OWL: Geometry-Aware Spatial Reasoning for Audio Large Language Models

Subrata Biswas, Mohammad Nur Hossain Khan, Bashima Islam

Worcester Polytechnic Institute

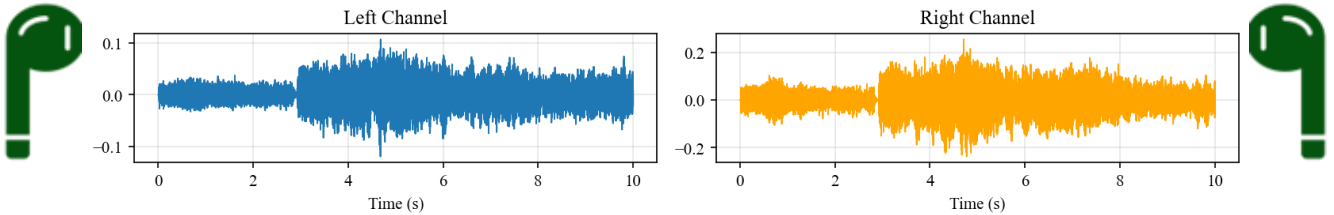
WPI

BASH LAB
#!/wpi/bash

ICLR

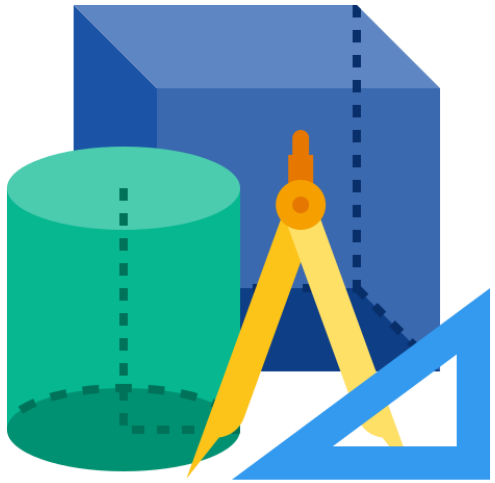
Current Audio LLMs Performance on Acoustic Localization

Q: Which sound source is left to the listener?



Gemini-2.5-Flash: From my listening position direction of the sound of the cat's meow is very much centered. It doesn't pan significantly left or right.

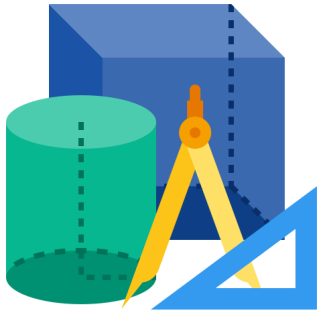
Why Audio LLMs Suffer in Spatial Reasoning



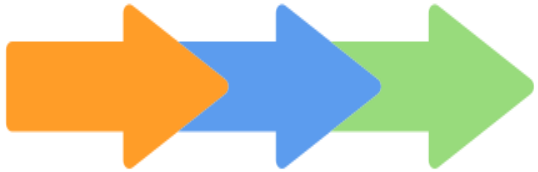
Audio encoders lack geometric grounding

- Capture only spectral and temporal patterns
- Overlook spatial cues

Why AudioLLMs Suffer in Spatial Reasoning



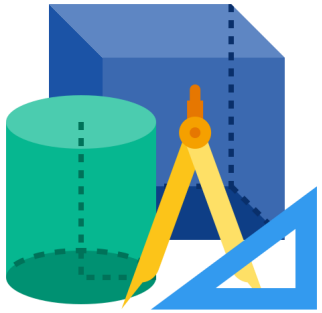
Audio encoders lack geometric grounding



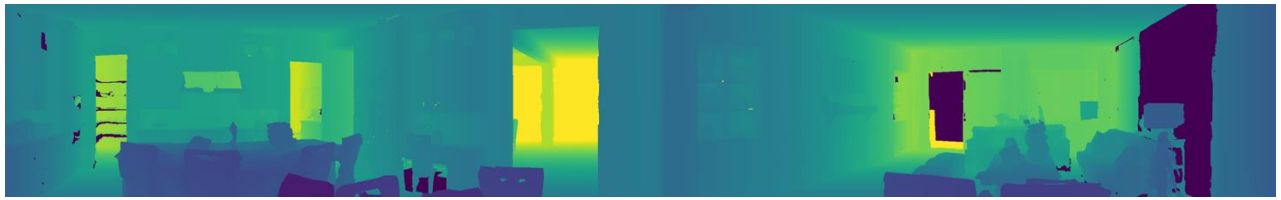
ALLMs use single-pass reasoning

- Map questions directly to answers without intermediate inference
- Do not decompose complex acoustic queries into smaller, interpretable subproblems

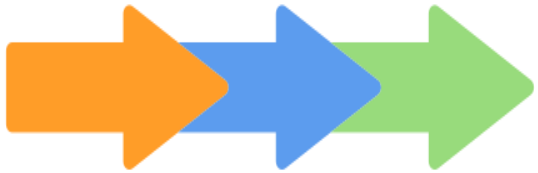
Why Audio LLMs Suffer in Spatial Reasoning



Audio encoders lack geometric grounding



Can we get geometric grounding from visual cues during training?

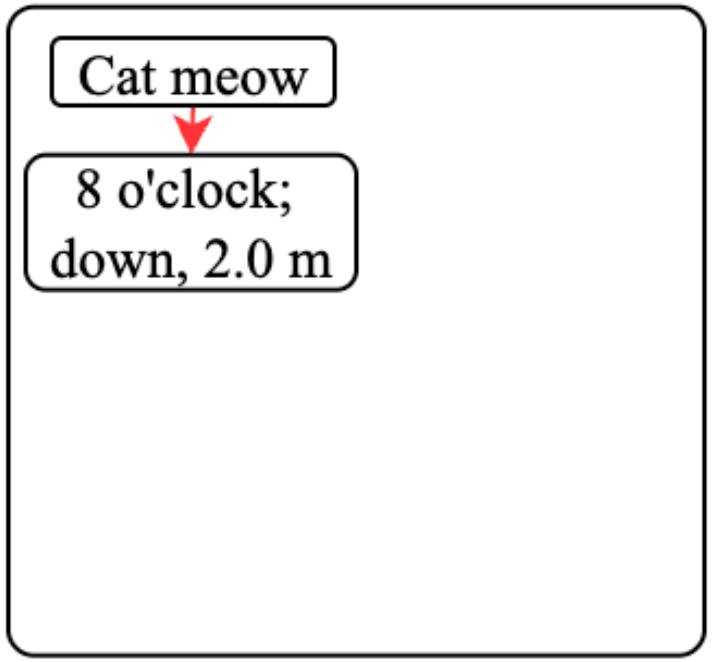


ALLMs use single-pass reasoning

Can Chain of Thought (CoT) improve the reasoning capability during acoustic localization?

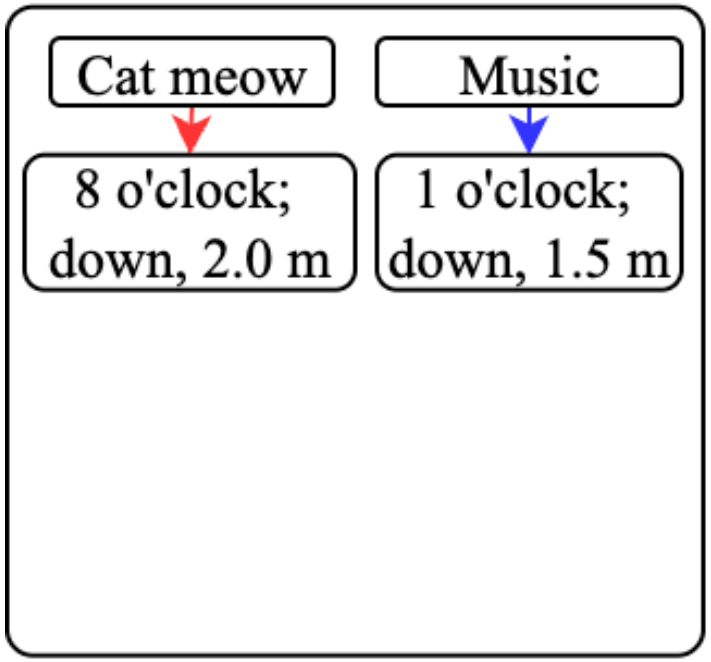
How OWL Functions?

Q: Which sound source is left to the listener?



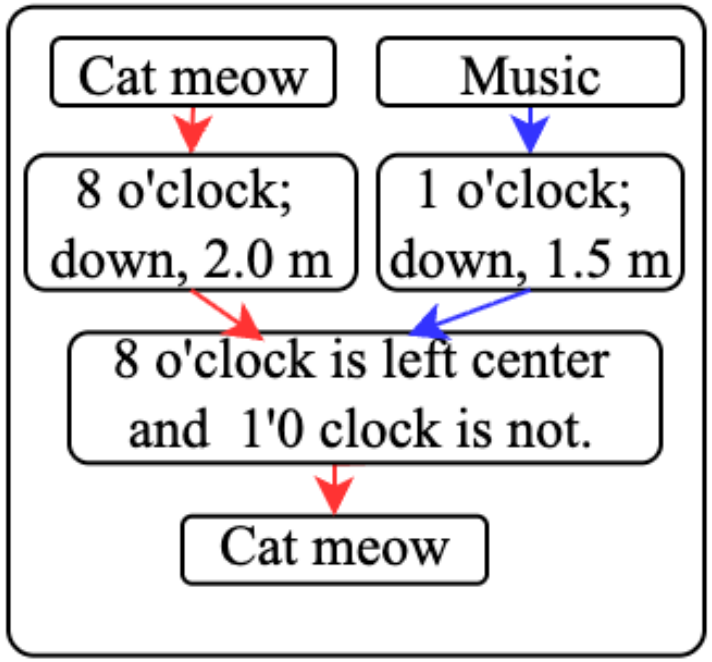
How OWL Functions?

Q: Which sound source is left to the listener?



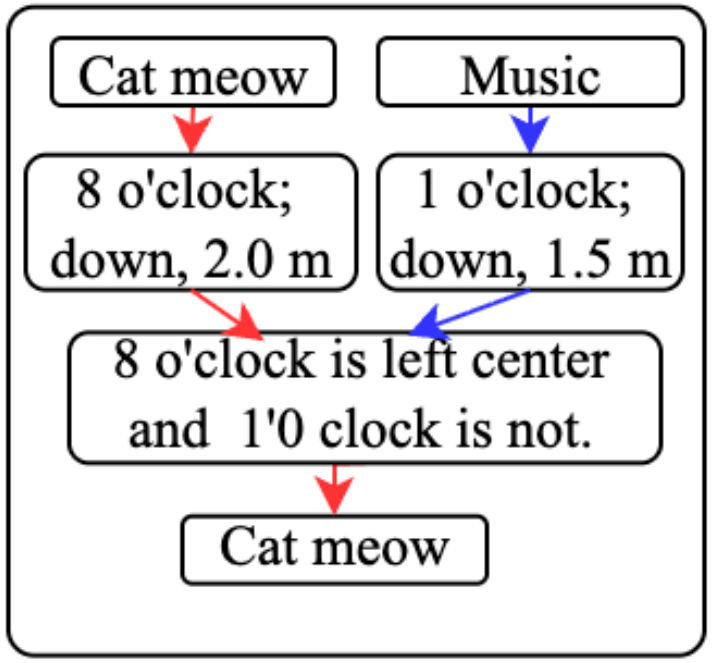
How OWL Functions?

Q: Which sound source is left to the listener?



How OWL Functions?

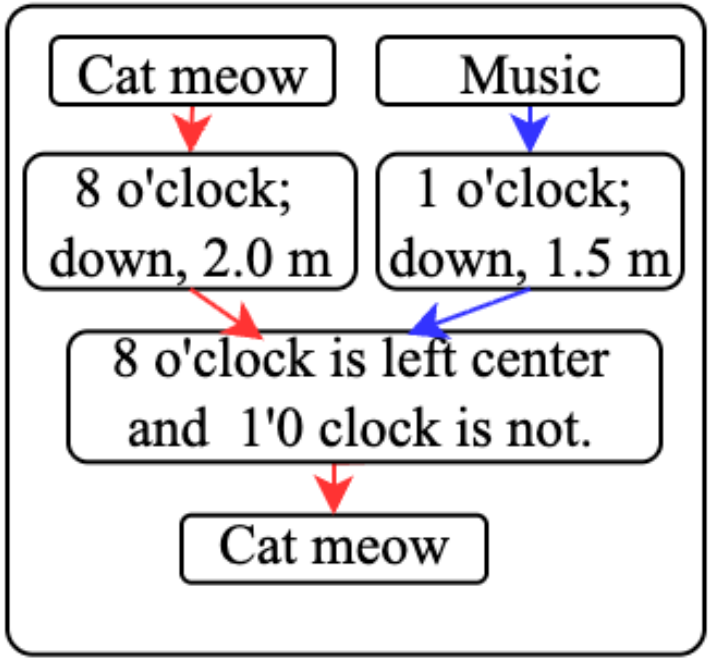
Q: Which sound source is left to the listener?



Multi-Step Reasoning Path

How OWL Functions?

Q: Which sound source is left to the listener?



Multi-Step Reasoning Path

Gemini-2.5-Flash: From my listening position direction of the sound of the cat's meow is very much centered. It doesn't pan significantly left or right.

OWL: Cat meow originates from 8 o'clock while music is at 1 o'clock. Therefore, Cat meow is on the left side of the listener.

Our Contribution

SAGE: Spatial-Acoustic Geometry Encoder

- A novel geometry-aware audio encoder.
- Trained with multimodal supervision.
- At inference only requires binaural audio

Our Contribution

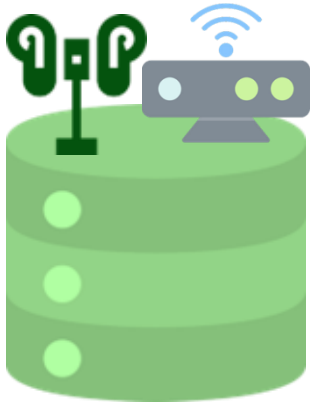
OWL: Spatial Audio-LLM with Chain-of-Thought Reasoning

- Integrates **SAGE** with **spatial Chain-of-Thought reasoning**.
- unifying event detection, localization, and structured inference.

SAGE: Spatial-Acoustic Geometry Encoder

Our Contribution

OWL: Spatial Audio-LLM with Chain-of-Thought Reasoning



SAGE: Spatial-Acoustic Geometry Encoder

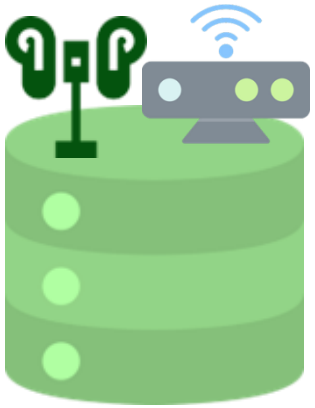
BiDepth Dataset

- Contains **binaural audio, RIRs, depth images, and QA**.
- First large-scale dataset with **1.1M QA pairs**.
- Enables **geometric grounding** for perception and **multi-step spatial reasoning**.

Our Contribution

OWL: Spatial Audio-LLM with Chain-of-Thought Reasoning

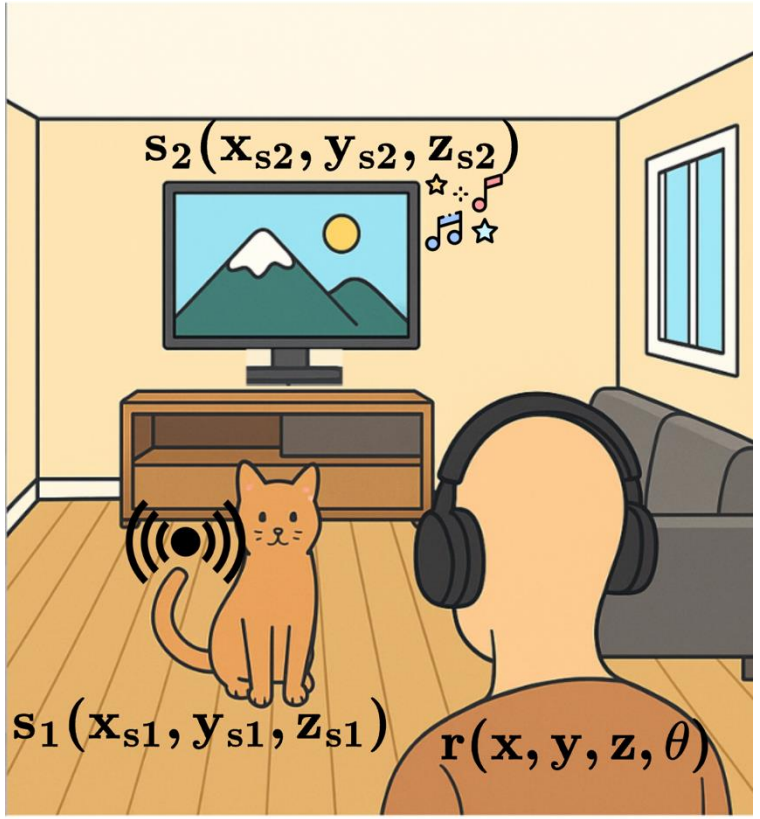
SAGE: Spatial-Acoustic Geometry Encoder



BiDepth Dataset

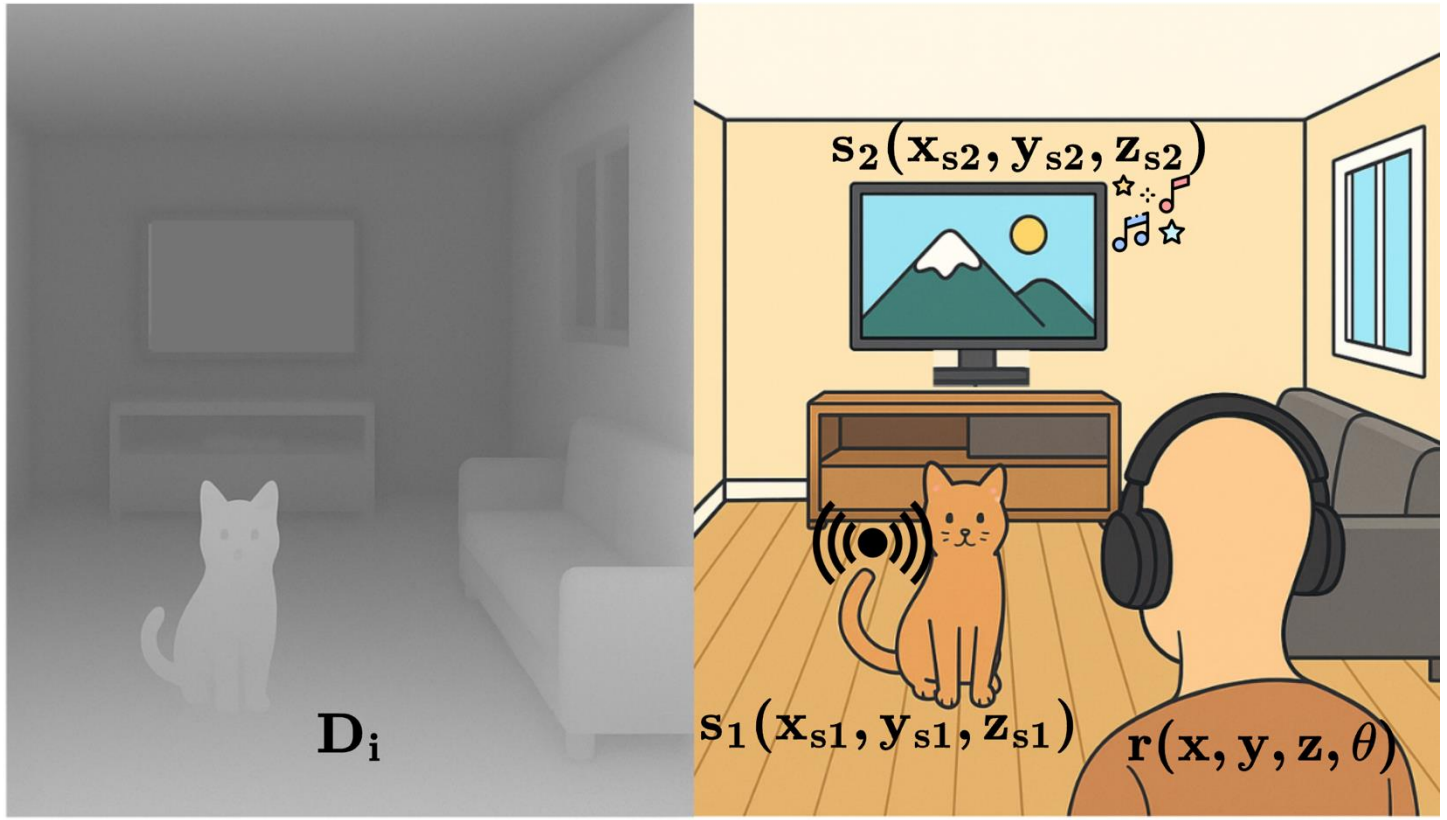
BiDepth Generation

$$B^r(t) = \begin{bmatrix} B_L^r(t) \\ B_R^r(t) \end{bmatrix} = \begin{bmatrix} \mathrm{RIR}_L(t, s, r, \gamma) \\ \mathrm{RIR}_R(t, s, r, \gamma) \end{bmatrix} \circledast M^s(t)$$

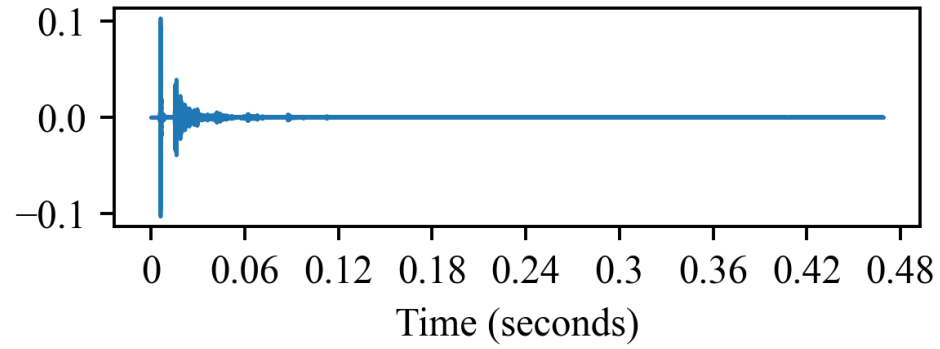


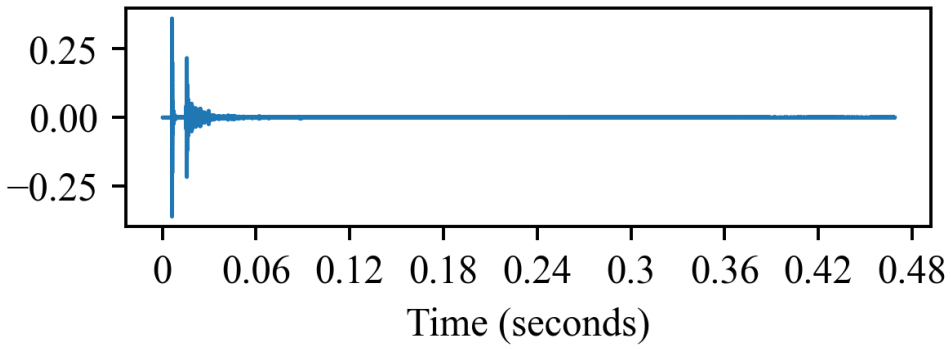
BiDepth Generation

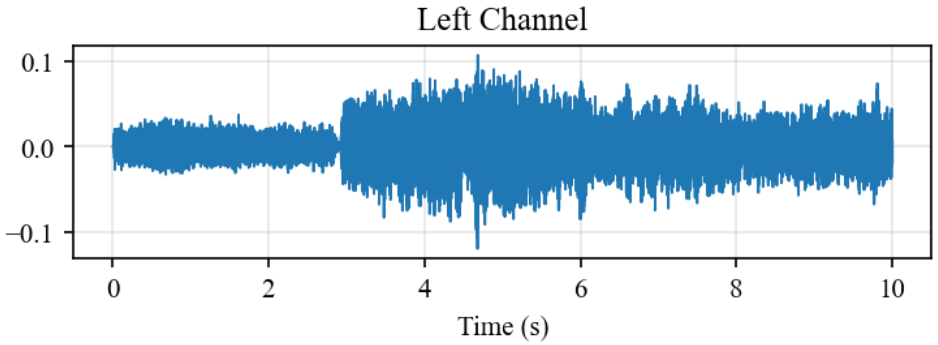
$$B^r(t) = \begin{bmatrix} B_L^r(t) \\ B_R^r(t) \end{bmatrix} = \begin{bmatrix} \mathrm{RIR}_L(t, s, r, \gamma) \\ \mathrm{RIR}_R(t, s, r, \gamma) \end{bmatrix} \circledast M^s(t)$$

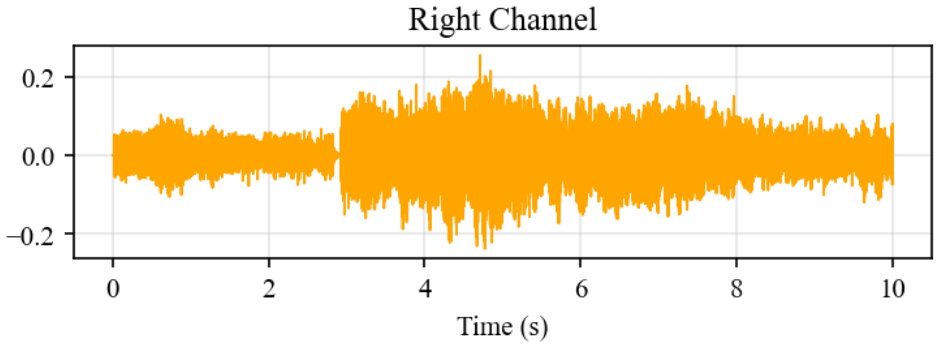


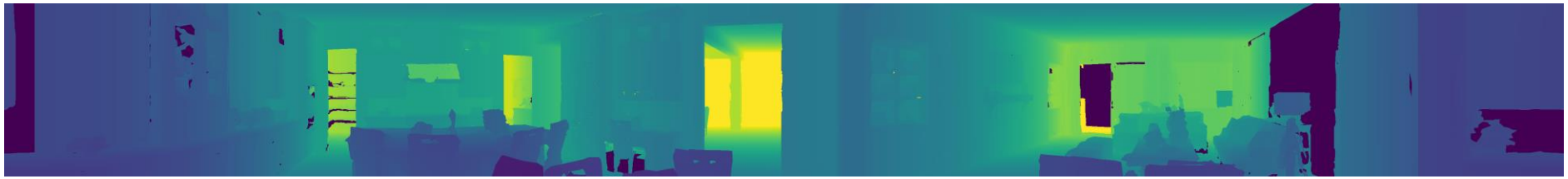
BiDepth Generation



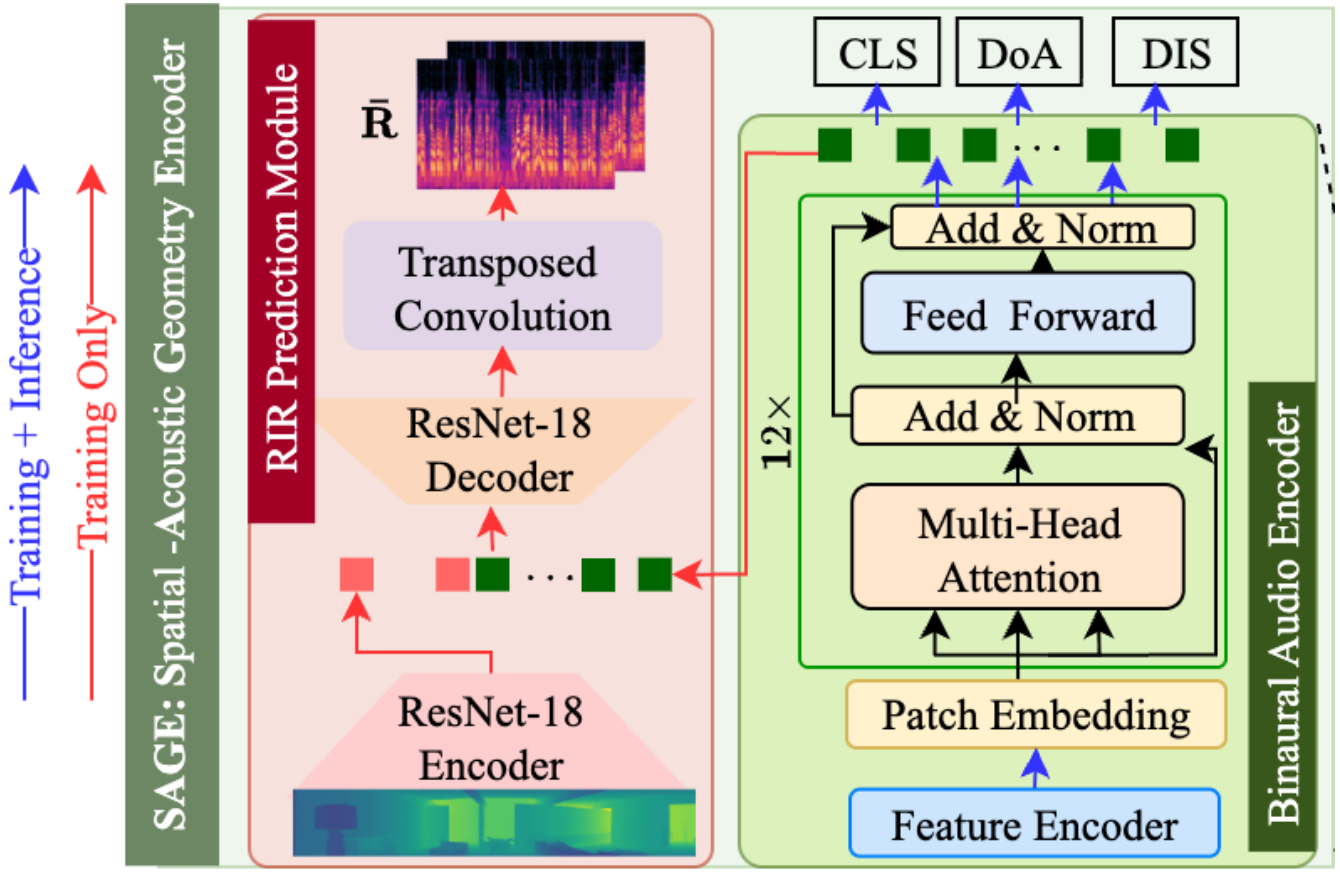




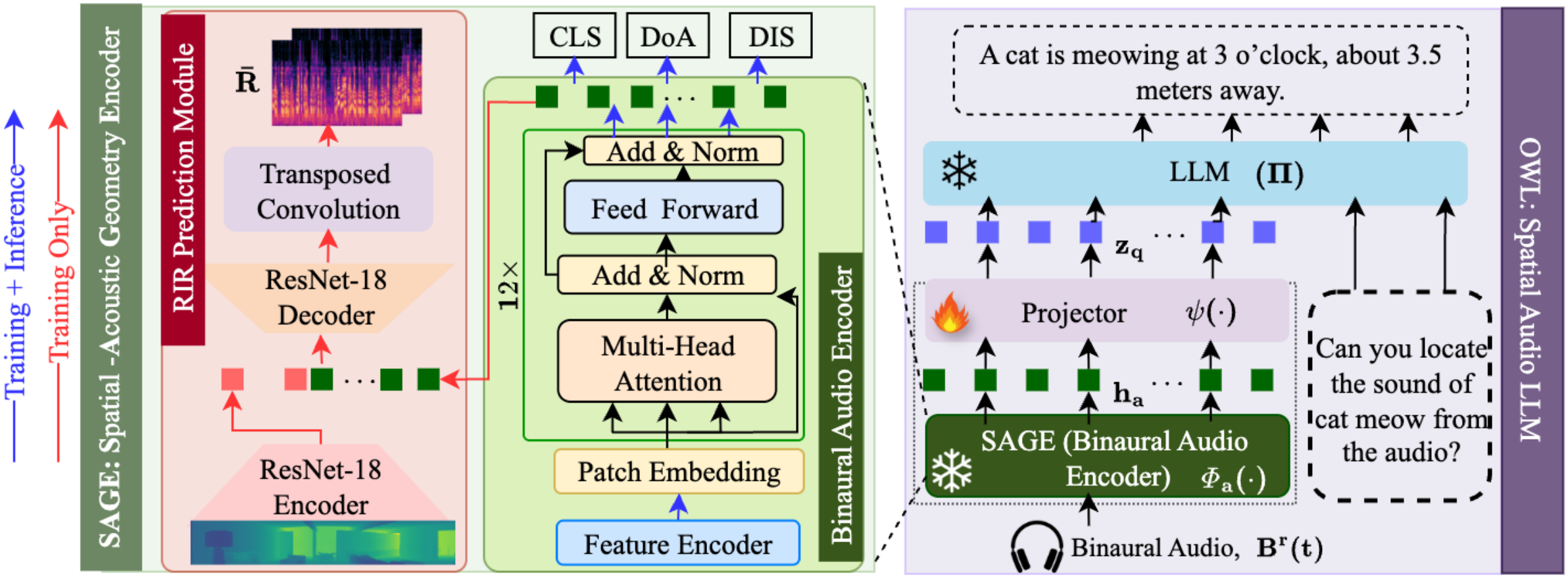




SAGE: Spatial-Acoustic Geometry Encoder



OWL Architecture



Performance of **SAGE** on SELD

Method	Modality		SpatialSound-QA (SSQA)				BiDepth			
	Audio	Depth	mAP ↑	$ER_{20°}$ ↓	MAE ↓	DER ↓	mAP ↑	$ER_{20°}$ ↓	MAE ↓	DER ↓
SELDNet	✓	✗	42.66	25.19	19.21	38.46	39.46	53.21	38.71	53.38
Spatial-AST[1]	✓	✗	**50.03**	<u>23.89</u>	**17.94**	<u>32.54</u>	48.97	45.29	32.99	47.82
Spatial-AST[2]	✓	✗	-	-	-	-	49.17	41.94	27.24	39.21
SAGE[3]	✓	✗	49.71	26.59	23.19	33.03	<u>49.75</u>	<u>36.89</u>	<u>26.32</u>	<u>17.11</u>
SAGE[4]	✓	✗	<u>49.94</u>	**23.67**	<u>18.26</u>	32.61	-	-	-	-
SAGE[5]	✓	✓	49.93	24.71	18.47	**17.84**	**49.81**	**28.13**	**21.67**	**14.32**

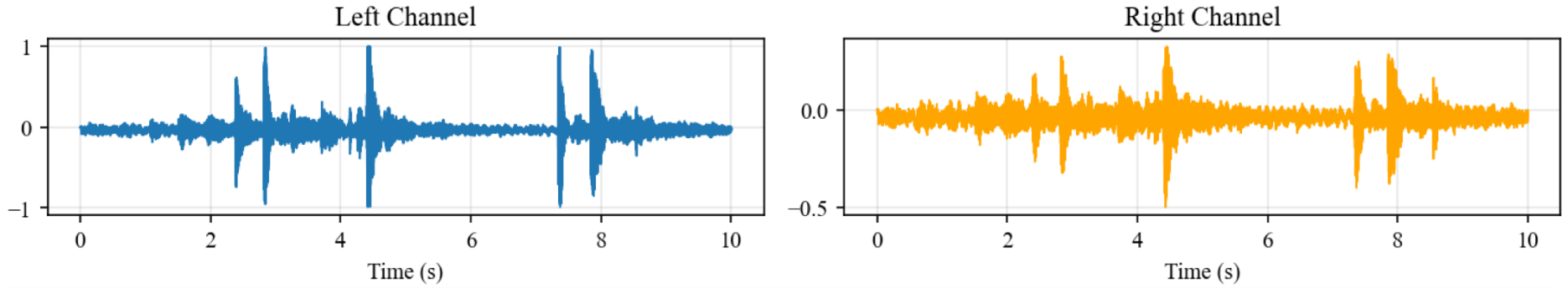
[1] Trained on SpatialSoundQA. [2] Trained on SpatialSoundQA and fine-tuned on BiDepth. [3] Trained on BiDepth audio only. [4] Pre-trained on BiDepth audio only, then fine-tuned on SpatialSoundQA. [5] Trained of BiDepth audio and depth.

Performance of **OWL** on QA Task (on BiDepth)

Method	TypeI				TypeII		TypeIII	TypeIV		
	Detection (mAP)↑		DoA (Acc) ↑		Distance (DER)↓					
	Single Source	Double source	Single Source	Double source	Single Source	Double source	BA ↑	Detection ↑	Direction ↑	BA ↑
Closed-source Models†										
Gemini1.5Pro	31.19	12.71	-	-	-	-	-	11.96	-	-
Gemini2.5Pro	32.47	12.17	-	-	-	-	-	12.01	-	-
Gemini2.5Flash	32.91	12.29	-	-	-	-	-	12.21	-	-
Open-source Models										
VideoLLaMA2	17.11	5.21	12.23	11.76	68.12	83.78	8.29	5.19	-	-
RAVEN	16.29	5.43	13.79	9.39	71.46	82.37	9.76	5.97	-	-
AudioFlamingo2	27.59	6.73	17.74	14.17	54.62	68.91	19.54	7.59	-	-
BAT	24.97	8.73	-|71.59 *	-|35.29*	28.61	45.79	69.46	71.62	78.27	61.29
OWL w/o CoT	33.31	17.24	46.15|77.21*	34.24|51.67*	24.67	31.29	74.29	-	-	65.27
OWL w CoT	**33.37**	**17.26**	**46.17**	**34.31**	**23.29**	**29.91**	**77.89**	**79.04**	**86.76**	**76.53**

Performance of **OWL** on QA Task (on SpatialSoundQA)

Model	Perception (Type ABCD)						Reasoning (Type E)		
	Detection (mAP) ↑		DoA (Acc) ↑		DP (DER) ↓		Direction ↑	Distances ↑	Avg ↑
	Type A	Type C	Type B	Type D	Type B	Type D			
Random	0.61	0.59	12.57	12.41	67.33	67.46	50.00	50.00	50.00
Mono BAT	24.15	6.42	14.31	11.93	34.17	56.26	57.69	51.36	54.33
BAT	26.34	9.89	75.54	37.65	29.16	47.90	69.77	84.04	76.89
OWL	**26.76**	**12.73**	**78.31**	**43.15**	**26.14**	**43.21**	**71.21**	**86.91**	**79.06**



Scene Description: 2 sounds (Music, Electric Piano) placed in a small room.

Question : From the receiver's perspective, what sound is coming from the left?

GT Answer : Since Music is at eight o' clock and Electric piano is at seven o' clock, both sources are on the left side of the receiver.

: Relative to the receiver, Keyboard (musical) and Music are detected at seven o' clock and eight o' clock. Thus, they both lie on the left.

Example of a left-right spatial reasoning question. Two concurrent sounds (Music and Electric Piano) are placed in a small room, and the system identifies both as originating from the left side of the receiver.