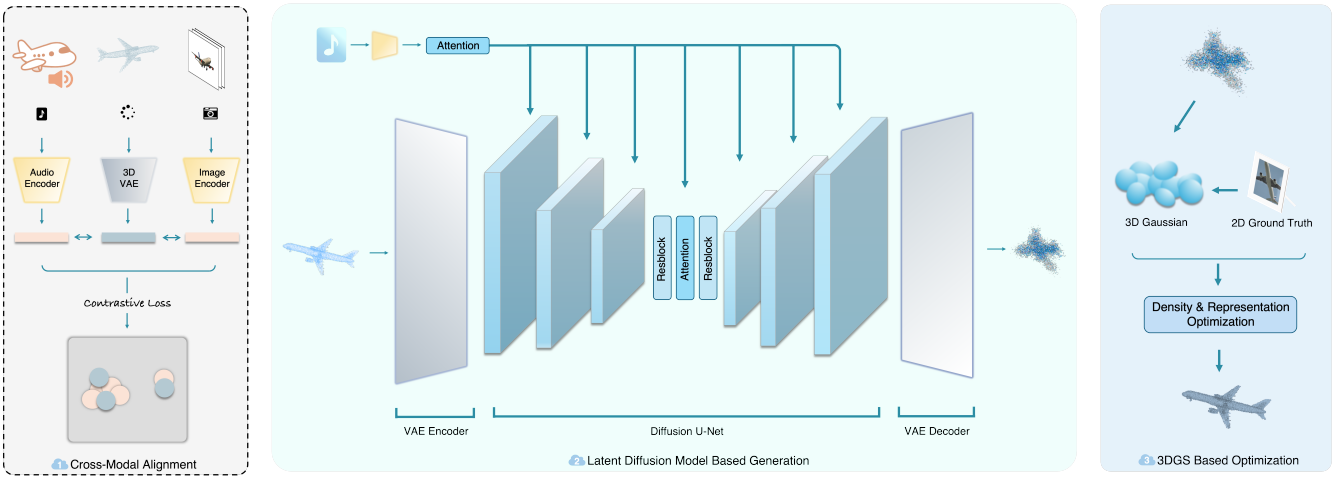

Fig. 2: Illustration of our audio-to-3D generation framework, which comprises three core components operating sequentially: (1) a cross-modal alignment module, (2) a latent diffusion model conditioned on aligned embeddings to generate initial 3D structures, and (3) a 3D Gaussian Splatting refinement stage that enhances the geometric details and fidelity of the generated shapes.

the audio domain $\mathscr{A}$ and a corresponding 3D shape s from the shape domain $\mathscr{S}$, our objective is to learn a generator $g : \mathscr{A} \to \mathscr{S}$ such that $g(a) = s$. This enables reconstruction of 3D geometry conditioned on acoustic cues.

To this end, we propose a novel architecture for audio-conditioned 3D point cloud generation that jointly captures semantic and spatial information across modalities. As shown in Figure 2, our approach integrates three main modules: cross-modal alignment, latent diffusion model based 3D shape generation, and 3DGS based refinement. First, we perform tri-modal alignment by mapping audio, image, and 3D shape representations into a unified embedding space via contrastive learning. In the second stage, conditioned on the aligned audio features, a latent diffusion model generates coarse 3D shapes utilizing a cross-attention mechanism that incorporates auditory cues during synthesis. Finally, to enhance geometric fidelity, we employ 3D Gaussian Splatting to refine the coarse shapes into high quality 3D structures with realistic surfaces and spatial continuity.

A. *Cross-modal Alignment*

To construct a semantically rich and robust embedding space, we propose a cross-modal alignment network consisting of three modality-specific encoders. The objective is to strengthen semantic alignment across modalities, with particular emphasis on aligning point clouds with the semantic space derived from audio and images.

The first step, termed Audio-Visual Alignment, aligns audio and visual semantics via contrastive learning. Specifically, we use Wav2CLIP [24] audio encoder and CLIP[9] image encoder to project audio inputs and images into a shared embedding space. Although these encoders have already been pretrained with some degree of alignment, we apply an additional contrastive objective to further refine this alignment, pulling semantically related cross-modal pairs closer. This unified space serves as the foundation for incorporating 3D point cloud representations.

Next, we introduce a PointNet-style[49] encoder enhanced with residual blocks to extract expressive per-point features. These features are aggregated via global pooling and passed through a lightweight MLP to produce an embedding suitable for contrastive learning. To align 3D geometry with the unified semantic space established in the first step, we adopt a teacher–student distillation framework: the pre-aligned audio and image encoders serve as teachers to guide the training of the point cloud encoder.

A tri-modal contrastive loss jointly optimizes the alignment among audio, image, and point cloud features, encouraging the point cloud encoder to map geometric structures into the shared semantic space.

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{audio-visual}} + \mathscr{L}_{\text{audio-pc}} + \mathscr{L}_{\text{visual-pc}} \tag{1}$$

This alignment enables the model to capture the semantic content of audio inputs in 3D form, producing latent embeddings that can be effectively leveraged through the generation pipeline.

B. *Latent Diffusion Model based 3D Point Cloud Generation*

After establishing cross-modal alignment between audio and point cloud, we introduce a latent diffusion model (LDM) built upon a Point Cloud Vector-Quantized Variational Autoencoder (PC-VQ-VAE) to enable audio-driven 3D shape generation.

Point Cloud Vector-Quantized Variational Autoencoder. Slow sampling speed and difficulty in generating well-structured shapes remain major challenges for 3D diffusion models. To address these limitations, we introduce a Point Cloud Vector-Quantized Variational Autoencoder (PC-VQ-VAE) as a prior learning module. This module learns compact and semantically meaningful latent representations

that serve as initialization for the diffusion process and as structural constraints to guide generation.

Specifically, our PC-VQ-VAE builds upon the point cloud encoder introduced in the previous alignment stage. However, unlike the alignment encoder, we omit pooling operations and retain per-point features to preserve geometric details. This design ensures that the quantized latent codes capture rich local structures essential for high-quality 3D generation. The raw point cloud $\mathbf{x} \in \mathbb{R}^{N\times 3}$ is processed to extract discriminative geometric features, which are then passed through an encoder to produce continuous latent representations.

$$\mathbf{z}_e = \mathrm{Enc}(\mathbf{x}) \in \mathbb{R}^d \tag{2}$$

These latent features are subsequently quantized by mapping them to the closest entries in a learned codebook $\mathscr{E} = \{\mathbf{e}_k \in \mathbb{R}^d\}_{k=1}^K$ of discrete latent vectors:

$$\mathbf{z} = \mathrm{Quantize}(\mathbf{z}_e, \mathscr{E}) \tag{3}$$

This quantization step enforces the model to learn a finite set of representative geometric patterns, forming a discrete and semantically meaningful latent space.

Diffusion Process. To model the complete audio-driven generation pipeline, we adopt a two-stage framework: a forward diffusion process that progressively adds noise, and a reverse denoising process guided by audio conditions. In the forward stage, a fixed Markov forward process gradually corrupts the latent representation of the point cloud $\mathbf{z} \in \mathbb{R}^d$ over T timesteps by injecting Gaussian noise:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathscr{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}\right) \tag{4}$$

This process transforms the clean data into an approximately Gaussian distribution, and the resulting noised samples are used as training inputs, enabling the model to learn to recover the original data from different noise levels.

In the denoising stage, we design a conditional UNet-style [50] architecture. This network employs a hierarchical downsampling–upsampling structure with attention modules. This design facilitates multi-scale feature extraction, long-range dependency modeling, and high-fidelity spatial detail reconstruction.

Specifically, the audio condition vector is injected into our UNet-style model via cross-attention modules, acting as a bridge between semantic information and the synthesized 3D structure. This is realized through a Transformer-based framework: it first captures long-range dependencies among point features via self-attention, then progressively incorporates audio semantic cues through cross-attention, enabling audio-conditioned generation while preserving fundamental geometric structure. For a query latent feature $\mathbf{Q}$ derived from the noisy point representation $\mathbf{z}_t$ and audio-derived key–value pairs $(\mathbf{K}_a, \mathbf{V}_a)$, the cross-attention output is computed as:

$$\mathrm{CrossAttn}(\mathbf{Q}, \mathbf{K}_a, \mathbf{V}_a) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_a^\top}{\sqrt{d_k}}\right)\mathbf{V}_a \tag{5}$$

The resulting audio-conditioned representation guides the entire generation process.

Next, our UNet-style denoising network $\varepsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_a)$ predicts the injected noise from the noisy latent $\mathbf{z}_t$, the current timestep t, and condition vector $\mathbf{c}_a$ obtained via cross-attention with audio features. The training objective minimizes the discrepancy between the true noise ε and the predicted noise, formulated as a simplified variational lower bound:

$$\mathscr{L}_{\mathrm{diff}} = \mathbb{E}_{\mathbf{z}_t, \varepsilon, t}\left[\left\|\varepsilon - \varepsilon_\theta\left(\mathbf{z}_t, t, \mathrm{CrossAttn}(\mathbf{z}_t, \mathbf{c}_a, \mathbf{c}_a)\right)\right\|_2^2\right] \tag{6}$$

During sampling, generation begins from a standard Gaussian noise vector and proceeds through a series of reverse diffusion steps. At each step, the trained UNet-style model predicts either the noise component or a clean estimate of the current noisy sample. Based on this prediction, the model performs a partial denoising update that reduces noise while preserving structural cues encoded in the latent space. By iterating this process across all timesteps, from the most noisy state back to the clean domain, the model gradually reconstructs a coherent 3D point cloud.

Throughout the reverse process, audio-conditioned guidance is applied at every iteration. This conditioning steers each denoising update toward regions of the latent space that are semantically consistent with the provided audio input, enabling progressive refinement and ensuring that the synthesized point cloud aligns well with the conditioning audio.

VAE Decoder. We develop a hierarchical decoder as the counterpart to the PC-VQ-VAE, leveraging progressive upsampling and attention-driven feature enhancement to enable effective 3D shape reconstruction from latent codes. The decoder begins with a latent projection layer, followed by upsampling blocks that simultaneously increase point resolution and refine feature representations. Spatial consistency and global geometric awareness are further reinforced by a multi-head self-attention module. Finally, a lightweight convolutional head predicts the 3D coordinates, yielding point cloud reconstructions. By reconstructing complete 3D structures from quantized latent codes, the decoder captures both local geometry and global structure, effectively bridging the gap between the discrete latent space and the continuous geometric domain.

C. 3DGS based Optimization

High-fidelity 3D reconstruction is achieved by refining an initially generated point cloud using a single reference image. Instead of relying on full 3D supervision or multi-view constraints, our pipeline utilizes 2D image observations to improve both the geometry and appearance of the generated shape. Importantly, the optimization process enhances geometric fidelity and adds missing high-frequency details without distorting the underlying coarse structure provided by the generative model.

Given a single input image, we first estimate the corresponding camera intrinsics and extrinsics using a structure-from-motion initialization strategy. Although only one view

is available, this estimation yields a valid camera pose for projecting and rendering the generated point cloud under the 3D Gaussian Splatting (3DGS) representation. Each point is represented as a Gaussian primitive parameterized by its mean $\mu_i \in \mathbb{R}^3$, covariance matrix $\Sigma_i \in \mathbb{R}^{3\times3}$, opacity α_i, and color $\mathbf{c}_i$.

Optimization is performed entirely in image space by minimizing the discrepancy between rendered views of the Gaussian model $\mathscr{G}$ and the reference image. The renderer synthesizes image predictions $\hat{I}$, and the reconstruction loss is formulated as:

$$\mathscr{L}_{\text{render}} = \left\| \hat{I} - I_{\text{gt}} \right\|_1$$

where I_{gt} is the target image. This image-space supervision implicitly guides the Gaussian primitives to refine geometry, color, and surface consistency without requiring explicit 3D ground truth. Since only visible regions are constrained, the coarse geometry provided by the generative model is preserved, while the optimization focuses on adding missing details and correcting local inconsistencies.

To further enhance fidelity without introducing redundant complexity, we employ adaptive density control. Regions with large residual errors in the rendered image trigger Gaussian duplication or subdivision, enhancing representational precision. Conversely, primitives that contribute little to the rendering are pruned to prevent over-parameterization. Through iterative refinement, the optimized Gaussians are transformed into a dense, high-quality point cloud that preserves the original generative structure while exhibiting significantly improved continuity, sharpness, and geometric detail.

IV. Experiment

TABLE I: Quantitative performance comparison on 3D cross-modal retrieval tasks. We report retrieval metrics including R@1, R@5, and mAP across multiple methods.

Metrics	Tasks			
	Audio-3D	3D-Audio	2D-3D	3D-2D
mAP	80.11%	82.73%	67.88%	73.36%
recall@1	49.07%	54.28%	34.64%	39.44%
recall@5	89.25%	91.47%	73.22%	77.19%
recall@10	96.05%	96.14%	83.71%	86.97%
0-shot mAP	20.12%	17.89%	23.34%	25.78%

A. Experimental Setup

Implementation Details. Our training pipeline consists of three key stages. First, we perform cross-modal pretraining using a tri-encoder architecture. This stage incorporates a frozen CLIP image encoder (ViT-B/32), a frozen Wav2CLIP audio encoder, and a PointNet-based point cloud encoder. The model is trained for 100 epochs with a contrastive loss, optimized using Adam (learning rate: 5e-4, weight decay: 1e-4, batch size: 64). Next, we optionally enhance point cloud representations by pretraining a Variational Autoencoder (VAE) with a latent dimension of 256. This VAE is optimized using the Chamfer Distance loss over 200 epochs. The final stage involves training a UNet-based diffusion model that processes inputs of B×512×2048. The diffusion model is trained for 200 epochs using the AdamW optimizer (learning rate: 1e-4, weight decay: 1e-2, batch size: 16), minimizing an L2 noise prediction loss with a multi-segment Lambda learning rate scheduler. All experiments were conducted on a single NVIDIA RTX 4090 GPU.

Cross Modal Retrieval. To evaluate the multi-modal alignment of our model, we conduct cross-modal retrieval tasks, including point cloud-to-image, image-to-point cloud, and audio-to-point cloud retrieval. Each modality is encoded separately and embedded into a shared space for retrieval. To assess generalization, we also perform zero-shot retrieval on unseen categories, demonstrating the model's ability to align modalities without category-specific supervision.

Reconstruction Analysis. Reconstruction capability is evaluated by passing input point clouds through the encoder-decoder pipeline and comparing the outputs with the original shapes. Experiments are conducted on both our curated multimodal dataset and the widely used ModelNet40 benchmark. Additionally, we perform latent space analysis by interpolating between latent codes and visualizing the resulting transitions in the reconstructed outputs.

TABLE II: Comparison of PC-VQ-VAE and PC-VAE for 3D representation learning.

Method	Class	Performance Metrics			
		CD ↓	EMD ↓	COV ↑	MMD ↓
PC-VAE	Violin	0.0597	0.1905	0.2251	0.1130
	Sax	0.0654	0.1920	0.3301	0.0705
	Harp	0.0498	0.1623	0.2998	0.1096
PC-VQ-VAE	Violin	0.0459	0.0898	0.2544	0.0384
	Sax	0.0468	0.0782	0.3853	0.0211
	Harp	0.0445	0.0731	0.4741	0.0107

Conditional Generation. We further evaluate the conditional generation capability of our model, focusing on audio-conditioned 3D shape synthesis. Experiments are conducted on our multimodal dataset, where each sample contains paired audio signal and 3D point cloud. To verify the semantic consistency between input audio and the generated 3D shape, we perform an additional audio-to-generated point cloud retrieval experiment. Using the pretrained joint embedding space, we treat the generated point clouds as the retrieval database and assess whether the corresponding audio queries can successfully retrieve their associated shape.

B. Metrics

We evaluate model performance using a comprehensive set of task-specific metrics tailored to different objectives. For cross-modal retrieval, we adopt mean Average Precision (mAP) and Recall@k to measure retrieval accuracy and top-ranked semantic matching. For 3D reconstruction, we assess geometric fidelity and distribution consistency using widely adopted point cloud metrics, including Chamfer Distance

Fig. 3: Qualitative visualization of audio-driven point cloud generation. Each row presents three samples, illustrating the model's capability to produce diverse and geometrically plausible geometries from acoustic inputs.

TABLE III: Performance on semantic consistency retrieval after generation,including audio-3D, 3D-audio, image-3D, 3d-image retrieval

Metrics	Tasks			
	Audio-3D	3D-Audio	2D-3D	3D-2D
mAP	73.94%	72.12%	71.08%	69.55%
recall@1	53.87%	47.92%	33.73%	36.52%

(CD), Earth Mover's Distance (EMD), Coverage (COV), and Minimum Matching Distance (MMD). CD and EMD capture point-wise geometric errors, while COV and MMD jointly evaluate how well the reconstructed distribution matches the real data distribution.

In ablation studies, we further examine reconstruction quality and generative sharpness by reporting Precision, average 1-Nearest Neighbor (1-NN) accuracy, and the L1 error of rendered images. Precision and 1-NN reflect local geometric correctness and point-wise fidelity, while the image-based L1 metric evaluates perceptual consistency between rendered views and ground-truth geometry.

C. Quantitative Analysis

Cross Modal Retrieval. We report quantitative results on multimodal retrieval tasks using our proposed dataset in Table I. Our approach demonstrates robust performance, achieving high mean Average Precision (mAP) scores in both conventional and novel retrieval scenarios. To the best of our knowledge, this is the first study to investigate cross-modal retrieval between audio signals and 3D point clouds. The consistently strong results across all four tasks underscore the model's remarkable generalization capability and effectiveness in zero-shot cross-modal retrieval.

Reconstruction Analysis. As shown in Table II, our model achieves strong reconstruction performance on the test set. These results indicate that our approach effectively preserves both the global structure and fine-grained geometric details of 3D shapes. Specifically, the low Chamfer Distance (CD) and Earth Mover's Distance (EMD) values reflect accurate spatial alignment, the high F-Score indicates strong point-wise correspondence with the ground truth, and the low 1-Nearest Neighbor (1-NN) average distance further confirms the local consistency between reconstructed and reference point clouds.

Semantic Consistency Retrieval. As shown in Table III, our model demonstrates strong performance, indicating that the generated shapes effectively retain meaningful audio-related semantics. Compared against both an upper bound established using real shapes and a lower bound based on random generations, our results confirm that audio-guided generation produces semantically aligned 3D outputs.

D. Qualitative Analysis

Conditional Generation. Figure 3 showcases compelling examples of audio-to-3D generation produced by our framework. Leveraging its robust generative capabilities, our approach takes audio inputs and synthesizes highly coherent and richly detailed 3D point clouds that faithfully reflect the semantic content of the audio. These results not only demonstrate the framework's ability to translate acoustic information into vivid 3D structures, but also highlight the strong alignment between audio features and 3D shape representations within the learned embedding space—validating the effectiveness of our generative mechanism in bridging these two distinct modalities.

E. Ablation Study

We perform ablation studies to analyze the contribution of each component. As shown in Table IV, our results highlight the importance of each module in achieving high-quality generation. Removing the cross-modal alignment stage leads

TABLE IV: Ablation study on the design choices of our architecture.

Model Configuration	Performance Metrics					
	EMD ↓	Precision ↑	L1 ↓	1-NN ↓	COV ↑	MMD ↓
W/O Cross Modal Alignment	0.9546	0.0201	0.1091	0.2778	0.0210	0.9594
W/O Attention Block	0.7966	0.1189	0.0972	0.4526	0.1077	0.6753
W/O Refinement Stage	0.4995	0.2961	0.0562	0.2185	0.7140	0.0368
Full Model	**0.4390**	**0.7910**	**0.0032**	**0.2284**	**0.8969**	**0.0211**

to semantically inconsistent shape generation, as the model fails to establish meaningful associations between audio and 3D structures. The attention block responsible for injecting audio conditions is equally critical. Its removal results in a distinct semantic discrepancy between the audio input and the generated shapes. The refinement stage also plays a key role in enhancing output quality; removing it leads to a noticeable loss of geometric detail and an overall decline in fidelity.

F. Limitations

Although our method demonstrates promising results, it struggles with highly complex shapes, often producing noisy outputs or structural artifacts. It may also exhibit category ambiguity when processing acoustically similar inputs from different categories. Addressing these limitations will be a key focus of our future work.

V. Conclusion

In this work, we present an audio-driven generative framework that unifies auditory, visual, and geometric modalities for high-quality 3D point cloud synthesis. By integrating contrastive alignment, latent diffusion generation, and differentiable Gaussian-based refinement, our model produces 3D outputs that are both semantically coherent and structurally faithful, conditioned on audio cues. While the framework demonstrates strong performance, it faces challenges with highly complex geometries or acoustically similar categories, which can occasionally result in structural artifacts or semantic ambiguities. Our work advances static audio-conditioned 3D reconstruction and provides a foundation for future research into dynamic generation, multimodal control, and real-world applications.

References

[1] C. Akkus, L. Chu, V. Djakovic, S. Jauch-Walser, P. Koch, G. Loss, C. Marquardt, M. Moldovan, N. Sauter, M. Schneider, *et al.*, "Multimodal deep learning," *arXiv preprint arXiv:2301.04856*, 2023.

[2] E. Tzinis, S. Wisdom, T. Remez, and J. R. Hershey, "Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation," in *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 2022, pp. 368–385.

[3] J. Liang, X. Liu, W. Wang, M. D. Plumbley, H. Phan, and E. Benetos, "Acoustic prompt tuning: Empowering large language models with audition capabilities," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1–15, 2025.

[4] A. Szot, B. Mazoure, O. Attia, A. Timofeev, H. Agrawal, D. Hjelm, Z. Gan, Z. Kira, and A. Toshev, "From multimodal llms to generalist embodied agents: Methods and lessons," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 10 644–10 655.

[5] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, "Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 15 768–15 780.

[6] Z. Chen, J. Tang, Y. Dong, Z. Cao, F. Hong, Y. Lan, T. Wang, H. Xie, T. Wu, S. Saito, L. Pan, D. Lin, and Z. Liu, "3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 26 576–26 586.

[7] Y. Chen, J. Zhang, Z. Xie, W. Li, F. Zhang, J. Lu, and L. Zhang, "S-nerf++: Autonomous driving simulation via neural reconstruction and generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4358–4376, 2025.

[8] C. Vachha, Y. Kang, Z. Dive, A. Chidambaram, A. Gupta, E. Jun, and B. Hartmann, "Dreamcrafter: Immersive editing of 3d radiance fields through flexible, generative inputs and outputs," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–13.

[9] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[10] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8552–8562.

[11] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9902–9912.

[12] J. Yang, A. Sax, K. J. Liang, M. Henaff, H. Tang, A. Cao, J. Chai, F. Meier, and M. Feiszli, "Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 21 924–21 935.

[13] D. Di, J. Yang, C. Luo, Z. Xue, W. Chen, X. Yang, and Y. Gao, "Hyper-3dg: Text-to-3d gaussian generation via hypergraph," *International Journal of Computer Vision*, vol. 133, no. 5, pp. 2886–2909, 2025.

[14] M. Wang, S. Yuan, X.-F. Han, and Z. Yi, "Draw what you hear: High-fidelity image generation and manipulation via soundadapter," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[15] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 180–15 190.

[16] W. Lei, Y. Ge, K. Yi, J. Zhang, D. Gao, D. Sun, Y. Ge, Y. Shan, and M. Z. Shou, "Vit-lens: Towards omni-modal representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 647–26 657.

[17] M. Shukor, C. Dancette, A. Rame, and M. Cord, "Unival: Unified model for image, video, audio and language tasks," *Transactions on Machine Learning Research Journal*, 2023.

[18] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023.

[19] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 515–16 525.

[20] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhahn, "Text to image generation with semantic-spatial aware gan," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18 187–18 196.

[21] Z. Wang, Z. Sha, Z. Ding, Y. Wang, and Z. Tu, "Tokencompose: Text-to-image diffusion with token-level supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8553–8564.

[22] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and B. Cui, "Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024, pp. 1–15.

[23] M. Yi, A. Li, Y. Xin, and Z. Li, "Towards understanding the working mechanism of text-to-image diffusion models," in *Advances in Neural Information Processing Systems(NeurIPS)*, vol. 37, 2024, pp. 55 342–55 369.

[24] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4563–4567.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[26] S.-B. Kim, A. Senocak, H. Ha, A. Owens, and T.-H. Oh, "Sound to visual scene generation by audio-to-visual latent alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6430–6440.

[27] S. Li, K. Kallidromitis, A. Gokul, Z. Liao, Y. Kato, K. Kozuka, and A. Grover, "Omniflow: Any-to-any generation with multi-modal rectified flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 13 178–13 188.

[28] G. Lin, J. Jiang, C. Liang, T. Zhong, J. Yang, Z. Zheng, and Y. Zheng, "Cyberhost: A one-stage diffusion framework for audio-driven talking body generation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[29] G. Yariv, I. Gat, S. Benaim, L. Wolf, I. Schwartz, and Y. Adi, "Diverse and aligned audio-to-video generation via text-to-video model adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6639–6647.

[30] M. Sun, W. Wang, Y. Qiao, J. Sun, Z. Qin, L. Guo, X. Zhu, and J. Liu, "Mm-ldm: Multi-modal latent diffusion model for sounding video generation," in *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM '24)*, 2024, pp. 10 853–10 861.

[31] Z. Xie, Q. He, Y. Zhu, Q. He, and M. Li, "Filmcomposer: Llm-driven music production for silent film clips," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 13 519–13 528.

[32] K. E. Smith and A. O. Smith, "Conditional gan for timeseries generation," *arXiv preprint arXiv:2006.16477*, 2020.

[33] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[34] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[35] B. C. Biner, F. M. Sofian, U. B. Karakaş, D. Ceylan, E. Erdem, and A. Erdem, "Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models," *arXiv preprint arXiv:2405.00878*, 2024.

[36] M. T. Islam and S. Nirjon, "Sound-adapter: Multi-source domain adaptation for acoustic classification through domain discovery," in *Proceedings of the 20th International Conference on Information Processing in Sensor Networks*, 2021, pp. 176–190.

[37] J. Li, C. Li, Y. Wu, and Y. Qian, "Unified cross-modal attention: Robust audio-visual speech recognition and beyond," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1941–1953, 2024.

[38] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3d point clouds," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 40–49.

[39] Y. Sun, Y. Wang, Z. Liu, J. Siegel, and S. Sarma, "Pointgrow: Autoregressively learned point cloud generation with self-attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 61–70.

[40] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[41] Y. Feng, X. Shi, M. Cheng, and Y. Xiong, "Diffpoint: Single and multi-view point cloud reconstruction with vit based diffusion model," *arXiv preprint arXiv:2402.11241*, 2024.

[42] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, pp. 99–106, 2021.

[43] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-nerf: Point-based neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5438–5448.

[44] J. Zhang, F. Zhang, S. Kuang, and L. Zhang, "Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7178–7186.

[45] M. Rafidashti, J. Lan, M. Fatemi, J. Fu, L. Hammarstrand, and L. Svensson, "Neuradar: Neural radiance fields for automotive radar point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 2488–2498.

[46] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, pp. 139:1–139:14, 2023.

[47] C. Liu, S. Chen, Y. S. Bhalgat, S. HU, M. Cheng, Z. Wang, V. A. Prisacariu, and T. Braud, "Gs-cpr: Efficient camera pose refinement via 3d gaussian splatting," in *The Thirteenth International Conference on Learning Representations*, 2024.

[48] P. Zheng, L. Wei, D. Jiang, and J. Zhang, "3d gaussian splatting against moving objects for high-fidelity street scene reconstruction," *arXiv preprint arXiv:2503.12001*, 2025.

[49] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.