3D Audio-Visual Segmentation

Artem Sokolov **Swapnil Bhosale** **Xiatian Zhu**
University of Surrey, UK
{as05633, s.bhosale, xiatian.zhu}@surrey.ac.uk

Abstract

Recognising the sounding objects in scenes is a longstanding objective in embodied AI, with diverse applications in robotics and AR/VR/MR. To that end, Audio-Visual Segmentation (AVS), taking as condition an audio signal to identify the masks of the target sounding objects in an input image with synchronous camera and microphone sensors, has been recently advanced. However, this paradigm is still insufficient for real-world operation, as the mapping from 2D images to 3D scenes is missing. To address this fundamental limitation, we introduce a novel research problem, *3D Audio-Visual Segmentation*, extending the existing AVS to the 3D output space. This problem poses more challenges due to variations in camera extrinsics, audio scattering, occlusions, and diverse acoustics across sounding object categories. To facilitate this research, we create the very first simulation based benchmark, *3DAVS-S34-O7*, providing photorealistic 3D scene environments with grounded spatial audio under *single-instance* and *multi-instance* settings, across 34 scenes and 7 object categories. This is made possible by re-purposing the Habitat simulator [32] to generate comprehensive annotations of sounding object locations and corresponding 3D masks. Subsequently, we propose a new approach, `EchoSegnet`, characterized by integrating the ready-to-use knowledge from pretrained 2D audio-visual foundation models synergistically with 3D visual scene representation through spatial audio-aware mask alignment and refinement. Extensive experiments demonstrate that `EchoSegnet` can effectively segment sounding objects in 3D space on our new benchmark, representing a significant advancement in the field of embodied AI. Project page: https://surrey-uplab.github.io/research/3d-audio-visual-segmentation/

1 Introduction

Human perception of the real world, both visual and acoustic, predominantly occurs in three dimensions. Prior psychology literature [37] has highlighted humans' remarkable ability to correspond across multiple modalities, often involving the association of events across these modalities. For instance, we can effortlessly ground emergent surround sound with its potential source in 3D visuals [26]. Inspired by this capability, a crucial aspect in the development of embodied AI systems is their ability to integrate cues from synchronous multimodal input streams and establish targets corresponding to their goals. In this work, we aim to build a machine model to achieve this multimodal correspondence, particularly targeted towards the task of audio-visual segmentation (AVS) in 3D.

Albeit AVS has been widely explored within audio-visual scene analysis and correspondence learning, prominent research in this field has focused on 2D environments involving mono (single channel) sound sources, thus devoid of spatial presence entirely. In this paper, we take the first step towards exploring 3D AVS and introduce a large benchmark, **3DAVS-S34-O7**. Our exploration is rooted in a fundamental grounding problem: given an embodied agent equipped with a camera and a binaural microphone, can we teach the agent to obtain fine-grained localization of potential sounding objects (generally by predicting a segment-level mask of the object in 3D) while also utilizing spatial audio

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

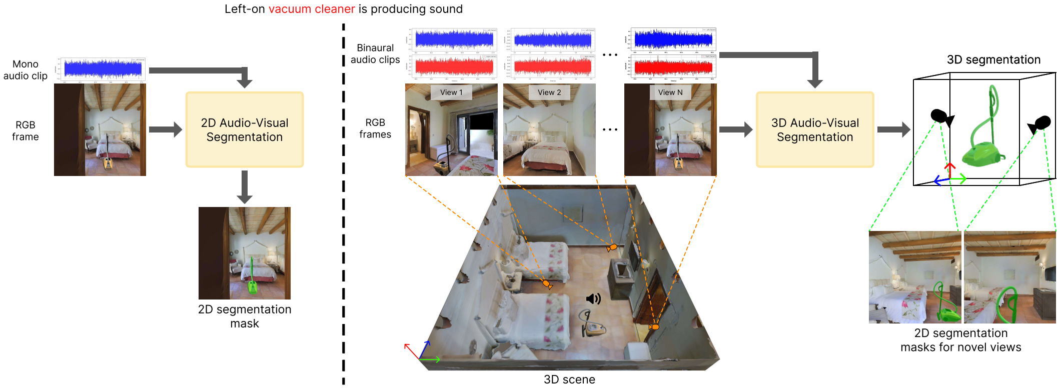
Left-on vacuum cleaner is producing sound
Mono audio clip
RGB frame
2D Audio-Visual Segmentation
2D segmentation mask
Binaural audio clips
RGB frames
View 1
View 2
View N
3D Audio-Visual Segmentation
3D segmentation
3D scene
2D segmentation masks for novel views

Figure 1: Comparison of the existing 2D AVS task with our proposed 3D AVS. Former task utilises single channel audio to generate pixel-level masks of the potential sounding object in the input RGB frame. 3D AVS on the other hand is aimed at generating 3D masks (from which multi-view consistent 2D masks can be rendered) while utilizing multichannel (spatial) audio.

cues? (see Fig. 1) Furthermore, we extend our benchmark to include a more competitive *multi-instance* setup where, although multiple instances of the same object might be present in the scene, the goal is to segment only the sounding instance. This setup helps us testify to the efficacy of spatial presence harnessed from the input binaural audio samples.

Recently, 3D Gaussian Splatting (3D-GS) [15] has emerged as a prospective method for modeling static 3D scenes directly from input RGB frames. Owing to its explicit Gaussian based representation, it has paved a natural pathway for 3D visual segmentation [13, 41, 33]. Deriving inspiration from human spatial memory in indoor environments, we design `EchoSegnet`, a purely training-free pipeline for 3D AVS within a 3D-GS representation. `EchoSegnet` leverages 2D foundation models (namely SAM [17] and Imagebind [9]) to first obtain 2D AVS masks on the input RGB frames. These 2D AVS masks are further used to segment the Gaussians in the learned 3D-GS representation to obtain multi-view masks to achieve a consistent 3D segmentation.

To summarize, we make the following *contributions*: (1) the first 3D audio-visual segmentation benchmark composing of fairly complex indoor room scenes with integrated spatial sound cues; (2) a training-free AVS framework, `EchoSegnet`, capable of syncing across sequential frames from 3D environments; (3) a novel Audio-Informed Spatial Refinement Module `AISRM`, designed to enhance 3D segmentation and resolve ambiguities in complex, multi-instance environments by leveraging spatial audio intensity maps. We perform a comprehensive evaluation of `EchoSegnet` on the proposed **3DAVS-S34-O7** for both *single-instance* and *multi-instance* scenarios, along with an ablative comparison with existing 2D AVS models, highlighting their shortcomings in aligning audio-visual cues within 3D scenes -establishing their adaptation to **3DAVS-S34-O7** as non-trivial.

2 Related Work

Audio-visual segmentation Existing AVS methods cater to 2D scenes with mono audio as inputs to identify audible visual pixels associated with a given audio signal [43, 12, 22, 27, 34, 42] and are typically trained on thousands of manually annotated 2D segmentation masks. Although there have been recent improvements on reducing the dependence on annotated AV masks using weakly supervised [26, 20] or entirely unsupervised methods [5, 19, 39, 7], there has not been any effort extending the task of AVS particularly to 3D scenes, with spatial audio cues. To address this gap, we propose the first benchmark for 3D AVS harnessing existing embodied AI platforms (Habitat simulator [32]) to capture visual and (binaural) acoustic cues for sounding objects placed in 3D indoor scenes. We believe this lays a prominent groundwork for systematic evaluation of future embodied systems for 3D segmentation.

3D scene representations Point-based rendering techniques, initiated by [10], utilize point-based explicit representation where each point affects a single pixel. Zwicker et al. [44] advanced this with ellipsoid-based rendering (splatting), allowing mutual overlap to fill image holes. In the absence of

Figure 2: Sample scenes from **3DAVS-S34-O7** dataset.

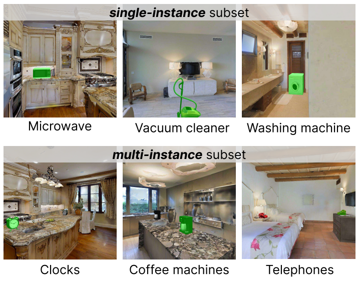

Table 1: Comparison with existing 3D visual segmentation benchmarks: NVOS [30] and SPIn-NeRF [25] support promptable 3D visual segmentation but lack spatial audio.

Benchmark	#(Objects)	#(Scenes)	Audio
NVOS [30]	8	8	✗
SPIn-NeRF [25]	6	10	✗
3DAVS-S34-O7 (Ours)	7	34	✓
single-instance	7	25	✓
multi-instance	7	9	✓

given geometry, Mildenhall et al. [24] explored neural implicit representation, NeRF, predicting view-dependent radiance via implicit density fields. 3D Gaussian Splatting (3D-GS) [15], a novel-view synthesis method, employs explicit point-based representation, contrasting with NeRF's volumetric rendering. Owing to its real-time high-quality rendering capabilities, 3D-GS has been applied to various domains, including simultaneous localization [14, 23], content generation [36], and 4D dynamic scenes [18, 38], among others. In this work, we utilize the explicit representation from learned 3D-GS and decompose 2D masks of potential sounding objects to obtain consistent 3D masks.

3 Dataset

Our proposed 3DAVS-S34-O7 is profoundly motivated towards simulating real-world indoor scenes, in terms of the visual quality of the scenes as well as the acoustic response generated by the objects placed within it. In the context of our 3D AVS task, we define an observation as O=$\{(v_i, a_i, m_i)\}_{i=1}^{n}$, where v_i, a_i represent the visual (RGB view, $\mathbb{R}^{1008\times1008\times3}$) and acoustic (1 second binaural audio, at 44.1kHz) cues respectively, captured by the embodied agent at i-th time. m_i represents a binary mask corresponding to v_i highlighting the sounding object. To record an observation, we load a randomly sampled scene from the Habitat-Matterport3D dataset [29] into the SoundSpaces 2.0 [4]. Next, we place a semantically relevant sounding object (for instance, *bathroom*$\leftrightarrow$*washing machine*, *kitchen*$\leftrightarrow$*microwave*, etc.) which emits a sound based on a mono audio (corresponding to the placed object and sourced from [8, 2, 28]). We capture $n = 120$ frames at 1 fps symbolizing different positions along the moving agent's path. Alongside the above *single-instance* setup, we also explore a slightly challenging *multi-instance* setup wherein, we place multiple instances of the sounding object, although only one instance is sound-emitting (Fig. 2). We split each observation into 1:7 for train:test split (following [1]). Details of the selected scanned spaces and sound-emitting objects are provided in Appendix A.3.

4 Method

Considering the complexity of the 3D AVS task and deriving inspiration from human spatial memory, we propose `EchoSegnet` (see Fig. 3), a training-free pipeline leveraging 2D foundation models. For each input view, v_i, we first obtain corresponding 2D AVS masks $\hat{m}_i$ using OWOD-BIND [3]. Particularly, OWOD-BIND prompts the SAM [17] model using bounding boxes obtained from class-agnostic object detection (CAOD) [21]. The mask proposals from SAM are further filtered based on maximum cosine similarity with a_i's audio embedding generated using ImageBind [9].

Please note, the masks $\hat{m}_i$ are confined to v_i however the main goal of the 3D AVS task is to obtain multi-view consistent masks of the potential sounding object for novel viewing positions (beyond v_i). Moreover, the sounding object may be fully or partly visible in the novel view. To achieve this, we propose to lift the 2D AVS masks $\hat{m}_i$ within an explicit 3D scene representation $\mathcal{G}$, generated using vanilla 3D-GS [16]. Although similar approaches exist for salient 3D visual segmentation (such as [13]), the sounding object in the context of 3D AVS, may not always be salient (i.e in the foreground). As a result, unlike [13], we opt to exclude out-of-view projections from the voting process for selecting underlying Gaussians as well as directly lift $\hat{m}_i$ (see Appendix A.5).

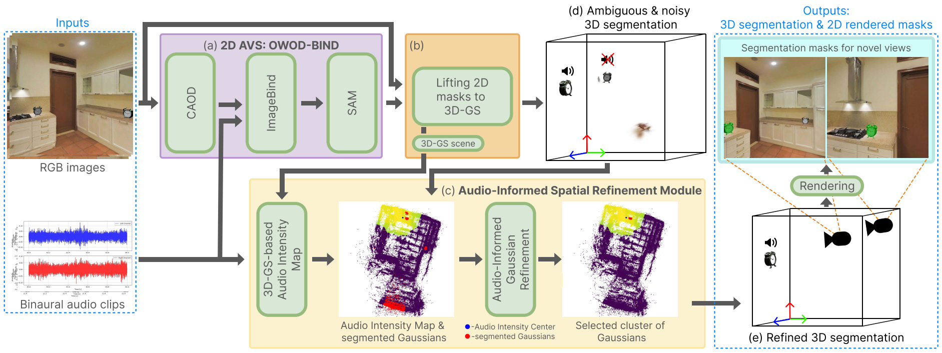
Inputs
RGB images
Binaural audio clips
(a) 2D AVS: OWOD-BIND
CAOD
ImageBind
SAM
(b)
Lifting 2D masks to 3D-GS
3D-GS scene
(d) Ambiguous & noisy 3D segmentation
Outputs: 3D segmentation & 2D rendered masks
Segmentation masks for novel views
Rendering
(c) Audio-Informed Spatial Refinement Module
3D-GS-based Audio Intensity Map
Audio Intensity Map & segmented Gaussians
● -Audio Intensity Center
● -segmented Gaussians
Audio-Informed Gaussian Refinement
Selected cluster of Gaussians
(e) Refined 3D segmentation

Figure 3: Overview of EchoSegnet: (a) 2D AVS pipeline OWOD-BIND [3] generates 2D masks. (b) These masks are lifted into a 3D-GS scene representation using [13] with a modified voting strategy. (d) The initial 3D segmentation may contain noise and ambiguities, as spatial relationships between objects and sound were not considered. (c) To address this, we apply the novel Audio-Informed Spatial Refinement Module (AISRM). (e) In the refined 3D segmentation, only the sound-emitting object instance is retained, and noise is filtered out.

Audio-Informed Spatial Refinement Module (AISRM) Although the above lifting process yields 3D segmentation masks, we observe certain ambiguities: (1) in the case of a *multiple-instance* setup, computation of $\hat{m}_i$, being devoid of spatial audio, is unable to accurately localize only the sound-emitting instance of the object (see Fig. 3(d)), and (2) due to errors in audio-visual alignment within the frozen ImageBind, $\hat{m}_i$ often includes other (silent) objects in the vicinity of the sound-emitting object.

To handle both the ambiguities, we start with a **3D-GS-based Audio Intensity Map**. Specifically, we introduce additional labels $I_{\mathbf{g}}$ on every Gaussian $\mathbf{g}$ within our scene representation $\mathcal{G}$ by weighing the root mean square (RMS) intensities on the agent's left and right audio channels, R^l, R^r respectively. For each Gaussian $\mathbf{g}$, we compute $I_{\mathbf{g}} = \sum_{i=1}^{t} \frac{|R_i^l - R_i^r|}{\max(R_i^l, R_i^r)} \cdot \mathbb{I}_{\text{RMS}}(\mathbf{g}_{center}, a_i)$, where $\mathbb{I}_{\text{RMS}}(.)$ equals 1 if the Gaussian center $\mathbf{g}_{center}$ is located on the side with the greater RMS intensity based on the binaural audio observation a_i. [40] proposed a similar intensity map but in two dimensions, and not grounded within an underlying 3D-GS representation.

We then perform an **Audio-Informed Gaussian Refinement** process through spatial clustering, guided by $I_{\mathbf{g}}$. We cluster the segmented 3D Gaussians using DBSCAN [6] and filter clusters with volumes $> \mu_v + 0.5\sigma_v$ where μ_v is the mean volume and σ_v is the standard deviation of all cluster volumes. Next we localize the audio intensity center by computing an average of the Gaussian center coordinates weighted by $I_{\mathbf{g}}$ (only Gaussians with $I_{\mathbf{g}} > \tau_{\text{ref}}$ are considered). We hypothesize that the cluster closest to the computed audio intensity center consists of Gaussians corresponding to the sound-emitting object, effectively filtering out both the inclusion of silent objects, as well as non-sound-emitting instances in the multi-instance setting.

5 Experiments

In this section, we demonstrate the effectiveness of EchoSegnet on the **3DAVS-S34-O7** benchmark, with a particular focus on the contribution of AISRM. Following [43], we adopt mIoU and F-Score as the metrics to estimate the segmentation performance. For implementation details, please refer to Appendix A.4. From Table 2, it is evident that removing the AISRM module results in a performance drop across both *single-instance* and *multiple-instance* settings. For the *single-instance* setting, mIoU decreases by 0.06, and F-Score drops by 0.10. Similarly, in the *multiple-instance* setting, mIoU drops by 0.04, and F-Score decreases by 0.11. As illustrated in Figure 4 (Left), omitting AISRM introduces noisy Gaussians representing silent objects (e.g., *the door*, view 3), negatively impacting performance across both subsets. In the *multiple-instance* setting, the inability to distinguish between sound-emitting and non-sound-emitting instances of the same object further reduces segmentation accuracy (e.g., both *clocks* are segmented, but only one is sound-emitting, view 1). Additionally, Gaussians in the vicinity of the sounding object (clock) are also incorrectly segmented (view 2).

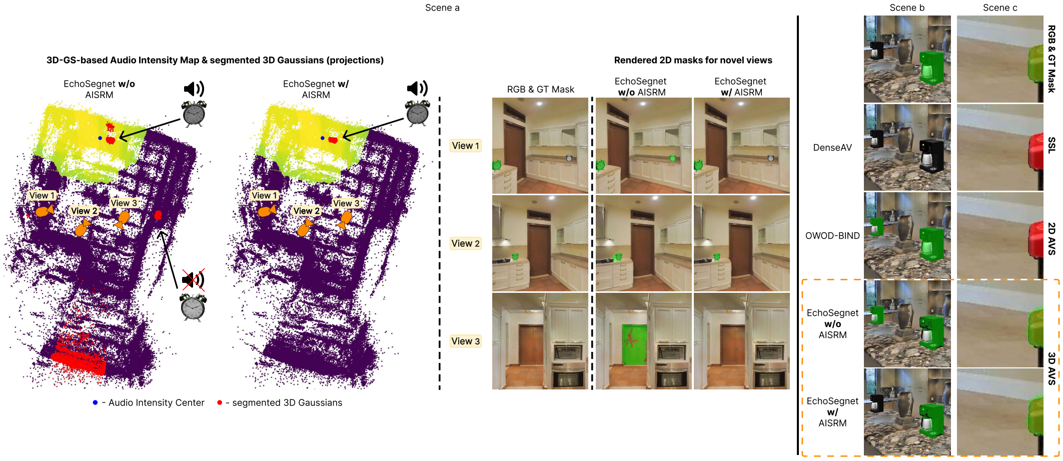

Figure 4: Left: (Scene a) Qualitative comparison of `EchoSegnet` performance with and without `AISRM`, illustrated through projected 3D-GS scene representation and renderings. Right: Comparison between DenseAV (SSL), OWOD-BIND (2D AVS) and `EchoSegnet`. (Scene b) OWOD-BIND [3] incorrectly segments the non-sound-emitting coffee machine. (Scene c) Both SSL and 2D AVS fail to handle a complex scenario where only a small part of the sound-emitting telephone is present in the view, whereas `EchoSegnet` successfully addresses this challenge.

Table 2: Performance comparison of `EchoSegnet` (3D AVS) with and without `AISRM`, and comparison against 2D AVS and SSL pipelines on both subsets of the **3DAVS-S34-O7** benchmark.

Approach	*single-instance*		*multi-instance*	
	mIoU ↑	F-Score ↑	mIoU ↑	F-Score ↑
`EchoSegnet w/o AISRM`	0.761	0.628	0.757	0.609
`EchoSegnet w/ AISRM`	**0.823**	**0.730**	**0.801**	**0.714**
DenseAV [11] (2D SSL)	0.426	0.023	0.436	0.023
OWOD-BIND [3] (2D AVS)	0.693	0.523	0.696	0.502

Comparison Between SSL, 2D, and 3D Audio-Visual Segmentation. We propose `EchoSegnet` as the first approach towards the novel 3D AVS task. Naturally, comparing the performance of existing 2D AVS (and Sound Source Localization (SSL)) approaches for the 3D AVS task is essential to establish the efficacy of `EchoSegnet`. From Table 2, it can be clearly observed that `EchoSegnet` consistently outperforms OWOD-BIND (a 2D AVS method) across both subsets, while DenseAV (a SSL method) shows significantly poorer and incomparable performance. The strength of `EchoSegnet` in performing 3D AVS lies in its ability to capture spatial relationships between objects and their sounds, which the existing 2D AVS methods lack, often resulting in segmentation of all visible instances (Figure 4, Right, Scene b).

6 Conclusion

In this work, we introduced 3D Audio-Visual segmentation (3D AVS) as a novel extension of the existing 2D AVS paradigm. We presented the **3DAVS-S34-O7** benchmark, the first simulation-based large dataset for 3D AVS, featuring photorealistic environments with spatial audio across 34 scenes and 7 object categories. Our proposed method, `EchoSegnet`, effectively segments sounding objects in 3D scenes in a training-free pipeline leveraging 2D audio-visual foundation models and 3D Gaussian Splatting. We believe this marks a significant advancement in bridging the gap between 2D and 3D audio-visual understanding, with broader implications for embodied AI. Looking ahead, we aim to explore diverse acoustic environments and dynamic objects as the future scope of this work.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022.

[2] BBC Sound Effects. BBC Sound Effects Archive. `https://sound-effects.bbcrewind.co.uk/`.

[3] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, and Xiatian Zhu. Leveraging foundation models for unsupervised audio-visual segmentation. *ArXiv*, abs/2309.06728, 2023.

[4] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

[5] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[7] Shun Fang, Qile Zhu, Qi Wu, Shiqian Wu, and Shoulie Xie. Audio–visual segmentation based on robust principal component analysis. *Expert Systems with Applications*, 256:124885, 2024.

[8] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*, pages 411–412, New York, NY, USA, October 21–25 2013. ACM.

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[10] Jeffrey P Grossman and William J Dally. Point sample rendering. In *Rendering Techniques' 98: Proceedings of the Eurographics Workshop*. Springer, 1998.

[11] Mark Hamilton, Andrew Zisserman, John R. Hershey, and William T. Freeman. Separating the "chirp" from the "chat": Self-supervised visual grounding of sound and language. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13117–13127, 2024.

[12] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2067–2075, 2024.

[13] Xu Hu, Yuxi Wang, Lue Fan, Junsong Fan, Junran Peng, Zhen Lei, Qing Li, and Zhaoxiang Zhang. Sagd: Boundary-enhanced segment anything in 3d gaussian via gaussian decomposition, 2024. arXiv preprint arXiv:2401.17857.

[14] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. *arXiv preprint arXiv:2312.02126*, 2023.

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[18] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023.

[19] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE Transactions on Multimedia*, 2024.

[20] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5604–5614, 2024.

[21] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *17th European Conference on Computer Vision (ECCV)*. Springer, 2022.

[22] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023.

[23] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. *arXiv preprint arXiv:2312.06741*, 2023.

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[25] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, and Alex Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023.

[26] Shentong Mo and Bhiksha Raj. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023.

[28] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015.

[29] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[30] Zhongzheng Ren, Aseem Agarwala†, Bryan Russell†, Alexander G. Schwing†, and Oliver Wang†. Neural volumetric object selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. († alphabetic ordering).

[31] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.

[32] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[33] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. *ECCV*, 2024.

[34] Zhaofeng Shi, Qingbo Wu, Hongliang Li, Fanman Meng, and Linfeng Xu. Cross-modal cognitive consensus guided audio-visual segmentation. *arXiv preprint arXiv:2310.06259*, 2023.

[35] Sketchfab. Sketchfab 3d models. `https://sketchfab.com`, 2024.

[36] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[37] Robert B Welch and David H Warren. Immediate perceptual response to intersensory discrepancy. *Psychological bulletin*, 88(3):638, 1980.

[38] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023.

[39] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27134–27143, 2024.

[40] Zeyuan Yang, Jiageng Liu, Peihao Chen, Anoop Cherian, Tim K. Marks, Jonathan Le Roux, and Chuang Gan. Rila: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16251–16261. IEEE, 2024.

[41] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *ECCV*, 2024.

[42] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7239–7257, 2022.

[43] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022.

[44] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Annual conference on Computer graphics and interactive techniques*, 2001.

A Appendix

A.1 Limitations & Failure Cases

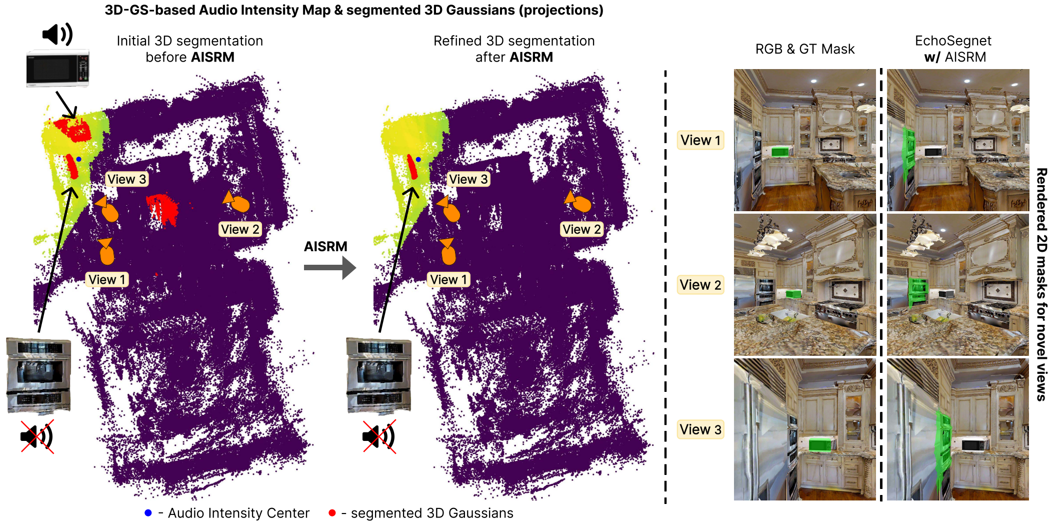
3D-GS-based Audio Intensity Map & segmented 3D Gaussians (projections)
Initial 3D segmentation before AISRM
Refined 3D segmentation after AISRM
AISRM
View 1
View 2
View 3
RGB & GT Mask
EchoSegnet w/ AISRM
Rendered 2D masks for novel views
● - Audio Intensity Center
● - segmented 3D Gaussians

Figure 5: Failure case: Due to the close proximity of the microwave and oven, the AISRM mistakenly refines the segmented Gaussians to those of the silent oven, discarding the Gaussians of the sound-emitting microwave.

Despite the evident improvements brought by AISRM, it struggles when objects are positioned too closely, which consequently impacts the overall performance of EchoSegnet. In Figure 5, both the microwave and oven are initially segmented in 3D due to a misalignment in ImageBind [9], even though only the microwave is emitting sound. While AISRM typically resolves such ambiguities, in this case, the 3D-GS-based Audio Intensity Map provided conflicting guidance due to the proximity of the objects. Consequently, the AISRM refinement process incorrectly retained the Gaussians corresponding to the silent oven, rather than the sound-emitting microwave, as seen in the rendered 2D masks for novel views.

A.2 Acknowledgement

The 3D models used in this research were sourced from Sketchfab [35] and are available under various open licenses, including Creative Commons, which permit their use in academic research.

A.3 Dataset: Sound-Emitting Objects and Scanned Spaces

The **3DAVS-S34-O7** dataset focuses on indoor environments, using scanned spaces from the Habitat-Matterport3D dataset [29]. Four scans were chosen based on their scanning quality and suitability for audio rendering (shown in Fig. 6, Right). Seven commonly found sound-emitting objects were selected: a washing machine, toilet, vacuum cleaner, microwave, coffee machine, clock, and telephone (shown in Fig. 6, Left). The 3D models of these objects were sourced from Sketchfab [35] and selected for their realism.

A.4 Implementation Details

In the OWOD-BIND [3] pipeline, each 1-second audio clip is extended to 2 seconds by appending 0.5 seconds of audio from neighboring clips before being input into the ImageBind [9] audio encoder, and the threshold τ_{BIND} is set to 0.2. To construct the 3D Gaussians Splatting [16] scene representation, the original image resolution of 1008x1008 is retained, and each scene is trained for 30,000 iterations. For the modified voting strategy in SAGD [13], the threshold τ_{voting} is set to 0.3, with the interval parameter for Gaussian Decomposition fixed at 4, as recommended by [13]. During the DBSCAN [31] clustering process, an epsilon value of 0.04 is used, with a minimum point count of 6, as

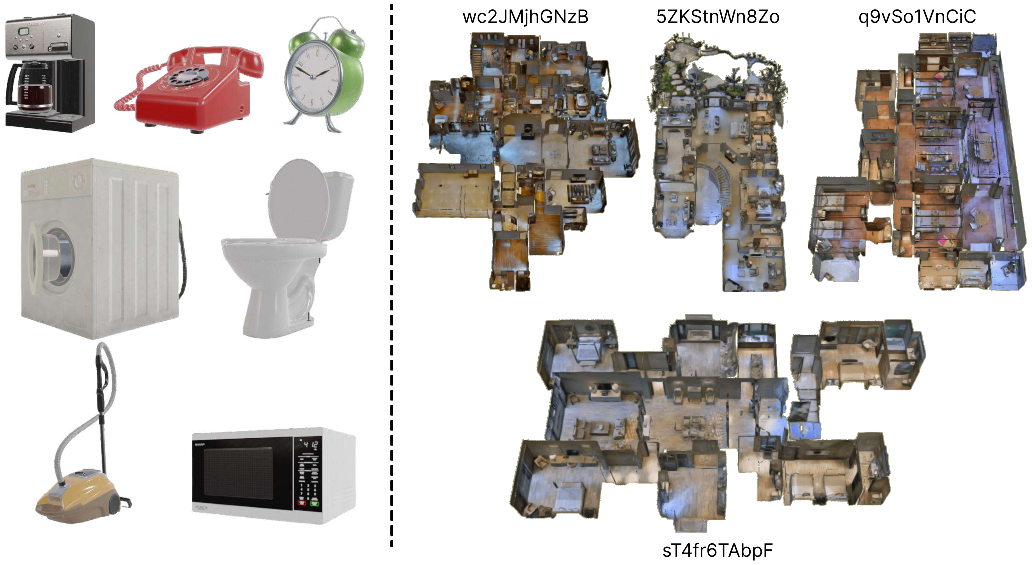

Figure 6: Left: 3D models of the seven selected sound-emitting objects (scale is not preserved). Right: Four selected scanned spaces from Habitat-Matterport3D [29] with corresponding dataset IDs.

suggested by *Sander et al.* [31]. For the 3D-GS-based Audio Intensity Map, we use a threshold τ_{ref} of 0.85, meaning only Gaussians with a normalized audio intensity greater than 0.85 are considered. All experiments were conducted using a GeForce GTX 1080 Ti GPU.

A.5 Modified Voting Strategy for SAGD

In contrast to SAGD's original voting strategy [13], which selects 3D Gaussians based on their projection into the 2D object mask more frequently than into the background or out of view, followed by thresholding, we exclude out-of-view projections from the voting process. Thresholding is applied solely based on the ratio of projections into the mask versus the background. This modification allows us to lift object masks as long as the object is consistently segmented in 2D, even if it appears in only a limited number of views. Since we apply additional refinement via `AISRM`, compared to the original SAGD [13], it is reasonable to use a lower thresholding value τ_{voting}. This approach prioritizes segmenting as many Gaussians as possible of the sound-emitting object, even if some Gaussians representing other objects are included, rather than risking the omission of Gaussians related to the sound-emitting object (demonstrated in Table 3). Additionally, we omit SAGD's original 3D prompt construction strategy, opting instead to directly lift the masks predicted by OWOD-BIND [3].

Table 3: `EchoSegnet` performance on a sample *single-instance* scene (sT4fr6TAbpF, bathroom with sound-emitting vacuum cleaner) with varying τ_{voting} thresholds. Values below 0.3 have little impact on accuracy due to `AISRM`, while higher values reduce performance.

τ_{voting}	mIoU ↑	F-Score ↑
0.9	0.333	0.028
0.8	0.333	0.028
0.7	0.335	0.028
0.6	0.352	0.104
0.5	0.396	0.241
0.4	0.897	0.948
0.3	**0.901**	**0.949**
0.2	0.901	0.949
0.1	0.901	0.948