JAEGER: Joint 3D Audio-Visual Grounding and Reasoning in Simulated Physical Environments

Zhan Liu [1,2] Changli Tang [1] Yuxin Wang [3,2] Zhiyuan Zhu [4,2] Youjun Chen [5,2] Yiwen Shao [2] Tianzi Wang [2] Lei Ke [2] Zengrui Jin [1] Chao Zhang [1 ♣]

[1]Tsinghua University [2]Tencent AI Lab [3]HKUST [4]Zhejiang University [5]The Chinese University of Hong Kong ♣corresponding author

Introduction

- Current audio-visual LLMs are confined to **2D RGB video** and **monaural audio**.
- **JAEGER** — the first open-source **end-to-end 3D audio-visual LLM**, taking **RGB-D** vision and **4-channel spatial audio** (First-Order Ambisonics, **FOA**) as input.
- We propose a **3D audio-visual data simulation pipeline** and release **SpatialSceneQA** to train and evaluate JAEGER.
- We propose the **Neural Intensity Vector** — a **learnable** FOA spatial audio representation replacing the hand-crafted Intensity Vector for robust **overlapping-source** grounding.
- **Neural IV** yields **stable azimuth/elevation** cues under overlapping sources and **better cross-scenario generalization**; JAEGER can perform **3D audio-visual reasoning** that 2D AV-LLM baselines are fundamentally incapable of.

Dataset: SpatialSceneQA 61K

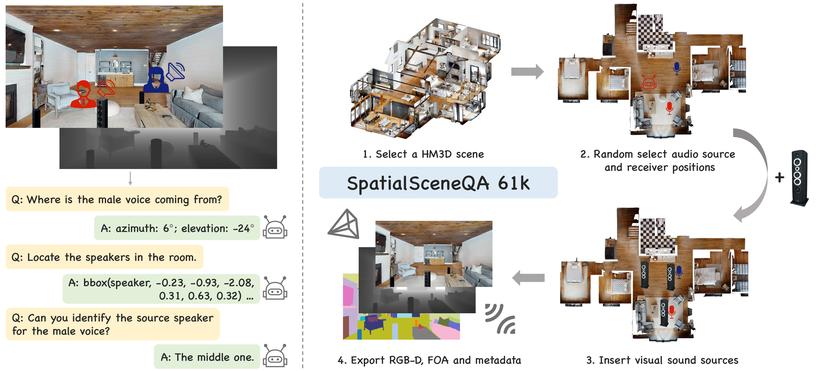

Figure 1. Data simulation pipeline: RGB-D + 4-channel FOA + dense 3D annotations, rendered with **SoundSpaces 2.0** and **Habitat-Sim** over **HM3D** scenes.

- **Acoustic sim: SoundSpaces 2.0** renders geometry-aware **room impulse responses** (reflection, diffraction, air absorption) for dry **LibriSpeech** speech.
- **Visual scenes:** in **Habitat-Sim**, load **HM3D** indoor scenes, insert **3D synthetic object meshes**, place a camera, and export RGB, depth, 3D bbox, distance, and camera intrinsics/extrinsics.
- **3D synthetic object meshes:** since static HM3D scans lack diverse sound-emitting objects, we insert 120 **loudspeaker meshes** (96/12/12 split) generated by Hunyuan3D.

Task Type	#Sources	Question & Answer Example
Perception		
A: Single-Source Audio DoA	1	**Q:** Based on the audio cues, please output the precise azimuth and elevation. **A:** azimuth: -7; elevation: -22
B: Overlap-Source Audio DoA	2	**Q:** Where is the female voice coming from? Output azimuth and elevation. **A:** azimuth: 28; elevation: -15
C: Visual Grounding	1 2 3	**Q:** Identify the 3D box for the speaker in this scene. **A:** bbox_0 = Bbox(speaker, 0.14, −0.48, −1.15, 0.33, 0.88, 0.32)
Reasoning		
D: Single-Source Multi-speakers Matching	2 3	**Q:** Determine which speaker corresponds to the audio source. **A:** Left
E: Overlap-Source Multi-speakers Matching	2 3	**Q:** Can you tell which speaker the male voice originates from? **A:** Center

Table 1. Task types in SpatialSceneQA 61K with representative Q&A prompts. **#Sources**: active audio sources (A–B) / inserted loudspeakers (C–E).

TL;DR: Using **RGB-D** vision and **4-channel spatial audio (FOA)**, **JAEGER** is the first **open-source E2E AV-LLM** for grounding and spatial reasoning in **simulated 3D scenes**.

Simulated Dataset | Model Architecture | Experiments

Model Architecture

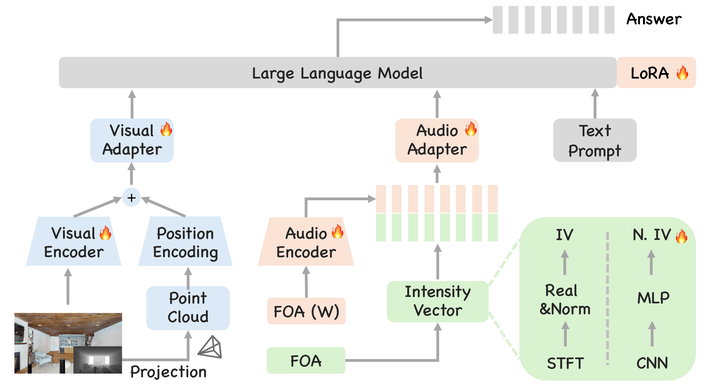

Figure 2. **JAEGER architecture.** A depth-aware visual stream and a dual-path audio stream (semantic + spatial) are fed into the AV-LLM for 3D grounding and reasoning.

- **LLM backbone:** Qwen2.5-Omni initialized, LoRA fine-tuned.
- **Visual stream:** RGB tokens fused with **3D-aware positional encodings**; depth back-projects pixels into a point cloud, adding sinusoidal (x, y, z) encodings: $\tilde{F}_{\mathrm{v}} = F_{\mathrm{v}} + F_{\mathrm{3D}}$.
- **Audio stream (dual-path):** semantic content from the omnidirectional FOA channel W; spatial cues from **Classical IV** or **Neural IV**.

Neural Intensity Vector

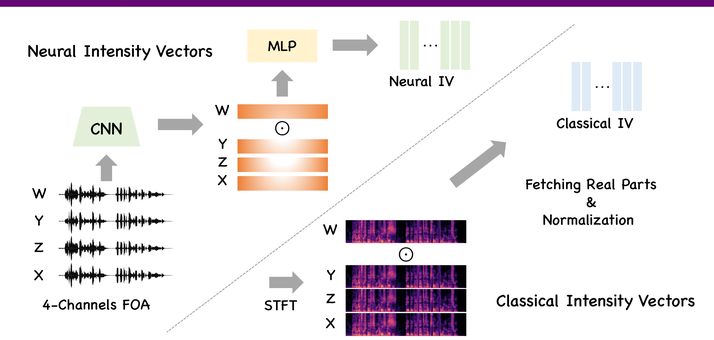

Figure 3. **Classical IV vs. Neural IV.** Neural IV replaces STFT with a learnable 1D-CNN on raw FOA waveforms, and the real-part extraction and L_2-norm steps with an MLP.

Experiments

Main results. JAEGER is the first model integrating RGB-D with 4-channel FOA; no direct baseline shares its modality setup.

Model	Modality	Perception				Reasoning	
		Audio DoA ↓	Overlap DoA ↓	3D IoU ↑	Vis. Offset ↓	1-Spk ↑	2-Spk ↑
Random	N/A	90°	90°	0.00	∞	45.6	47.4
Open-source omni model							
Qwen2.5-Omni	RGB+Mono	–	–	0.00	2.40	35.8	44.0
Open-source specialized models							
BAT	Binaural	**2.16°**	19.09°	–	–	–	–
Qwen3-VL-8B	RGB	–	–	0.01	1.11	–	–
N3D-VLM	RGB-D	–	–	0.00	2.04	–	–
JAEGER (Classical IV)	RGB-D+FOA	2.95°	6.44°	0.32	0.16	99.5	98.6
JAEGER (Neural IV)	RGB-D+FOA	2.21°	**4.11°**	**0.32**	**0.16**	**99.5**	**99.2**

Table 2. Visual offset in m; reasoning in %; "–" indicates the task is not supported.

- **Neural IV** matches BAT on single-source DoA and outperforms it on overlapping audio.
- JAEGER performs **3D audio-visual reasoning** that 2D omni models cannot.

Ablation Studies

IV Cross-Evaluation — Median DoA (°)

Test \ Train	Classical IV		Neural IV	
	Single	Overlap	Single	Overlap
Single Src.	2.95°	18.35°	**2.21°**	**14.91°**
Overlap Src.	19.25°	6.44°	**14.85°**	**4.11°**

- **Neural IV generalizes** better across mismatched train/test source configurations.
- **Architecture matters:** with identical inputs, a naive fusion baseline collapses on reasoning.

Real-World Transfer — STARSS23 Median Error (°)

Model	Elevation ↓		Azimuth ↓	
	Pretrain	Scratch	Pretrain	Scratch
Neural IV	**4.8°**	7.3°	76.7°	94.9°
Classical IV	7.0°	8.0°	74.2°	99.7°

- **Transfers to real audio:** SpatialSceneQA pretraining consistently improves real-world FOA localization over training from scratch.

Takeaways

- Explicit **3D audio-visual modeling** is essential.
- **Architecture matters:** JAEGER's design beats naive fusion.
- **Neural IV** is more robust across acoustic conditions.

Get our code, model checkpoints and dataset on our GitHub repo!