PHYSICS-AWARE NOVEL-VIEW ACOUSTIC SYNTHESIS WITH VISION-LANGUAGE PRIORS AND 3D ACOUSTIC ENVIRONMENT MODELING

Congyi Fan[1], *Jian Guan*[1,*], *Youtian Lin*[2], *Dongli Xu*[3], *Tong Ye*[1], *Qiaoxi Zhu*[4], *Pengming Feng*[5], *Wenwu Wang*[6]

[1] Group of Intelligent Signal Processing, Harbin Engineering University, Harbin, China
[2] School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[3] Processing Speech and Images, KU Leuven, Leuven, Belgium
[4] Acoustics Lab, University of Technology Sydney, Ultimo, Australia
[5] State Key Laboratory of Space Information System and Integrated Application, Beijing, China
[6] Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK

ABSTRACT

Spatial audio is essential for immersive experiences, yet novel-view acoustic synthesis (NVAS) remains challenging due to complex physical phenomena such as reflection, diffraction, and material absorption. Existing methods based on single-view or panoramic inputs improve spatial fidelity but fail to capture global geometry and semantic cues such as object layout and material properties. To address this, we propose Phys-NVAS, the first physics-aware NVAS framework that integrates spatial geometry modeling with vision–language semantic priors. A global 3D acoustic environment is reconstructed from multi-view images and depth maps to estimate room size and shape, enhancing spatial awareness of sound propagation. Meanwhile, a vision–language model extracts physics-aware priors of objects, layouts, and materials, capturing absorption and reflection beyond geometry. An acoustic feature fusion adapter unifies these cues into a physics-aware representation for binaural generation. Experiments on RWAVS demonstrate that Phys-NVAS yields binaural audio with improved realism and physical consistency.

Index Terms— Novel-view acoustic synthesis, physics-aware feature representation, vision-language priors

1. INTRODUCTION

Spatial audio, conveying sound position, direction, and distance in 3D space, is essential for immersive applications such as augmented/virtual reality (AR/VR), gaming, and interactive media [1, 2, 3]. A fundamental task is novel-view acoustic synthesis (NVAS), which generates binaural audio for arbitrary listener positions given a mono input and scene observations [4]. Realistic NVAS requires accurate modeling of sound propagation, governed by direct sound, early reflections, and reverberation, all influenced by scene geometry, object layout, and materials [5].

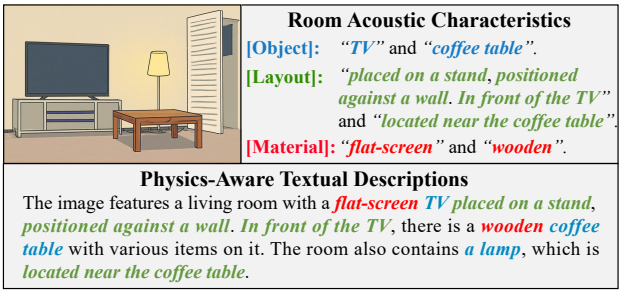

Fig. 1. Illustration of scene semantics influencing acoustics. Different materials (e.g., *wooden*, *flat-screen*) affect absorption, while objects and their layouts (e.g., *In front of the TV, there is a wooden coffee table*) modify reflection paths. These semantic cues are often ignored in existing NVAS methods, leading to physically inconsistent audio.

The main challenge is that these factors jointly produce complex effects such as reflection, diffraction, and absorption; without explicitly modeling them, synthesized audio lacks spatial realism and physical consistency. Recent works thus leverage visual cues, motivated by the intuition that a scene's appearance provides critical information for acoustics [4, 6, 7, 8, 9].

AV-NeRF [4] first conditioned a neural acoustic field on single-view images and depth, establishing a strong baseline on the Real-World Audio-Visual Scene (RWAVS) dataset. However, its reliance on listener-centric single views limited spatial awareness, and it did not explicitly incorporate semantic cues such as objects, layouts, and materials. Subsequent methods further enriched vision priors. For example, SOAF [6] modeled occlusion, AV-GS [7] exploited 3D Gaussian Splatting (3DGS) [10] for geometric fidelity, AV-Surf [8]

*Corresponding author: j.guan@hrbeu.edu.cn

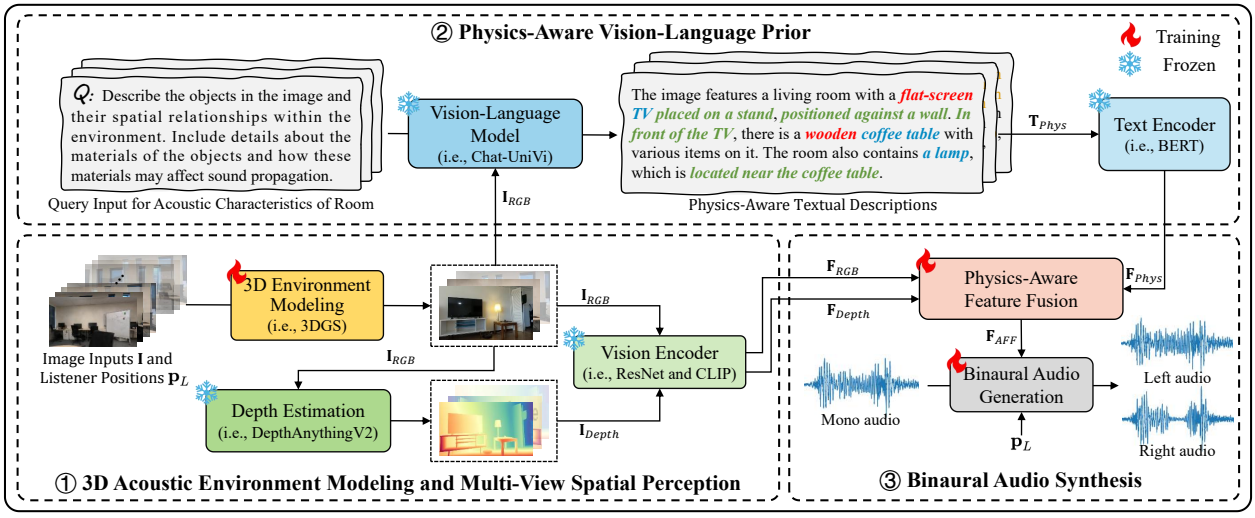

Fig. 2. Overview of the proposed physics-aware NVAS framework. 3D acoustic environment modeling with 3DGS and depth estimation on multi-view images, enhancing spatial awareness by recovering room geometry and size. Physics-aware vision–language priors further enrich acoustic modeling with object, layout, and material cues that capture absorption and reflection effects. Finally, geometric and semantic features are fused into a unified physics-aware feature representation, enabling realistic and physically consistent binaural audio generation.

refined structural details with surface normals, and SoundVista [9] introduced panoramic context. While these methods improved geometry and occlusion handling, ***they overlooked how object layout and material properties affect absorption and reflection.*** For instance, geometry-based models cannot distinguish the dampening of carpet versus tile, while panoramic inputs are weak in object-level semantics. As illustrated in Fig. 1, materials such as *"wooden"* or *"flat-screen"*, and layouts such as *"In front of the TV"*, directly alter acoustic behavior. Ignoring such factors often leads to physically inconsistent audio synthesis.

In this paper, we propose Phys-NVAS, the first NVAS framework to incorporate physics-aware vision-language priors into scene acoustic modeling. Phys-NVAS integrates 3D acoustic environment reconstruction with physics-aware vision semantics to achieve more realistic spatial audio synthesis. Specifically, a global 3D acoustic scene is reconstructed with 3D Gaussian Splatting (3DGS) and depth estimation, enabling multi-view spatial perception and providing explicit structural cues on room geometry and size that guide direct sound and early reflections. Meanwhile, a vision-language model is employed to extract physics-aware priors describing objects, layouts, and materials, capturing absorption and reflection effects beyond geometry alone. We further propose an acoustic feature fusion adapter that integrates these geometric and semantic cues into a unified physics-aware representation for binaural generation. Experiments demonstrate that combining spatial geometry with semantic information yields more realistic and physically consistent novel-view acoustic synthesis.

2. PROPOSED METHOD

This section presents the proposed Phys-NVAS framework, illustrated in Fig. 2. First, 3D acoustic environment modeling with 3DGS and depth estimation recovers global scene geometry and size, enhancing spatial perception via multi-view inputs. Second, a vision-language model extracts physics-aware priors such as objects, layouts, and material properties, providing semantic cues for absorption and reflection beyond geometry. Finally, an acoustic feature fusion adapter unifies geometric and semantic features into a physics-aware representation, enabling realistic and physically consistent binaural synthesis.

2.1. 3D Acoustic Environment Modeling and Multi-View Spatial Perception

3D Environment Modeling: To capture the global geometry of the acoustic environment, we generate multi-view RGB images $\mathbf{I}_{RGB}$ from a few image inputs $\mathbf{I}$ and the listener position $\mathbf{p}_L \in \mathbb{R}^5$, with 3D Gaussian Splatting (3DGS) [10] as image generator $\mathcal{G}(\cdot)$:

$$\mathbf{I}_{RGB} = \mathcal{G}(\mathbf{I}, \mathbf{p}_L). \tag{1}$$

We then apply a pre-trained vision encoder (i.e., ResNet-18 [11] and CLIP [12]) to extract geometric features $\mathbf{F}_{RGB} \in \mathbb{R}^M$ from $\mathbf{I}_{RGB}$, where M denotes the feature dimension. Leveraging multiple viewpoints, $\mathbf{F}_{RGB}$ encodes global shape information of the scene that is relevant to spatial sound propagation.

Multi-View Depth Perception: While RGB features describe global structure, they lack explicit geometric cues such as dis-

tances between listener and objects. To address this, we apply a depth estimation model $\mathcal{D}(\cdot)$ (i.e., DepthAnythingV2 [13]) to generate multi-view depth maps $\mathbf{I}_{Depth}$ as follows:

$$\mathbf{I}_{Depth} = \mathcal{D}(\mathbf{I}_{RGB}). \tag{2}$$

The depth maps are then encoded into structural features $\mathbf{F}_{Depth} \in \mathbb{R}^M$. With the estimated distances between the listener and surrounding objects, $\mathbf{F}_{Depth}$ provides explicit cues about room size and shape, thereby complementing $\mathbf{F}_{RGB}$ for modeling direct sound propagation and early reflections in the acoustic environment.

2.2. Physics-Aware Vision-Language Priors

While multi-view geometry provides global structural cues, it fails to capture semantic properties such as objects, layouts, and materials that critically affect sound absorption and reflection. To address this limitation, we introduce a physics-aware vision-language prior that extracts semantic descriptions from the rendered multi-view images.

Specifically, we employ a vision-language model (VLM), i.e., Chat-UniVi [14], to generate textual descriptions $\mathbf{T}_{Phys}$ containing object identities, spatial layout, and material attributes from $\mathbf{I}_{RGB}$. These physics-aware textual priors explicitly describe scene configurations relevant to acoustic modeling (e.g., "*a flat-screen TV*", "*a wooden coffee table*", and their relative positions). To obtain such descriptions, we provide the VLM with a fixed query Q, which instructs the model to identify objects, layouts, and materials, as illustrated in the Physics-Aware Vision-Language Priors block of Fig. 2. Formally, this process can be expressed as:

$$\mathbf{T}_{Phys} = \text{VLM}(\mathbf{I}_{RGB}, Q). \tag{3}$$

The generated $\mathbf{T}_{Phys}$ is then encoded by a pre-trained text encoder $\mathcal{T}(\cdot)$ (i.e., BERT [15]) to obtain semantic feature $\mathbf{F}_{Phys} \in \mathbb{R}^M$, as follows:

$$\mathbf{F}_{Phys} = \mathcal{T}(\mathbf{T}_{Phys}). \tag{4}$$

which provides physics-aware cues about material-dependent absorption and reflection, as well as object-level layout, thereby complementing the geometric features $\mathbf{F}_{RGB}$ and $\mathbf{F}_{Depth}$ in acoustic environment modeling.

2.3. Binaural Audio Synthesis with Physics-Aware Feature Representation

Physics-Aware Feature Fusion: To obtain a unified representation of the acoustic environment, we propose an acoustic feature fusion adapter $\mathcal{A}(\cdot)$ that integrates geometric and semantic features. Specifically, the adapter fuses the multi-view geometric features $\mathbf{F}_{RGB}$ and $\mathbf{F}_{Depth}$ with the physics-aware semantic features $\mathbf{F}_{Phys}$:

$$\mathbf{F}_{AFF} = \mathcal{A}(\mathbf{F}_{RGB}, \mathbf{F}_{Depth}, \mathbf{F}_{Phys}), \tag{5}$$

where $\mathbf{F}_{AFF}$ denotes the fused physics-aware feature representation. Concretely, we concatenate the geometric features $\mathbf{F}_{RGB}$ and $\mathbf{F}_{Depth}$ and feed them into a multi-layer perceptron (MLP) to extract a geometric embedding, while the semantic features $\mathbf{F}_{Phys}$ are fed into another MLP to extract a semantic embedding. The outputs of these two MLPs are then added together to obtain the final fused feature $\mathbf{F}_{AFF}$. This unified embedding jointly encodes room size, shape, object layout, and material properties, providing a physics-aware representation of the acoustic environment.

Binaural Audio Generation: Finally, we employ the binaural audio generator $\mathcal{B}(\cdot)$ from AV-NeRF [4], conditioned on the mono audio input $\mathbf{a}_{mono}$, the listener position $\mathbf{p}_L$, and the fused representation $\mathbf{F}_{AFF}$. The binaural signals are generated as:

$$\mathbf{a}_{bi} = \mathcal{B}\left(\mathbf{a}_{mono} \mid \mathbf{F}_{AFF}, \mathbf{p}_L\right), \tag{6}$$

where $\mathbf{a}_{bi} = \{\mathbf{a}_l, \mathbf{a}_r\}$ denotes the synthesized left and right audio channels. This design enables Phys-NVAS to generate binaural audio that reflects both geometric and semantic acoustic cues, leading to improved spatial realism and physical consistency.

3. EXPERIMENTS AND RESULTS

3.1. Experiment Setup

Dataset: We evaluate our method on the RWAVS dataset [4], which provides multimodal samples including camera poses, images, and high-quality binaural audio. The dataset covers four scenes (office, house, apartment, and outdoor), with recordings of 10–25 minutes sampled at 1 fps. Each frame is paired with one-second binaural and source audio, forming a complete data sample. Following AV-NeRF [4], we adopt the official 80/20 split, resulting in 9850 training and 2469 validation samples after pre-processing.

Evaluation Metrics: We follow AV-NeRF and report two widely used metrics for spatial audio, i.e., magnitude distance (MAG) [19], which measures spectral amplitude differences, and envelope distance (ENV) [20], which measures envelope structure differences. Lower scores indicate better perceptual alignment with the ground truth.

Baseline Methods: To validate the effectiveness of Phys-NVAS, we compare it with both signal-based and learning-based methods. The signal-based methods include *Mono-Mono* (duplicating mono to both channels), *Mono-Energy* (scaling mono by average energy), and *Stereo-Energy* (constructing stereo from known energy priors). The learning-based methods are INRAS [16], NAF [17], ViGAS [18], and AV-NeRF [4]. As Phys-NVAS is the first to explicitly incorporate physics-aware vision-language priors, our experiments emphasize validating its effectiveness rather than exhaustively benchmarking all AV-NeRF variants. We adopt the same binaural generator as AV-NeRF for direct comparison, ensuring fairness and representativeness.

Table 1. Performance comparison in terms of MAG and ENV. Lower MAG/ENV is better.

Methods	Modality		Office ↓		House ↓		Apartment ↓		Outdoors ↓		Overall ↓	
	Audio	Visual	MAG	ENV	MAG	ENV	MAG	ENV	MAG	ENV	MAG	ENV
Mono-Mono	✓	✗	9.269	0.411	11.889	0.424	15.120	0.474	13.957	0.470	12.559	0.445
Mono-Energy	✓	✗	1.536	0.142	4.307	0.180	3.911	0.192	1.634	0.127	2.847	0.160
Stereo-Energy	✓	✗	1.511	0.139	4.301	0.180	3.895	0.191	1.612	0.124	2.830	0.159
INRAS [16]	✓	✗	1.405	0.141	3.511	0.182	3.421	0.201	1.502	0.130	2.460	0.164
NAF [17]	✓	✗	1.244	0.137	3.259	0.178	3.345	0.193	1.284	0.121	2.283	0.157
ViGAS [18]	✓	✓	1.049	0.132	2.502	0.161	2.600	0.187	1.169	0.121	1.830	0.150
AV-NeRF [4]	✓	✓	0.930	0.129	2.009	0.155	2.230	0.184	0.845	0.111	1.504	0.145
Phys-NVAS	✓	✓	**0.856**	**0.126**	**1.984**	**0.154**	**2.098**	**0.180**	**0.787**	**0.109**	**1.431**	**0.142**

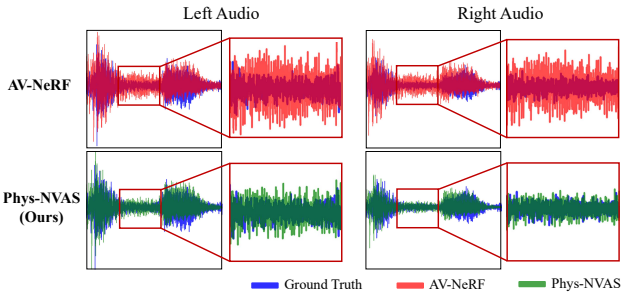

Fig. 3. Comparison of reconstructed binaural waveforms at a target listener position using AV-NeRF and our Phys-NVAS.

3.2. Performance Comparison

Table 1 reports the results across all environments. Our proposed Phys-NVAS consistently outperforms the compared baseline methods with the lowest MAG and ENV scores. Compared with AV-NeRF [4], Phys-NVAS gains stem from multi-view 3D acoustic modeling, which improves spatial awareness of direct sound and early reflections, and physics-aware vision–language priors, which capture object layout and material properties shaping reverberation and fine acoustic details. These complementary pieces of information provide physically grounded enhancements, validating the effectiveness of our Phys-NVAS with physics-aware feature representation in generating more realistic spatial audio[1].

3.3. Visualization Analysis

Fig. 3 compares binaural waveforms generated by AV-NeRF and the proposed Phys-NVAS with the ground truth. Phys-NVAS more closely follows the reference in both energy envelope and temporal dynamics, producing peaks with consistent amplitude and timing while exhibiting lower background fluctuations. In addition, it better preserves interaural differences, with channel energy asymmetry more closely aligned to the ground truth. These results indicate that Phys-NVAS captures scene acoustic characteristics more accurately and provides perceptually more reliable spatial cues for binaural audio generation.

[1]Demo examples are available at: https://physnvas.github.io/.

Table 2. Ablation study of feature components.

Feature Sources			Overall ↓	
RGB	DEP	SEM	MAG	ENV
			1.638	0.146
✓			1.463	0.142
	✓		1.476	0.143
		✓	1.602	0.145
✓	✓		1.448	0.143
✓		✓	1.463	0.143
	✓	✓	1.468	0.143
✓	✓	✓	**1.431**	**0.142**

3.4. Ablation Study

We evaluate the contribution of three feature sources, where RGB, DEP, and SEM denote $\mathbf{F}_{RGB}$ (multi-view appearance feature from 3DGS), $\mathbf{F}_{Depth}$ (depth-based structural feature), and $\mathbf{F}_{Phys}$ (semantic prior from the VLM), respectively. The baseline (first row in Table 2) uses only the binaural audio generator from AV-NeRF, conditioned on the mono audio input and listener position, without incorporating any additional priors. As shown in Table 2, each feature individually improves performance, combining any two yields further gains, and using all three achieves the best results, confirming that geometric (RGB, DEP) and semantic (SEM) cues are complementary for accurate spatial audio synthesis.

4. CONCLUSION

This paper presented a physics-aware framework for novel-view acoustic synthesis that integrates multi-view spatial perception and vision–language priors to jointly model scene geometry, object layout, and material properties. Unlike prior audio-only or single-view visual methods, Phys-NVAS provides a unified representation that captures both geometric and physics-aware semantic cues critical for realistic acoustics. Experiments on the RWAVS dataset show consistent gains across environments, validating the effectiveness of Phys-NVAS and the complementarity of geometric and semantic cues.

5. ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 3072025YY0601).

6. REFERENCES

[1] M. A. Gigante, "Virtual Reality: Definitions, History and Applications," in *Virtual Reality Systems*, 1993.

[2] J. Carmigniani and B. Furht, "Augmented Reality: An Overview," *Handbook of Augmented Reality*, 2011.

[3] X. Lu, Y. Chen, Z. Chen, J. Wang, M. Liu, H. Hu, C. Zheng, S. Bleeck, and J. Sang, "Deep Learning for Personalized Binaural Audio Reproduction," *arXiv preprint arXiv:2509.00400*, 2025.

[4] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "AV-NeRF: Learning Neural Fields for Real-World Audio-Visual Scene Synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[5] S. Maluski and B. Gibbs, "The effect of construction material, contents and room geometry on the sound field in dwellings at low frequencies," *Applied Acoustics*, 2004.

[6] H. Gao, J. Ma, D. Ahmedt-Aristizabal, C. Nguyen, and M. Liu, "SOAF: Scene Occlusion-Aware Neural Acoustic Field," *arXiv preprint arXiv:2407.02264*, 2024.

[7] S. Bhosale, H. Yang, D. Kanojia, J. Deng, and X. Zhu, "AV-GS: Learning Material and Geometry Aware Priors for Novel View Acoustic Synthesis," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[8] H. Baek, H. Shin, J. Seo, C. Kim, S. Kim, H. Kim, and S. Kim, "AV-Surf: Surface-Enhanced Geometry-Aware Novel-View Acoustic Synthesis," *arXiv preprint arXiv:2503.12806*, 2025.

[9] M. Chen, I. D. Gebru, I. Ananthabhotla, C. Richardt, D. Markovic, J. Sandakly, S. Krenn, T. Keebler, E. Shlizerman, and A. Richard, "SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[10] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph. (TOG)*, 2023.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[13] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[14] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. NAACL)*, 2019.

[16] K. Su, M. Chen, and E. Shlizerman, "INRAS: Implicit Neural Representation for Audio Scenes," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[17] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning Neural Acoustic Fields," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[18] C. Chen, A. Richard, R. Shapovalov, V. K. Ithapu, N. Neverova, K. Grauman, and A. Vedaldi, "Novel-View Acoustic Synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.

[19] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, "Visually Informed Binaural Audio Generation without Binaural Audios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[20] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-Supervised Generation of Spatial Audio for 360 Video," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018.