# 3D Audio-Visual Segmentation

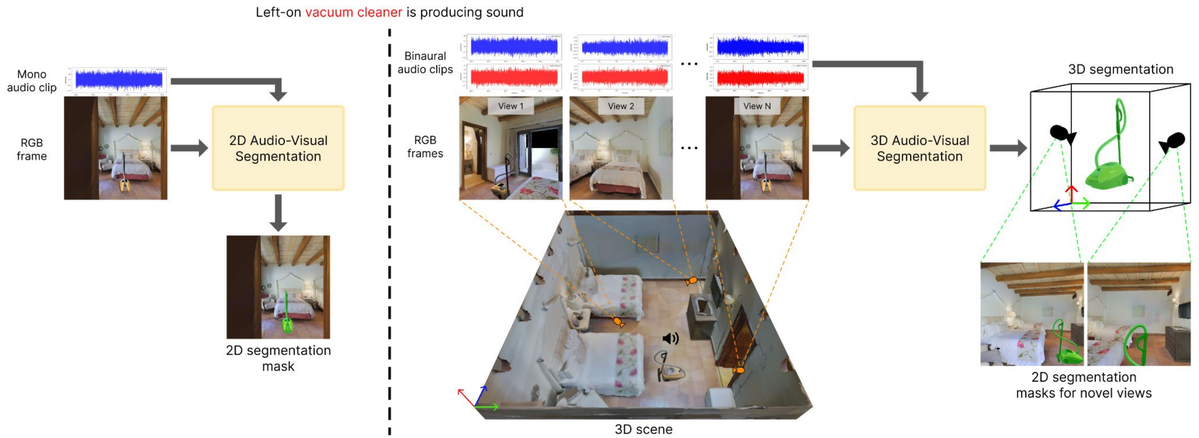
*Artem Sokolov, Swapnil Bhosale, Xiatian Zhu*
*University of Surrey*

**Audio Imagination: Workshop on**
**AI-Driven Speech, Music, and Sound Generation**
NEURAL INFORMATION PROCESSING SYSTEMS

## Motivation & Novel Research Problem:

- **Human perception** of the world predominantly occurs in **three dimensions**
- In **2D Audio-Visual Segmentation** (AVS), **mapping** from **2D** images to **3D** scenes **is missing** and **spatial audio** is **not** taken into **consideration**

=> **2D Audio-Visual Segmentation is insufficient** for real world operations

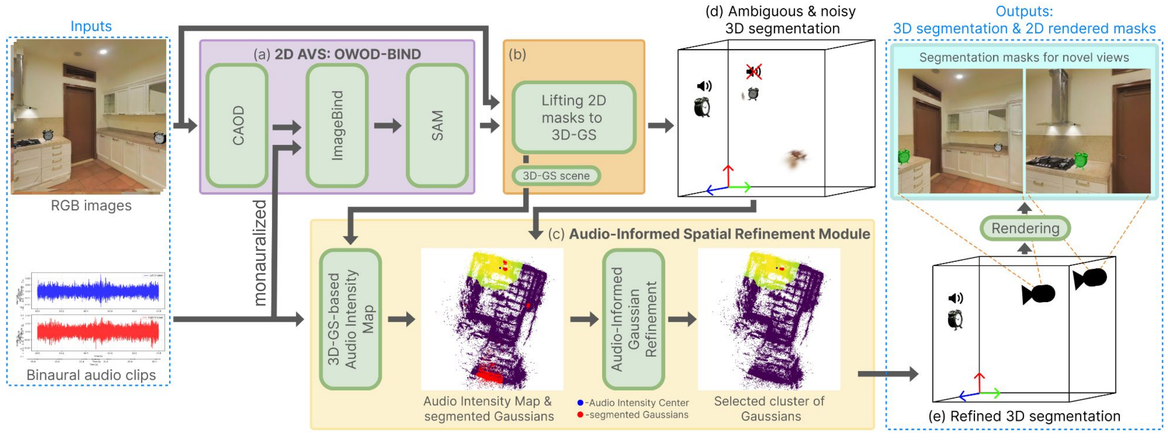



We propose a novel **3D Audio-Visual Segmentation** problem:

**Given**: 3D audio-visual scene (represented by sequence of frames with visual and spatial audio cues)

**Goal**: to obtain consistent 3D mask of the sound-emitting object

## Method: EchoSegnet




**Stage 1:**
Generate 2D Audio-Visual Segmentation masks using OWOD-BIND[3] (Class-agnostic object detection + SAM + ImageBind)

**Stage 2:**
Lift[4] these 2D AVS masks into built 3D Gaussian Splatting[5] scene representation

**Stage 3:**
Apply `AISRM` to the initial 3D segmentation to retain only 3D Gaussians of the sound-emitting object

## Dataset: 3DAVS-S34-O7


We propose the **first** simulation-based **3D Audio-Visual Segmentation benchmark 3DAVS-S34-O7**:

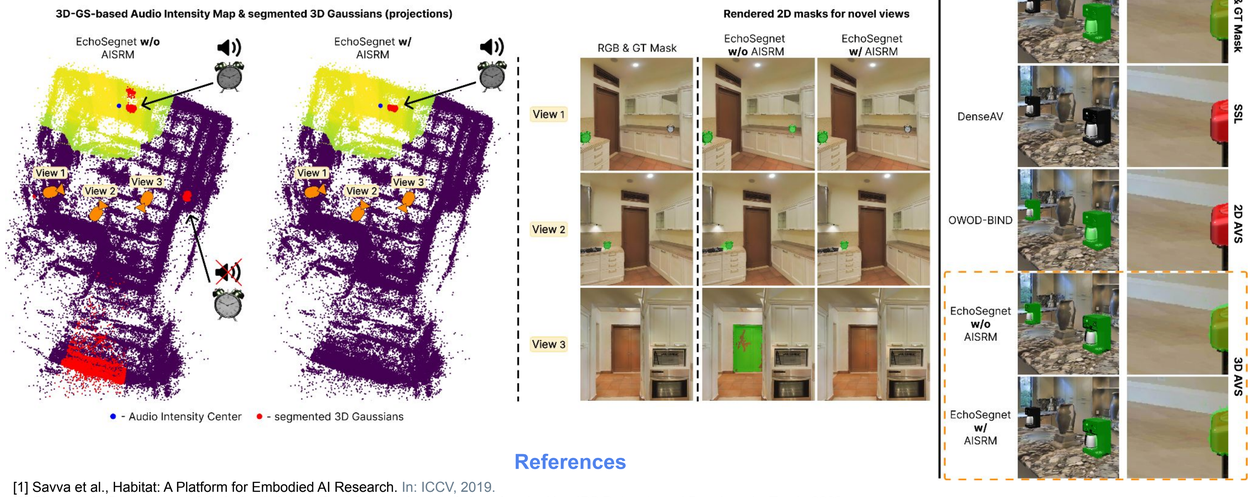
- 34 photorealistic, semantically meaningful indoor 3D scenes with visual and grounded spatial sound cues across 7 objects
- Created using Habitat[1] and SoundSpaces[2] 2.0 platforms
- Two benchmarking subsets:
  - *single-instance*
  - *multi-instance (with the goal to segment **only the sound-emitting** object **from multiple** instances)*

For each scene we capture 120 frames at 1 fps (symbolizing embodied agent's path).
Each frame includes: **RGB** view, 1 second **binaural audio**, and **semantic mask,** highlighting sounding object.

## Experimental Results:

| 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 |

- `AISRM` improves accuracy of `EchoSegnet` across both *single-* and *multi-instance* subsets
- 2D AVS pipeline OWOD-BIND[3] cannot address the 3D AVS task due to its inability to capture spatial relationships between objects and their sound
- `EchoSegnet` is the only method to segment sounding objects partially visible in frame




## References

[1] Savva et al., Habitat: A Platform for Embodied AI Research. In: ICCV, 2019.
[2] Chen et al., Soundspaces 2.0: A simulation platform for visual-acoustic learning. In: NeurIPS Datasets and Benchmarks Track, 2022.
[3] Bhosale et al. Leveraging foundation models for unsupervised audio-visual segmentation. In: ICCV Workshop AV4D, 2023.
[4] Hu et al. SAGD: Boundary-enhanced segment anything in 3d gaussian via gaussian decomposition. In: arXiv, 2024.
[5] Kerbl et al., 3d gaussian splatting for real-time radiance field rendering. In: ACM Transactions on Graphics, 42(4), 2023.