Learning Spatially-Aware Language and Audio Embeddings

Bhavika Devnani, Skyler Seto, Zak Aldeneh, Alessandro Toso, Elena Menyalenko, Barry-John Theobald, Jonathan Sheaffer, Miguel Sarabia

NeurIPS I Apple I 2024

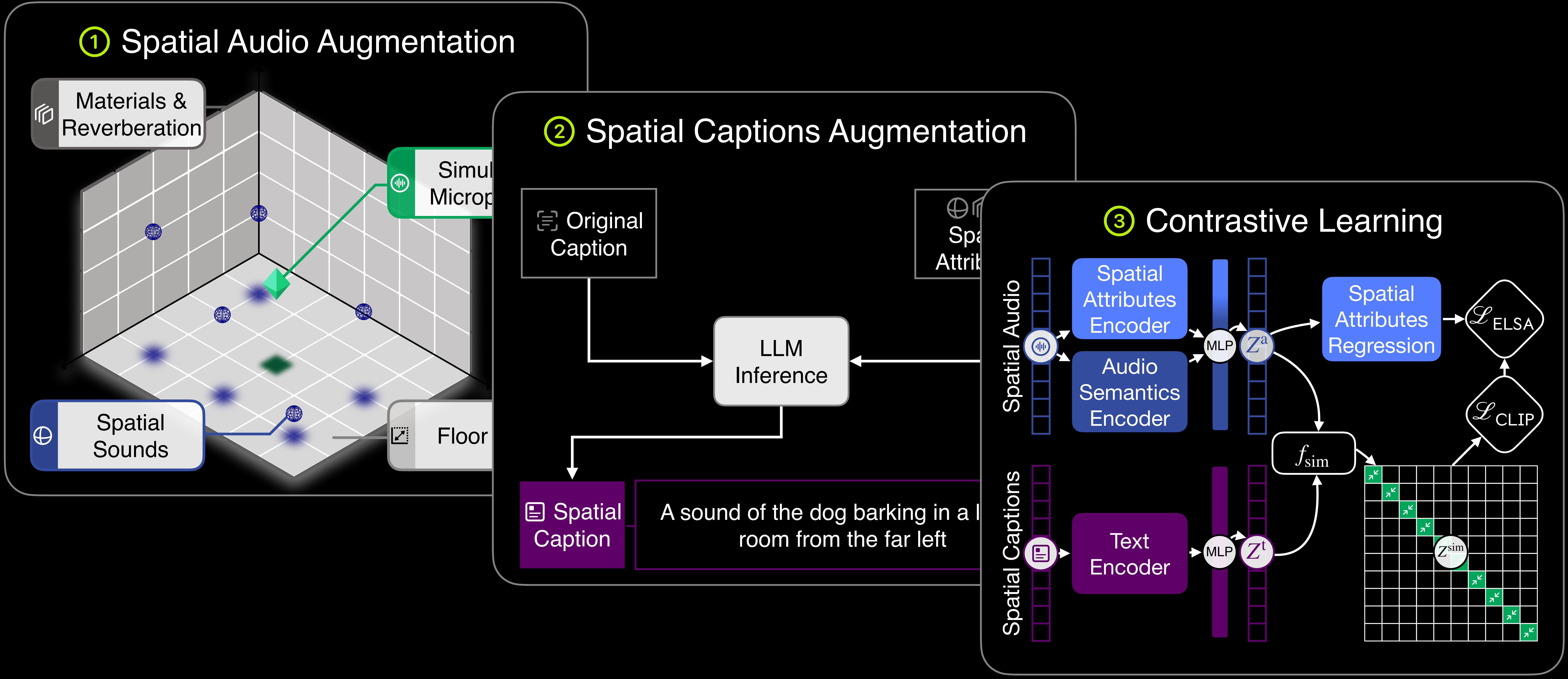
Motivation

Picture this situation:

The semantics of the audio are as important as the spatial attributes!

OBJECTIVE: align **semantic** and **spatial** attributes of audio with **natural language**

Learning ELSA Embeddings



① Spatial Audio Augmentation

Materials & Reverberation

Simul

Microp

Spatial Sounds

Floor

② Spatial Captions Augmentation

Original Caption

Spa

Attrib

LLM Inference

Spatial Caption

A sound of the dog barking in a l
room from the far left

③ Contrastive Learning

Spatial Audio

Spatial Attributes Encoder

Audio Semantics Encoder

MLP

Z^a

Spatial Attributes Regression

$\mathscr{L}_{\text{ELSA}}$

$\mathscr{L}_{\text{CLIP}}$

f_{sim}

Spatial Captions

Text Encoder

MLP

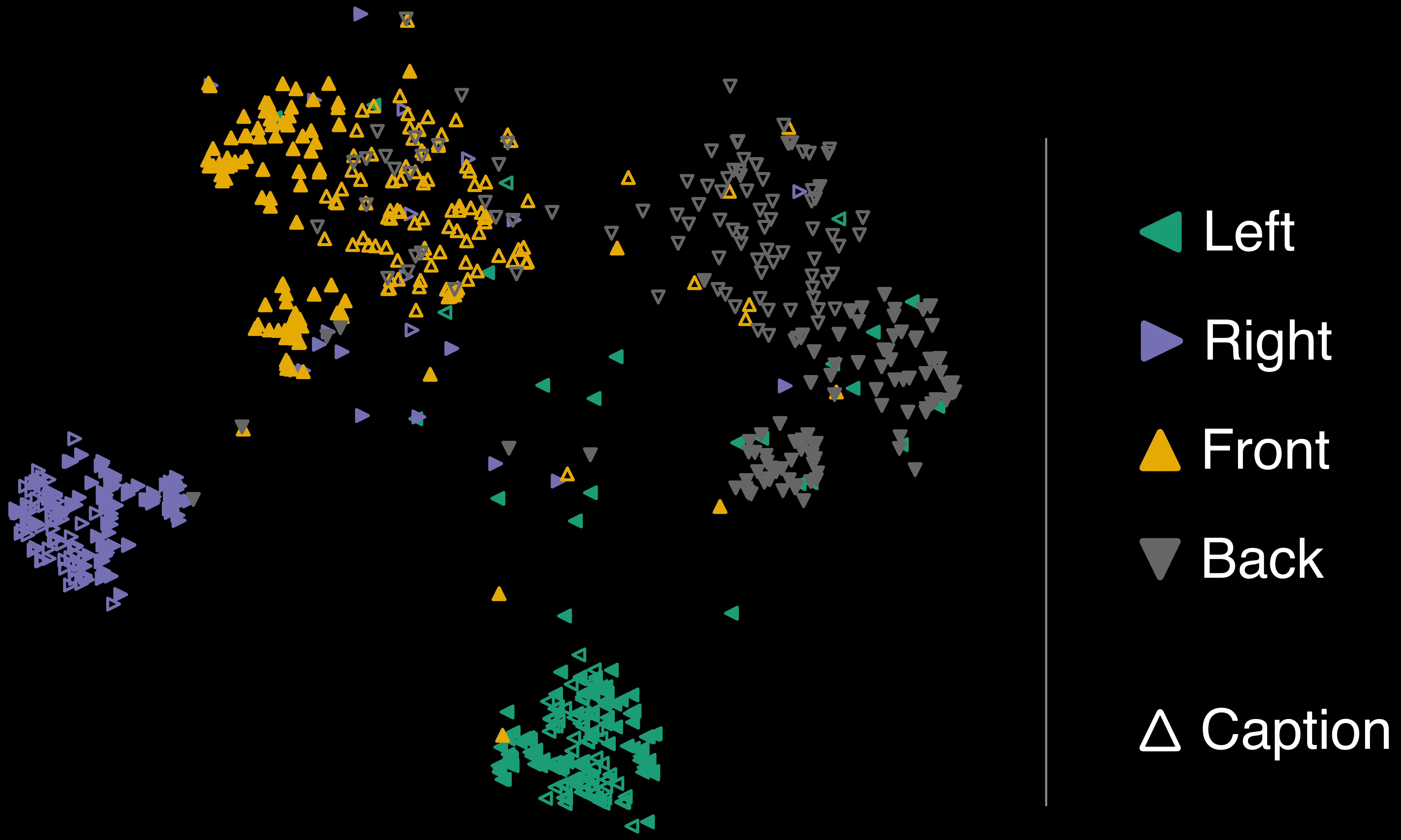
Z^t

Z^{sim}

ELSA Embeddings Evaluation

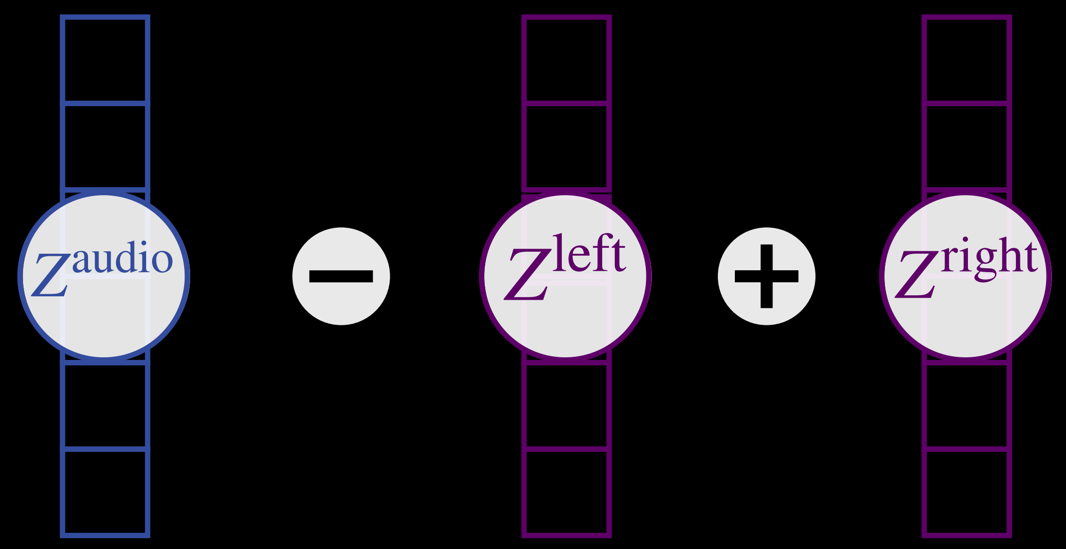
	Semantic Capabilities	Spatial Capabilities	Semantic Retrival mAP@10 ↑ AudioCaps	3D Source Localization Mean Absolute Error ↓ REAL TUT Sound Events 2018
SeldNET	✗ Fixed Vocabulary	✓	✗	26.6°
PILOT	✗ Fixed Vocabulary	✓	✗	4.2°
LAION CLAP	✓ Open Vocabulary	✗	43.8%	95.3°
ELSA (ours)	✓ **Open Vocabulary**	✓	**44.2%**	**15.0°**

Learning ELSA Embeddings



Understanding ELSA Embeddings

Swapping Spatial Directions



99.6% direction accuracy
& maintains audio semantics

Automatic Captioning

ELSA spatial audio embeddings
to GPT2 input tokens

Generated Caption from Audio

The sound of water flowing and splashing is emanating from the front of a room.