3D SoundMask: Connecting Audio and 3D Scenes for Interactive Objective Segmentation

Minseo Kim
Seoul National University
ms05251@snu.ac.kr

Yujin Kim
Seoul National University
yujin.k@snu.ac.kr

Hochan Jeong
Seoul National University
jhchnewage114@snu.ac.kr

Abstract

We present 3D SoundMask, a novel framework for Audio-Visual Segmentation (AVS) that extends 2D sound masks into fully view-consistent 3D reconstructions. First, we introduce a dynamic synthetic dataset: starting from original 2D AVS datasets, we generate videos via an image-to-video foundation model, and we produce stereo recordings and ground-truth masks through Unity-based simulation. Second, our pipeline integrates per-chunk 2D Audio-Visual Instance Segmentation(AVIS), stereo time-delay-of-arrival depth estimation, 4D Gaussian Splatting for dynamic scene reconstruction, and mask-informed Gaussian segmentation (SAGA) to yield precise 3D sound source masks. Experiments demonstrate that our method produces multi-view consistent segmentations and handles time-variant scenes effectively. Our code is publicly available at https://github.com/minseo25/3D-SoundMask.

1. Introduction

The integration of visual and auditory information is essential for robust video-based scene understanding. Audio-visual segmentation (AVS) methods extract frame-level masks by correlating audio embeddings with image features, and have demonstrated significant progress in identifying sounding objects in 2D frames. However, conventional 2D AVS remains confined to the image plane and cannot recover true spatial depth, limiting its ability to represent real-world scenes.

Recent advances in 3D reconstruction, most notably 3D Gaussian Splatting (3DGS) [11], can render high-resolution novel views far more efficiently than NeRF. Yet 3DGS relies on static Structure-from-Motion, making it unsuitable for dynamic content. To address moving scenes, 4D Gaussian Splatting (4DGS) [13] augments each Gaussian with time-dependent parameters, enabling real-time reconstruction of deforming or moving objects. However, existing pipelines still treat audio and vision independently, and no prior work delivers a unified 4D audio-visual segmentation.

In this paper, we introduce ***3D SoundMask***, the first public end-to-end framework for 4D audio-visual segmentation that jointly reconstructs dynamic geometry and segments sound sources in 3D. Specifically, our contributions are:

1. **Synthetic Moving-Camera Dataset:** We convert static AVS clips into dynamic stereo videos with a foundation model, then generate ground-truth 3D masks and spatialized audio via Unity simulation.
2. **Unified 4D AVS Pipeline:** We split the audio into active segments, run a 2D AVIS model for per-frame masks [8], estimate depth via stereo time-delay-of-arrival (TDoA) [12], reconstruct dynamic geometry with 4DGS [13], and segment Gaussians with SAGA [1] to produce view-consistent 3D sound masks.

Our experiments demonstrate that 3D SoundMask achieves multi-view consistent segmentation in dynamic scenes.

2. Related Works

2.1. Audio Feature Extraction

Learning rich audio embeddings is a foundational step for any audio-driven segmentation task. AST [7] employs a spectrogram transformer trained on large labeled corpora to produce frame-level representations, and AV-SepFormer [14] jointly trains cross-modal attention between audio and video streams. In contrast, CLAP [24] uses contrastive language-audio pretraining, so that semantically related audio and text share nearby vectors in an embedding space. In our experiments on ESC-50 and FSDnoisy18K, CLAP's contrastively trained embeddings most consistently grouped similar acoustic events, so we adopt CLAP as our audio encoder for 4D audio-visual segmentation.

2.2. Sound Event Detection

Transformer-based models have recently pushed the frontier of sound event detection by improving both accuracy

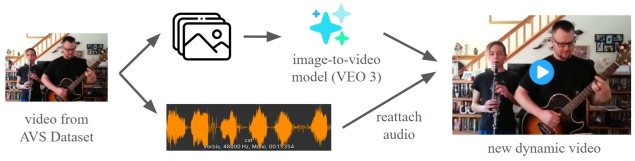

Figure 1. Synthetic video dataset generation via foundation model.

(a) Recorded Video (b) Recorded Video with GT sound mask

Figure 2. Synthetic video dataset generation using Unity. (a) recorded video, (b) video with GT sound mask (yellow)

and latency. SAT [5] processes short (1-2 s) audio segments with a lightweight memory module, achieving low-latency streaming tagging and competitive AudioSet metrics. PretrainedSED [17] provides an AudioSet-pretrained transformer backbone that delivers strong frame-level detection in a compact architecture. HTS-Audio-Transformer [3] hierarchically fuses Swin Transformer blocks with a token-semantic module to generate precise class-wise feature maps. ATST-SED [18] fine-tunes a lightweight ATST model using feature-level mix-up and pseudo-labeling. After assessing ease of installation, benchmark performance, and dependency stability, we adopt SAT as our backbone for our 4D audio-visual segmentation pipeline.

2.3. 2D Audio-Visual Segmentation

Mapping audio embeddings onto image features is the core task in 2D sound-source instance segmentation. Zhou et al. [26] introduce TPAVI, which computes pixel-wise audio–visual correlations over short video clips to generate instance masks. Guo et al. [8] extend this by combining frame-level audio-guided masks with video-wide object tracking to maintain sound-source identities across time. Subsequent works such as AVSegFormer [6], Multimodal VAE [15], and audio-query segmentation [9] improve representation learning and mask quality, but they remain limited by the scarcity and homogeneity of existing annotated datasets. Moreover, all 2D AVS approaches operate strictly in the image plane, making them unable to exploit 3D geometry or render segmentation masks from novel viewpoints.

2.4. 4D Reconstruction

4D Gaussian Splatting extends static 3DGS by endowing each Gaussian with time-dependent parameters, enabling real-time dynamic scene rendering. MoSca [13] converts 2D motion tracks, depth estimates, and pixel-likelihoods into a sparse 4D "motion scaffold" that deforms and fuses Gaussians to reconstruct moving objects in monocular videos. The original 4DGS framework [22] incorporates a 4D neural-voxel grid and a lightweight MLP to decode per-timestamp deformations, allowing training in minutes and real-time rendering. SC-GS [10] introduces sparse 6-DoF control points whose interpolated transforms drive dense Gaussians to produce temporally coherent, editable dynamic reconstructions. Finally, Yuan et al.'s 1000+FPS variant [25] prunes short-lived Gaussians based on spatio-temporal variation and masks inactive ones during rasterization, boosting throughput to over 1000 FPS.

2.5. 3D Object segmentation

Extending 2D masks into 3D has led to methods that leverage LiDAR point clouds and geometric priors. LiDAR-MOS approaches first project multi-sweep scans into 2D, perform instance segmentation, and then back-project the resulting labels into 3D. The 2D-PASS-MOS model achieves state-of-the-art moving-object segmentation on SemanticKITTI and Apollo [16, 20]. A different line of work reconstructs Gaussian-based scene representations from LiDAR data to enable novel-view segmentation [21]. Unsupervised frameworks exploit dynamic scene information by training segmentation networks with self-supervised scene flow under consistency and smoothness constraints [19]. Another approach lifts 2D SAM masks into a 3D radiance field by alternating inverse rendering and cross-view prompting, producing coherent volumetric masks [2]. Despite these advances, no single method consistently excels across all 3D segmentation challenges.

3. Method

3.1. Synthetic Dataset Generation

Public audio–visual segmentation datasets pose some challenges for 3D reconstruction and spatial audio modeling. First, existing AVS benchmarks use fixed cameras and only very short clips (5s), which prevents reliable Structure-from-Motion. Second, uncurated, user-generated videos often contain background music, narration, and frequent scene cuts that corrupt both visual and audio consistency.

Foundation Model–based Generation To introduce camera motion into original AVS clips, we use an image-to-video foundation model. We decompose each clip into audio chunks and image frames, feed the frames through the model to synthesize dynamic sequences, and then reattach the original audio. This pipeline (see Fig. 1) extends static clips into more realistic camera trajectories, but the resulting videos remain short and lack true spatial audio cues, since the original mono audio is simply reattached.

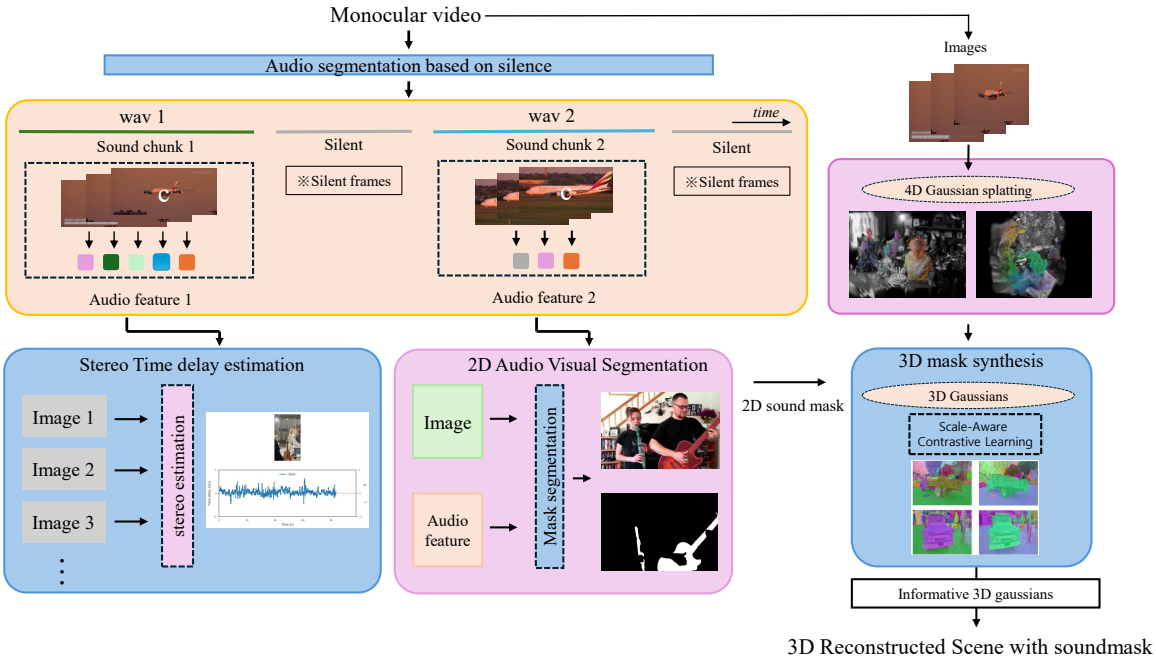

Figure 3. Main pipeline of 3D SoundMask.

Unity Simulation–based Generation For fully controllable, spatialized data, we build 3D environments (e.g., living room, store) in Unity. We attach stereo audio sources to scene objects along custom timelines and record video with genuine spatial audio. The Unity-generated examples (see Fig. 2) include (a) the raw recorded video and (b) the same frame overlaid with the ground-truth sound mask in yellow. This approach yields longer clips with accurate spatial cues and exact segmentation labels for robust training and evaluation.

3.2. Pipeline

Conventional 2D AVS uses *mono* audio to predict pixel-level masks of sounding objects in each RGB frame. Our 3D SoundMask generalizes this task: it exploits *stereo* cues to produce view-consistent 3D masks that can be re-rendered from any novel viewpoints (Fig. 3). The pipeline proceeds in six stages:

1. **Audio chunking**: We detect silent intervals in the input video and split the waveform into audio-active segments following SAT [5].
2. **Audio embedding and 2D AVS**: For every active segment we extract an audio embedding using CLAP [24]. In parallel, a 2D audio-visual instance-segmentation network (AVISeg + Mask2Former + Detectron2) [4, 8, 23] predicts a binary mask of the sounding object in each frame.
3. **Stereo depth cue**: We estimate time-delay-of-arrival (TDoA) with StereoTDE [12], providing a depth prior for each sounding instance.
4. **4D scene reconstruction**: All RGB frames are fed to MoSca-based 4D Gaussian Splatting [13]. MoSca optimizes a time-aware set of Gaussians whose per-timestamp deformations capture both static background and moving objects, yielding a dynamic representation that renders in real time.
5. **Gaussian segmentation**: We apply Segment Any Gaussian (SAGA) [1] directly on the 4D Gaussian volume to partition it into object–specific subsets.
6. **Sound localization**: For every frame, we back-project each segmented Gaussian component into the image plane and measure its IoU with the corresponding 2D sound mask. Components whose IoU exceeds a threshold form the candidate set. When multiple candidates remain, we break ties with a stereo TDoA depth check.
7. **Embedding & tagging**: The selected Gaussian subset is tagged with its CLAP audio embedding and timestamp, producing a sound-annotated 3D object that can be queried, edited, or rendered from novel viewpoints.

This unified pipeline converts a monocular stereo video into an editable 4D representation that supports interactive object segmentation and multi-view rendering.

Figure 4. Results of 3D SoundMask.

4. Results

Figure 4 presents a representative outcome of 3D SoundMask. From left to right, the original source video frame is shown first, followed by the 2D sound mask overlay that highlights the flute and guitar in green and red respectively; next comes the reconstructed 3D scene generated by our pipeline; and finally, the 3D sound mask is visualized from a novel viewpoint as a heatmap-like overlay on the geometry. By projecting the sound mask onto the reconstructed scene, we obtain a spatially consistent heatmap that roughly covers each sounding object, enabling segmentation and visualization of audio sources from arbitrary viewpoints.

5. Limitations

Despite its promising results, 3D SoundMask has several limitations:

Incomplete quantitative evaluation We have not yet performed a full end-to-end assessment (e.g., Mask IoU vs. baseline, latency measurements). For synthetic data, one can compare predicted masks to ground truth via mesh-intersection metrics; for real videos, pseudo-labels from 2D AVS and novel-view reprojection IoU can estimate 3D accuracy.

Generalization brittleness Our train-free pipeline stitches together multiple state-of-the-art modules without joint optimization. While each component was chosen empirically, the assembled system can fail on unseen scenes, due to hyperparameter sensitivity (e.g., Gaussian-segmentation thresholds) and suboptimal fusion of heterogeneous models. Additionally, the 4D reconstruction and 3D Gaussian segmentation steps are time-consuming, limiting real-time applicability.

Audio complexity Real-world audio often contains off-camera noise and overlapping sources, leading the 2D AVIS module to produce false positives or incomplete masks. In particular, 2D AVIS models pretrained mostly on musical instruments struggle to classify wild sound events without waveform priors. Fine-tuning AVIS on our synthetic stereo dataset may improve robustness to these challenging audio conditions.

6. Conclusion

We have presented 3D SoundMask, the first public end-to-end framework for 4D audio-visual segmentation that jointly reconstructs dynamic scene geometry and segments sound sources in 3D. To address dataset limitations, we introduced two synthetic generation strategies: a foundation model–based pipeline that injects camera motion into existing AVS clips, and a Unity simulation approach that yields long stereo recordings with precise ground-truth masks. Building on this data, our unified pipeline combines per-chunk 2D AVIS, stereo TDoA depth estimation, 4D Gaussian Splatting for dynamic reconstruction, and mask-informed Gaussian segmentation (SAGA) to produce multi-view consistent 3D sound masks. Empirical results confirm that 3D SoundMask successfully handles time-varying scenes with view-consistent segmentation.

References

[1] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1971–1979, 2025. 1, 3

[2] Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields. *International Journal of Computer Vision*, 2025. 2

[3] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '22)*, pages 646–650, 2022. 2

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 3

[5] Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo

Zhang, Yujun Wang, and Bin Wang. Streaming audio transformers for online audio tagging. In *Proceedings of Interspeech 2024*, 2024. 2, 3

[6] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI '24)*, pages 12155–12163, 2024. 2

[7] Y. Gong, Y. A. Chung, and J. Glass. Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 1

[8] Ruohao Guo, Xianghua Ying, Yaru Chen, Dantong Niu, Guangyao Li, Liao Qu, Yanyu Qi, Jinxing Zhou, Bowei Xing, Wenzhen Yue, Ji Shi, Qixun Wang, Peiliang Zhang, and Buwen Liang. Audio-visual instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13550–13560, 2025. 1, 2, 3

[9] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23)*, pages 875–883, 2023. 2

[10] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4220–4230, 2024. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 1

[12] S. S. Kushwaha, I. R. Roman, M. Fuentes, and J. P. Bello. Sound source distance estimation in diverse and dynamic acoustic conditions. In *Proceedings of the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2023. 1, 3

[13] Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3

[14] Jiuxin Lin, Xinyu Cai, Heinrich Dinkel, Jun Chen, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Zhiyong Wu, Yujun Wang, and Helen Meng. Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '23)*, pages 1–5, 2023. 1

[15] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '23)*, pages 954–965, 2023. 2

[16] Akos Madaras Rozsa, Zoltan and Tamas Sziranyi. Efficient moing object segmentation in lidar point clouds using minimal number of sweeps. *IEEE Open Journal of Signal Processing*, 2025. 2

[17] Florian Schmid, Tobias Morocutti, Francesco Foscarin, Jan Schlüter, Paul Primus, and Gerhard Widmer. Effective pretraining of audio transformers for sound event detection. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '25)*, 2025. 2

[18] Nian Shao, Xian Li, and Xiaofei Li. Fine-tune the pretrained atst model for sound event detection. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '24)*, pages 911–915, 2024. 2

[19] Ziyang Song and Bo Yang. Ogc: Unsupervised 3d object segmentation from rigid dynamics of point clouds. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. 2

[20] Fangzhou Tang, Bocheng Zhu, and Junren Sun. Gradient enhancement techniques and motion consistency constraints for moving object segmentation in 3d lidar point clouds. *Remote Sensing*, 17(2):195, 2025. 2

[21] Weizhe Wei, Ozan Ülger, Farrokh K. Nejadasl, Theo Gevers, and Marc R. Oswald. 3d-avs: Lidar-based 3d auto-vocabulary segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8910–8920, 2025. 2

[22] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2

[23] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[24] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large–scale contrastive language-audio pretraining with feature fusion and keyword–to–caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 1, 3

[25] Yuheng Yuan, Qiuhong Shen, Xingyi Yang, and Xinchao Wang. 1000+ fps 4d gaussian splatting for dynamic scene rendering. *arXiv preprint arXiv:2503.16422*, 2025. 2

[26] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, 133(4):1644–1664, 2025. 2