NeuroAI

NeurIPS 2026 Spotlight

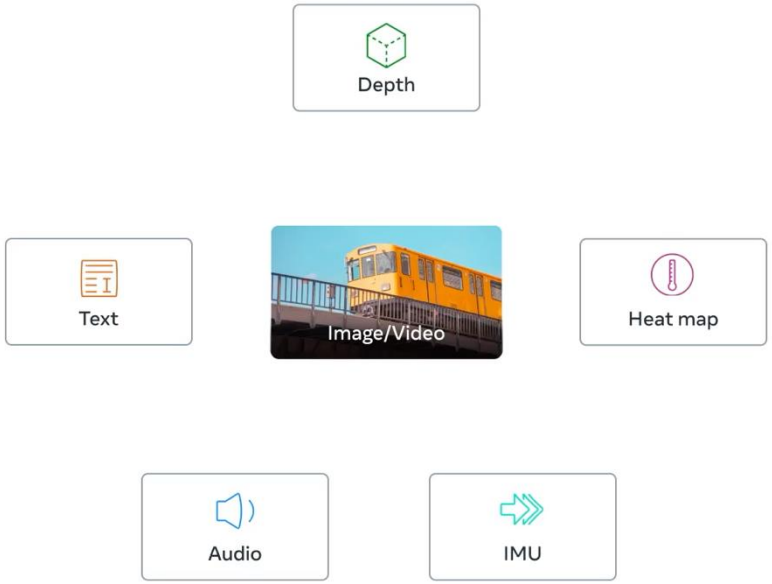
SceneBind: Binding What and Where Across Vision, Audio and Language

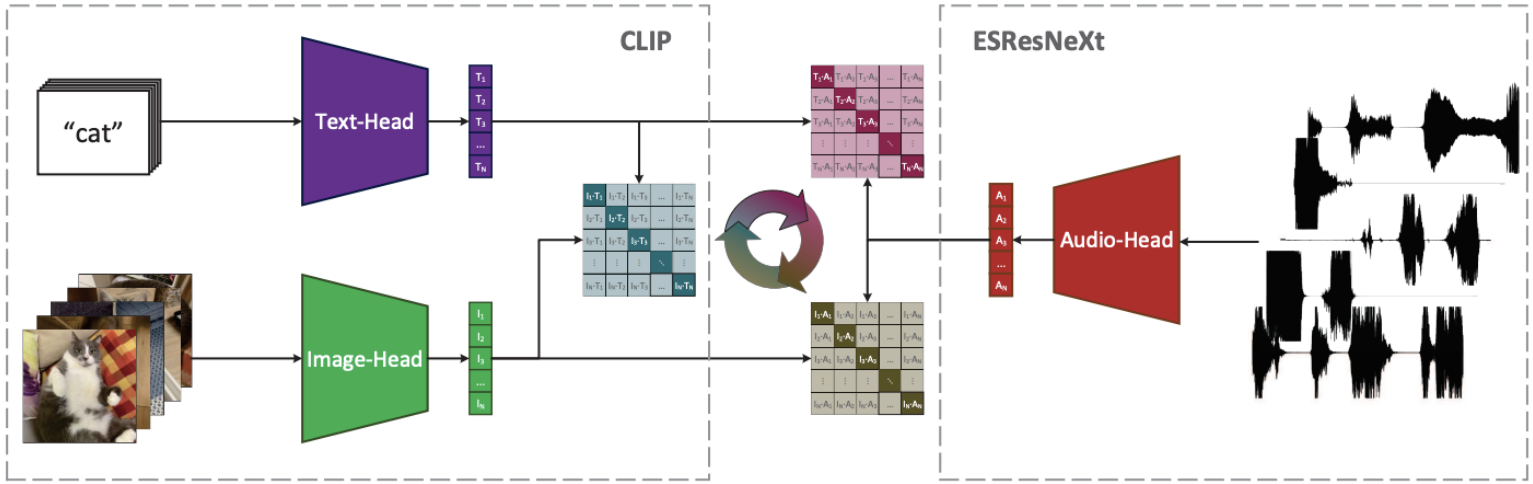
Mingfei Chen[1], Zijun Cui[2], Ruoke Zhang[1], Hyeonggon Ryu[3], Eli Shlizerman[1]

[1]University of Washington [2]University of Texas at Dallas [3]Hankuk University of Foreign Studies

Audio-Visual-Text Representation

Semantic Only!





Bridge via Visual: ImageBind [1]

Bridge via Text: AudioClip [2]

[1] R. Girdhar, et al. ImageBind: One Embedding Space To Bind Them All. CVPR23.
[2] A. Guzhov, et al. AudioCLIP: Extending CLIP to Image, Text and Audio. GCPR21.

SceneBind: Binding What and Where Across Vision, Audio and Language





"A white truck passes a pedestrian on a narrow urban street."

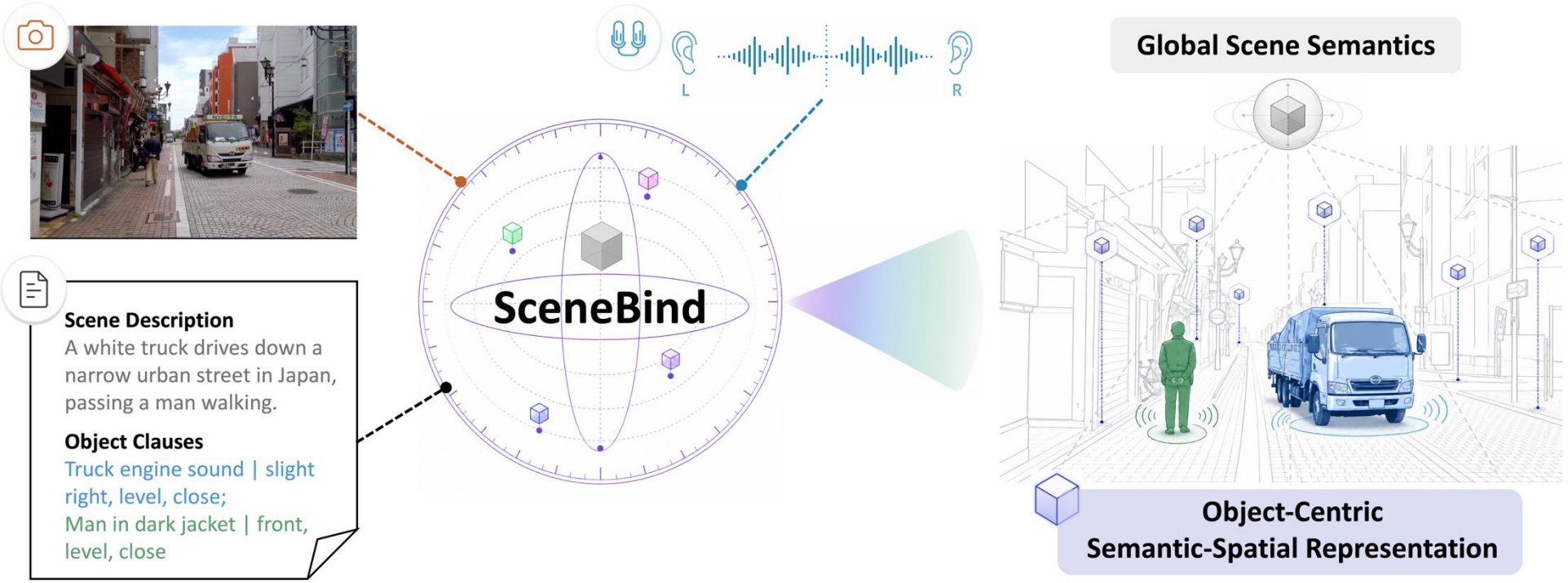
Truck engine sound

slight right | level | close

Man in dark jacket

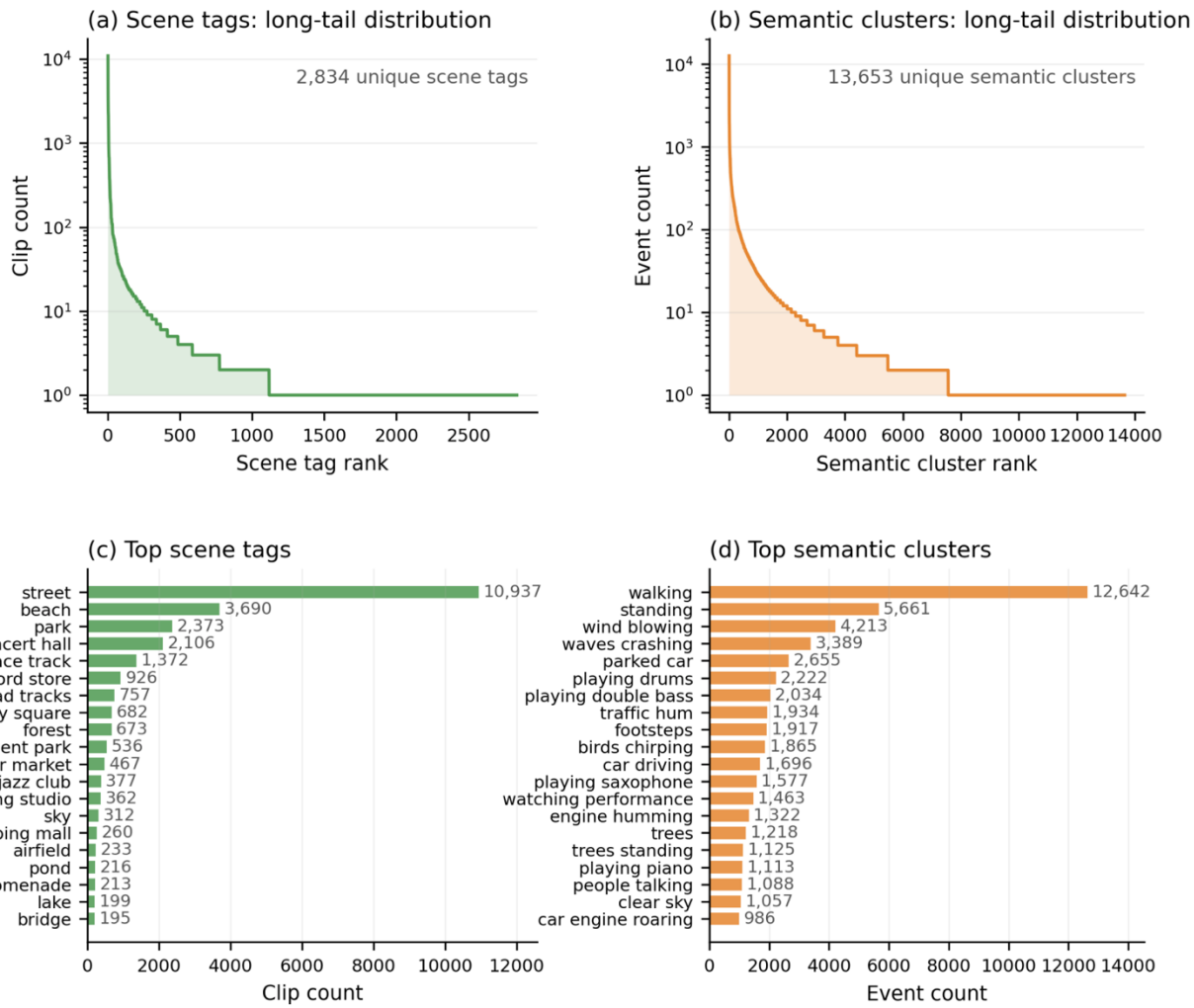
front | level | close

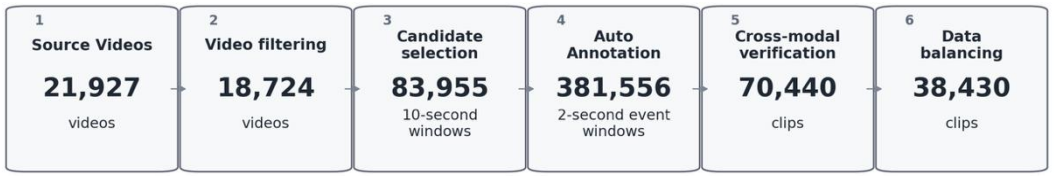
SceneBind: Binding What and Where Across Vision, Audio and Language

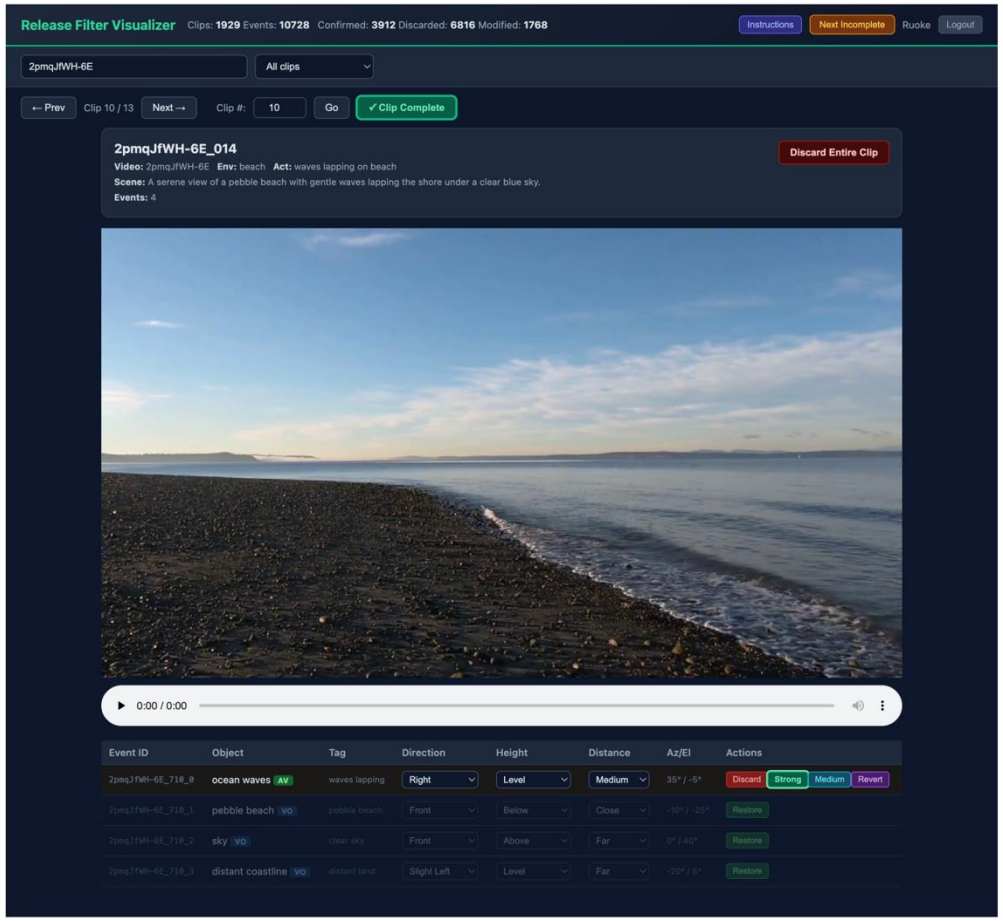


- ✓ Omni-modal
- ✓ Real Scenes
- ✓ Spatial-Semantic

Dataset Curation

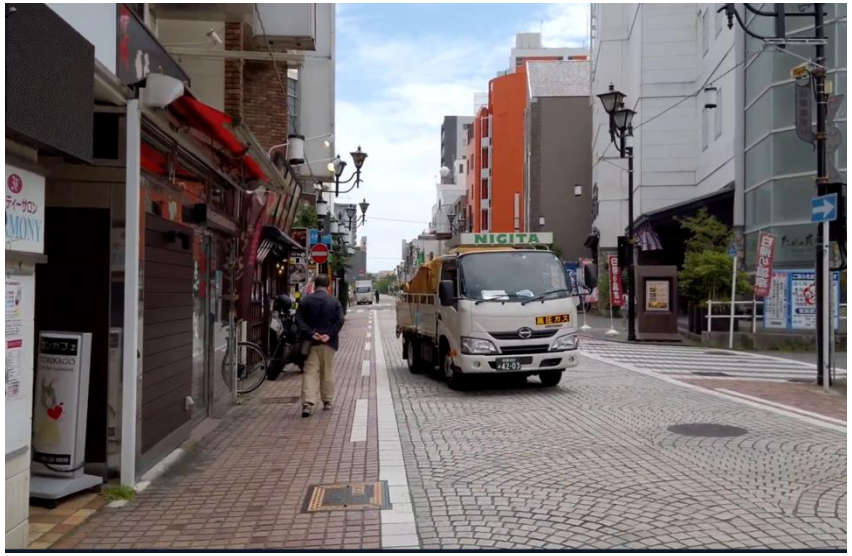






Dataset Curation

38,430 Data Clips





2s

"A white truck passes a pedestrian on a narrow urban street."

Truck engine sound

slight right | level | close

Man in dark jacket

front | level | close

SceneBind Spatial-Semantic Representation



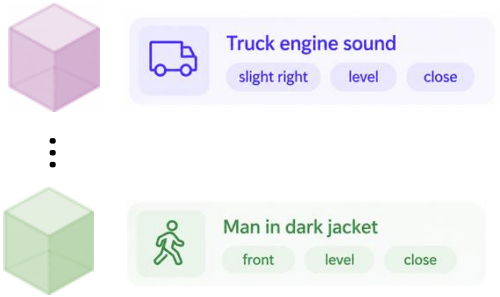


"A white truck passes a pedestrian on a narrow urban street."

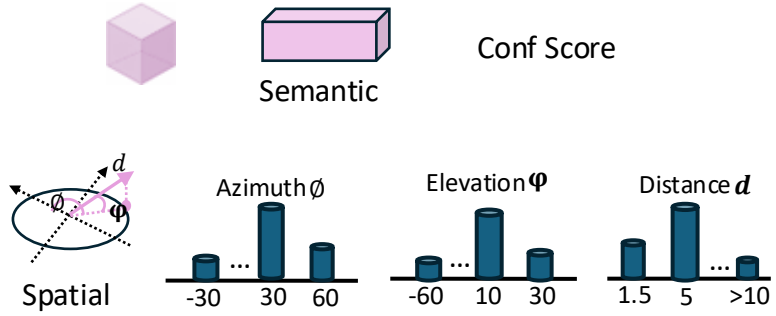
Global Semantic

+

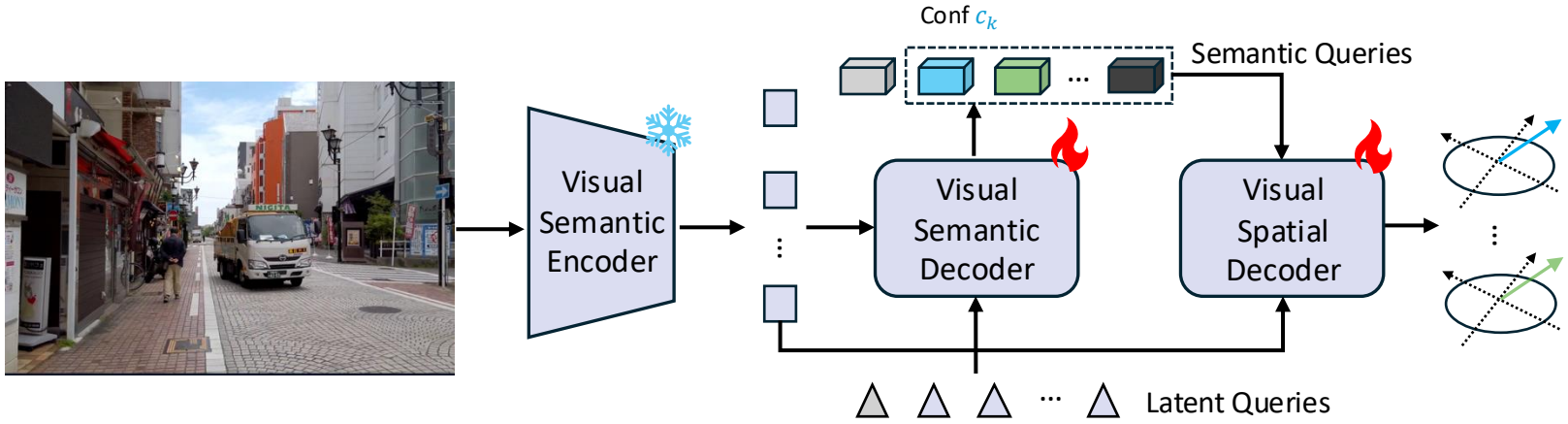
K Object-centric Slots (Spatial-Semantic)



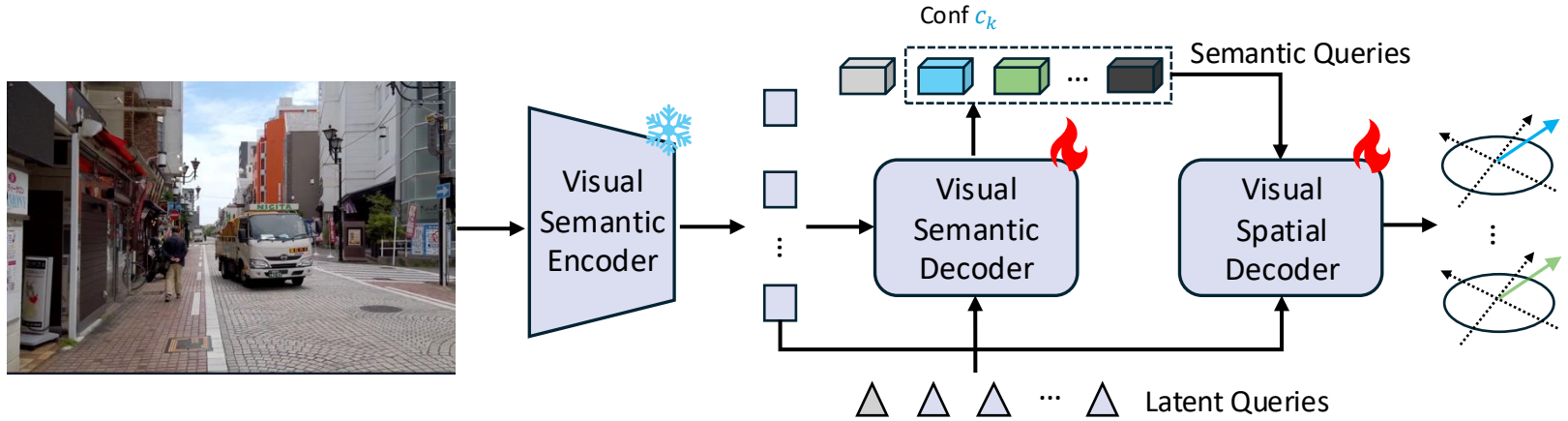


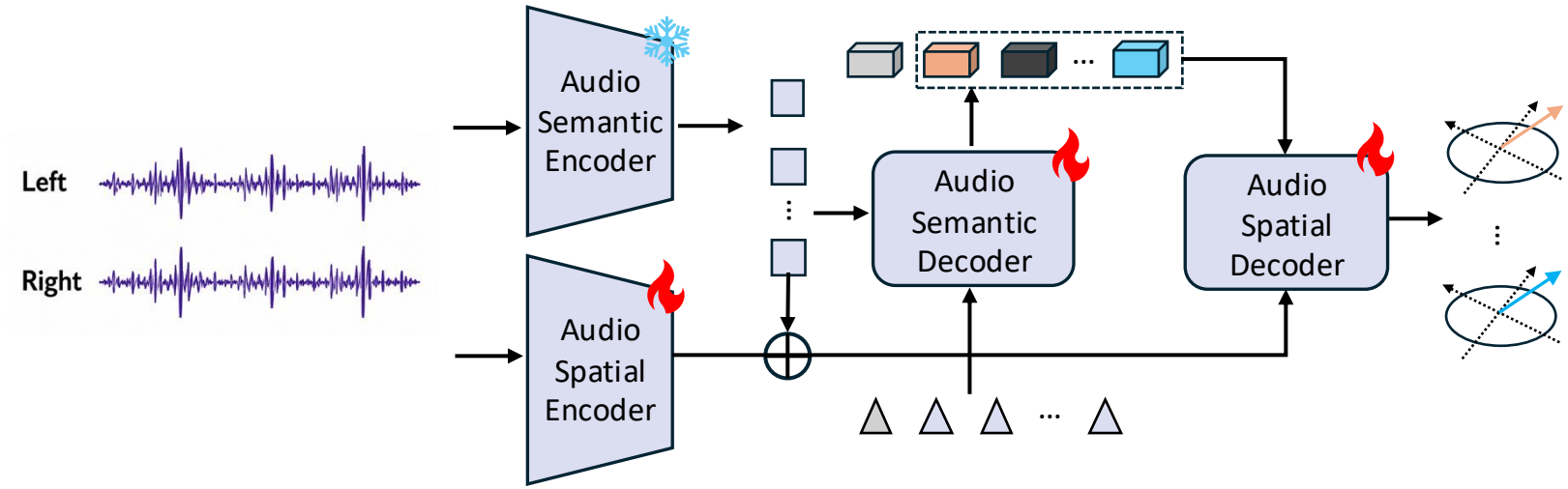


SceneBind Encoding



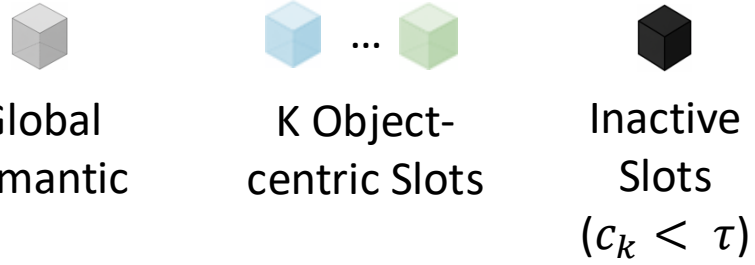
SceneBind Encoding



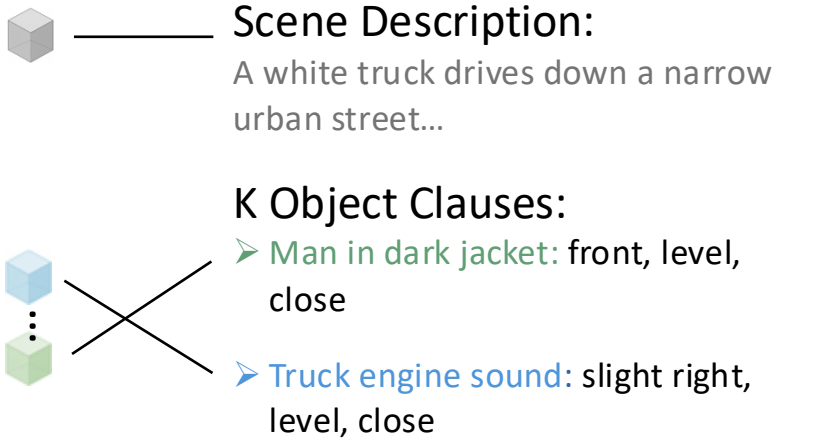


SceneBind Supervision

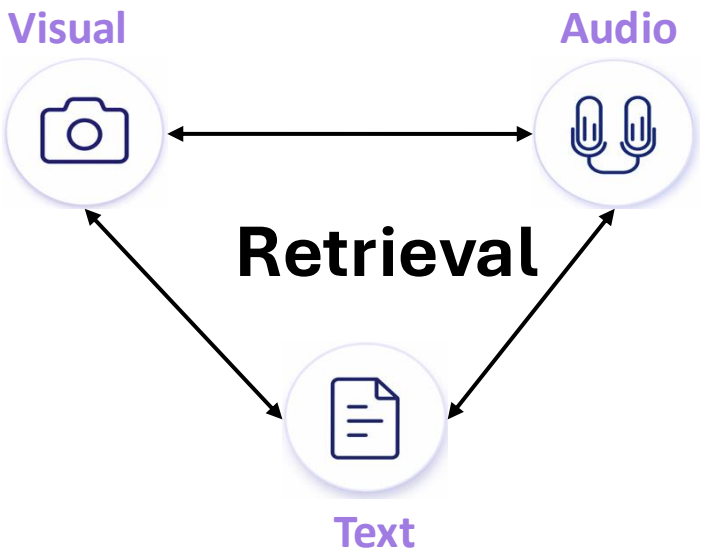
SceneBind Semantic-Spatial Representation

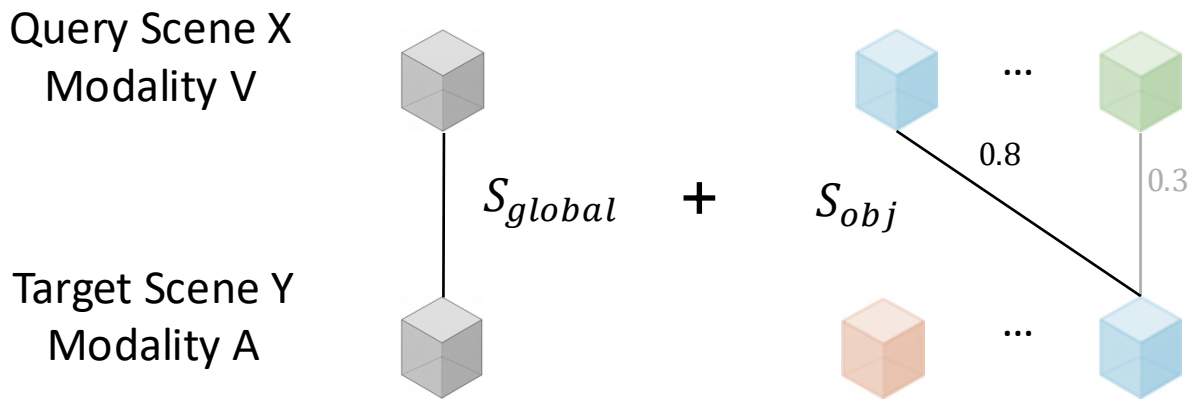


Bipartite Matching Supervision



Spatial-Semantic omni-modal retrieval and scorer





Cross-Modal Spatial-Semantic Retrieval



AUDIO TO IMAGE

Waves on the right, buildings on the left

A wide shot of a beach with waves crashing on the shore, buildings in the background, and a fishing rod in the foreground.

GLOBAL	OBJECT	JOINT
0.598	0.234	1.050

0:00 / 0:02

Cross-Modal Spatial-Semantic Retrieval



AUDIO TO IMAGE

Waves on the right, buildings on the left

A wide shot of a beach with waves crashing on the shore, buildings in the background, and a fishing rod in the foreground.

GLOBAL	OBJECT	JOINT
0.598	0.234	1.050

0:00 / 0:02



TEXT TO AUDIO

Saxophone left, guitar right

A woman plays the saxophone while a man plays the electric guitar on a stage with a red curtain backdrop and a banner.

GLOBAL	OBJECT	JOINT
0.435	0.302	1.455

0:00 / 0:02

Cross-Modal Spatial-Semantic Retrieval



AUDIO TO IMAGE

Waves on the right, buildings on the left

A wide shot of a beach with waves crashing on the shore, buildings in the background, and a fishing rod in the foreground.

GLOBAL	OBJECT	JOINT
0.598	0.234	1.050

0:00 / 0:02



TEXT TO AUDIO

Saxophone left, guitar right

A woman plays the saxophone while a man plays the electric guitar on a stage with a red curtain backdrop and a banner.

GLOBAL	OBJECT	JOINT
0.435	0.302	1.455

0:00 / 0:02



TEXT TO IMAGE

River below and to the right

A first-person perspective of walking along a paved sidewalk next to a stone-walled river on a cloudy day.

GLOBAL	OBJECT	JOINT
0.399	0.275	1.398

0:00 / 0:02



IMAGE TO TEXT

Church bells above the square

A bustling public square in front of a large church with bells ringing and a fountain with an obelisk.

GLOBAL	OBJECT	JOINT
0.340	0.260	1.443

0:00 / 0:02

Cross-Modal Spatial-Semantic Retrieval

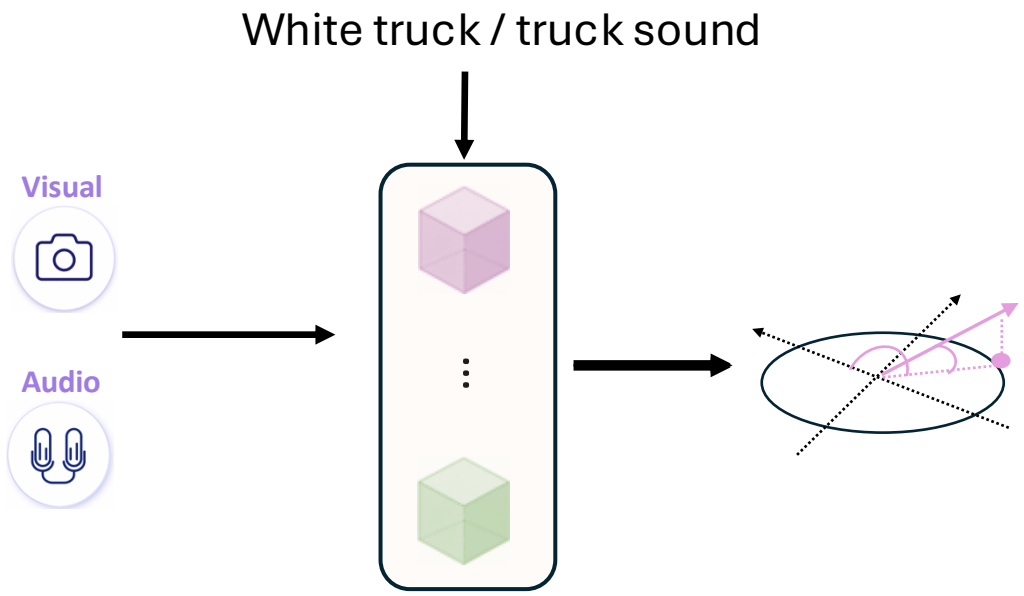
Encoder		Scene Retrieval (R@1 ↑)				Spatial Retrieval (↑)		
Audio–Text	Visual–Text	A↔V	A↔T	V↔T	Avg	R@1	mAP	Avg
AudioCLIP [11]	AudioCLIP [11]	0.3	0.3	20.6	7.1	9.3	28.1	18.7
SpatialCLAP [10]	SigLIP2 [5]	0.4	1.9	51.2	17.8	10.1	28.0	19.0
LAION-CLAP [6]	SigLIP2 [5]	0.5	5.9	51.2	19.2	10.1	28.7	19.4
M2D-CLAP [7]	SigLIP2 [5]	0.3	9.0	51.2	20.2	11.4	29.6	20.5
ImageBind [8]	ImageBind [8]	7.8	7.6	48.0	21.2	10.6	28.7	19.6
M2D-CLAP* [7]	SigLIP2* [5]	19.3	11.2	42.8	24.4	10.1	28.9	19.5
ImageBind* [8]	ImageBind* [8]	19.3	10.9	35.4	21.9	10.4	29.4	19.9
SpatialCLAP* [10]	SigLIP2* [5]	16.1	9.3	44.2	23.2	10.6	29.2	19.9
SceneBind (Ours)	**SceneBind (Ours)**	**21.8**	**17.0**	**65.3**	**34.7**	**28.9**	**48.0**	**38.4**

Cross-modal scene retrieval (R@1) and object-level spatial retrieval on unseen Binaural test set. SceneBind outperforms all baselines, with largest gains on text-conditioned retrieval

Fast omni-modal perception: Object grounding





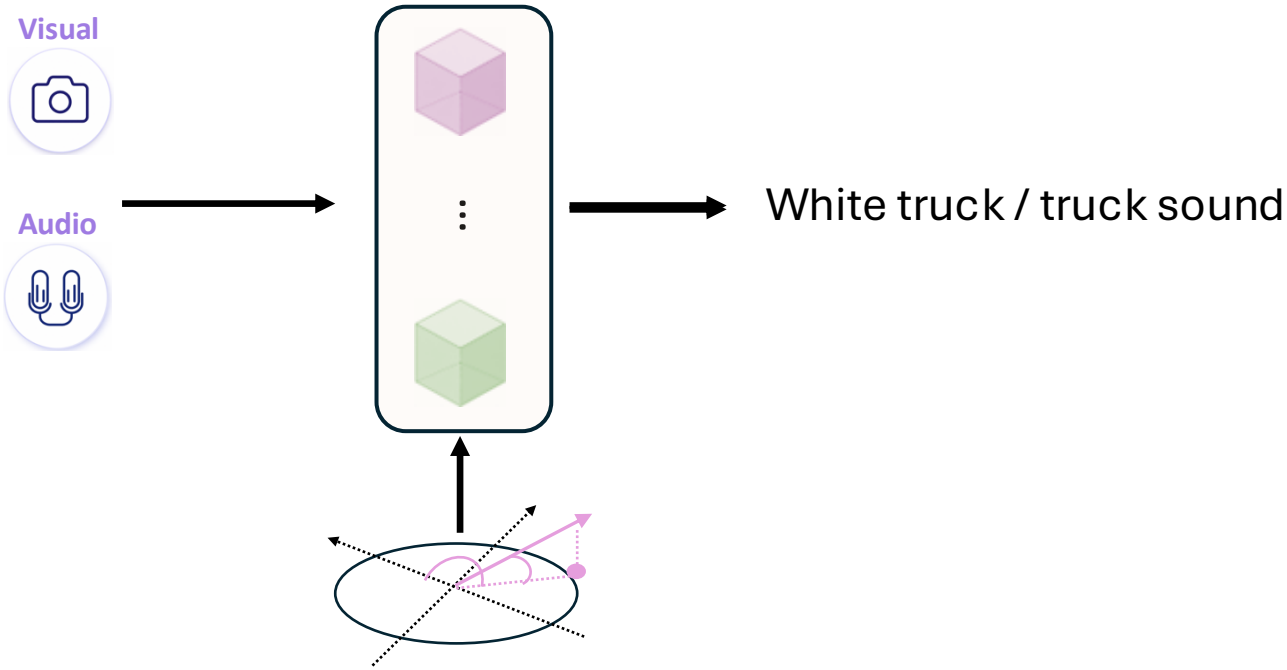


Safe, without need of heavy LLM and compute!!

Fast omni-modal perception: Spatial querying

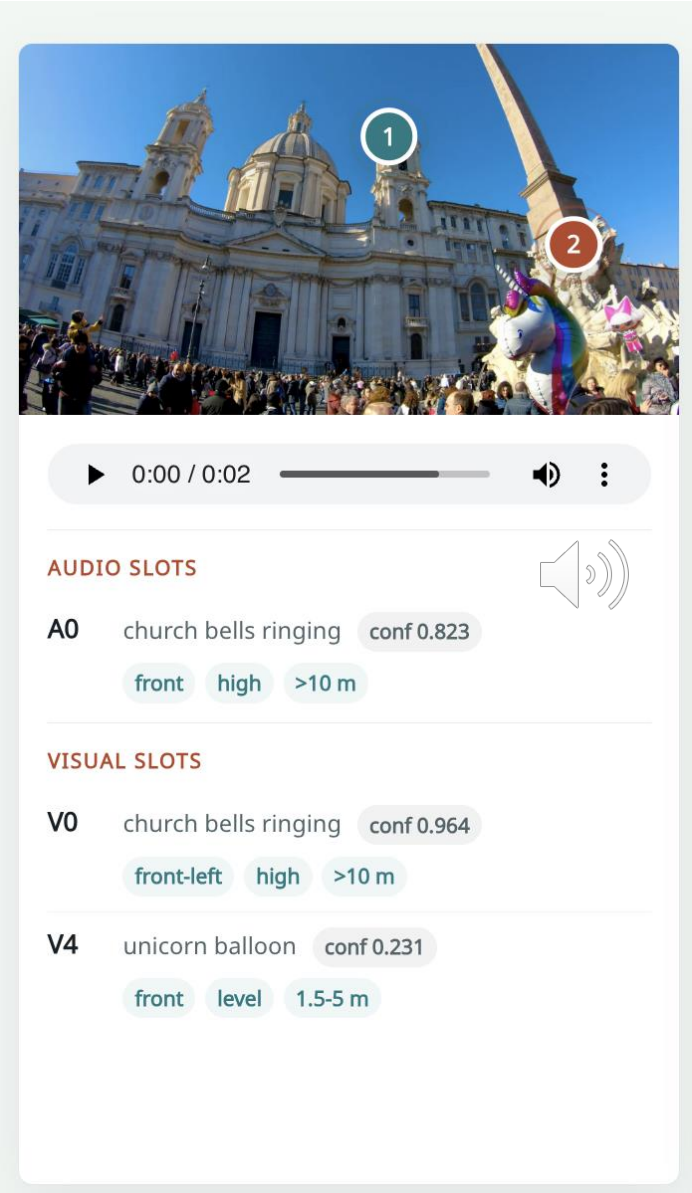






Safe, without need of heavy LLM and compute!!

Fast omni-modal perception



Fast omni-modal perception



0:00 / 0:02

AUDIO SLOTS

A0 church bells ringing conf 0.823
front high >10 m

VISUAL SLOTS

V0 church bells ringing conf 0.964
front-left high >10 m

V4 unicorn balloon conf 0.231
front level 1.5-5 m



0:00 / 0:02

AUDIO SLOTS

A0 person playing saxophone conf 0.986
front-left level 1.5-5 m

A1 person playing electric guitar conf 0.937
front-right level 1.5-5 m

VISUAL SLOTS

V0 person playing electric guitar conf 0.987
front-right level 1.5-5 m

V1 person playing saxophone conf 0.917
front-left level 1.5-5 m

Fast omni-modal perception



0:00 / 0:02

AUDIO SLOTS

A0 church bells ringing conf 0.823
front high >10 m

VISUAL SLOTS

V0 church bells ringing conf 0.964
front-left high >10 m

V4 unicorn balloon conf 0.231
front level 1.5-5 m



0:00 / 0:02

AUDIO SLOTS

A0 person playing saxophone conf 0.986
front-left level 1.5-5 m

A1 person playing electric guitar conf 0.937
front-right level 1.5-5 m

VISUAL SLOTS

V0 person playing electric guitar conf 0.987
front-right level 1.5-5 m

V1 person playing saxophone conf 0.917
front-left level 1.5-5 m



0:00 / 0:02

AUDIO SLOTS

A0 river conf 0.776
front-right level >10 m

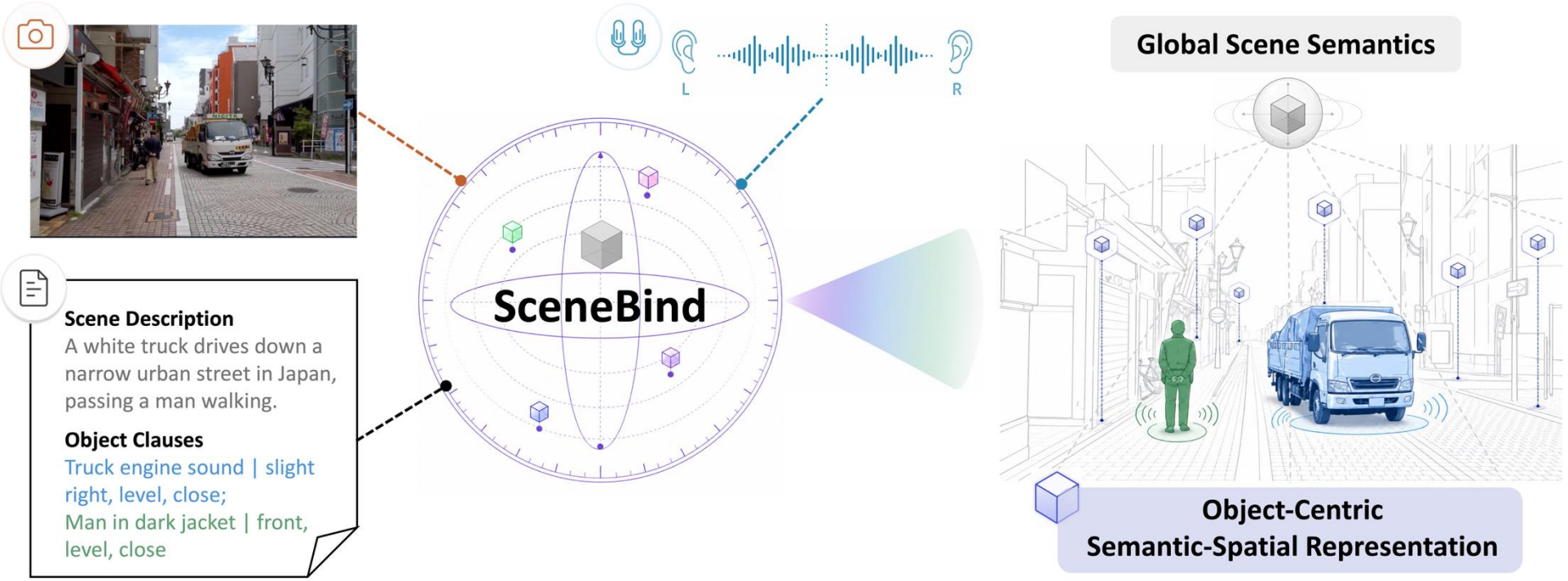
VISUAL SLOTS

V0 river conf 0.956
front-right low 5-10 m

V1 parked cars conf 0.863
front-right level >10 m

SceneBind: Binding What and Where Across Vision, Audio and Language

https://scenebind.github.io/



- ✓ Omni-modal
- ✓ Real Scenes
- ✓ Spatial-Semantic