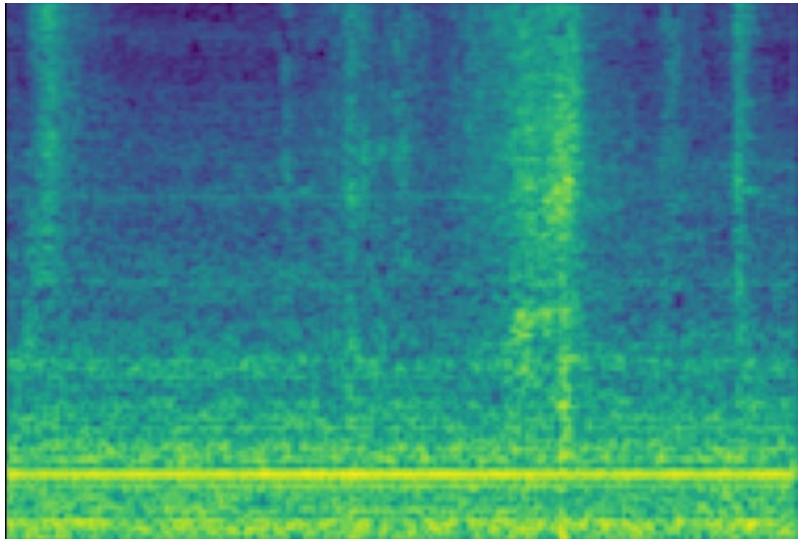
Generating sounds for actions and scenes

Kristen Grauman

University of Texas at Austin

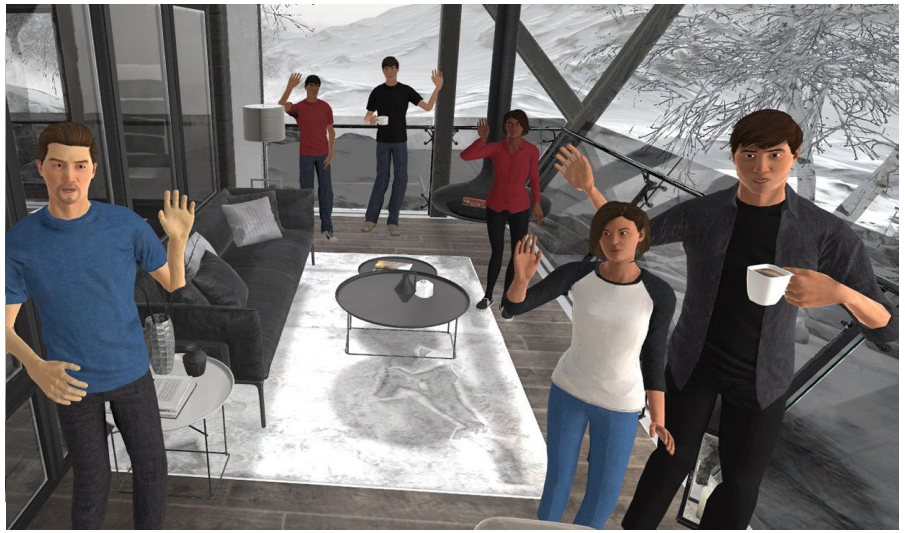
Generating visually consistent sounds

Silent video



Plausible audio

Immersive audio-visual (re-)generation



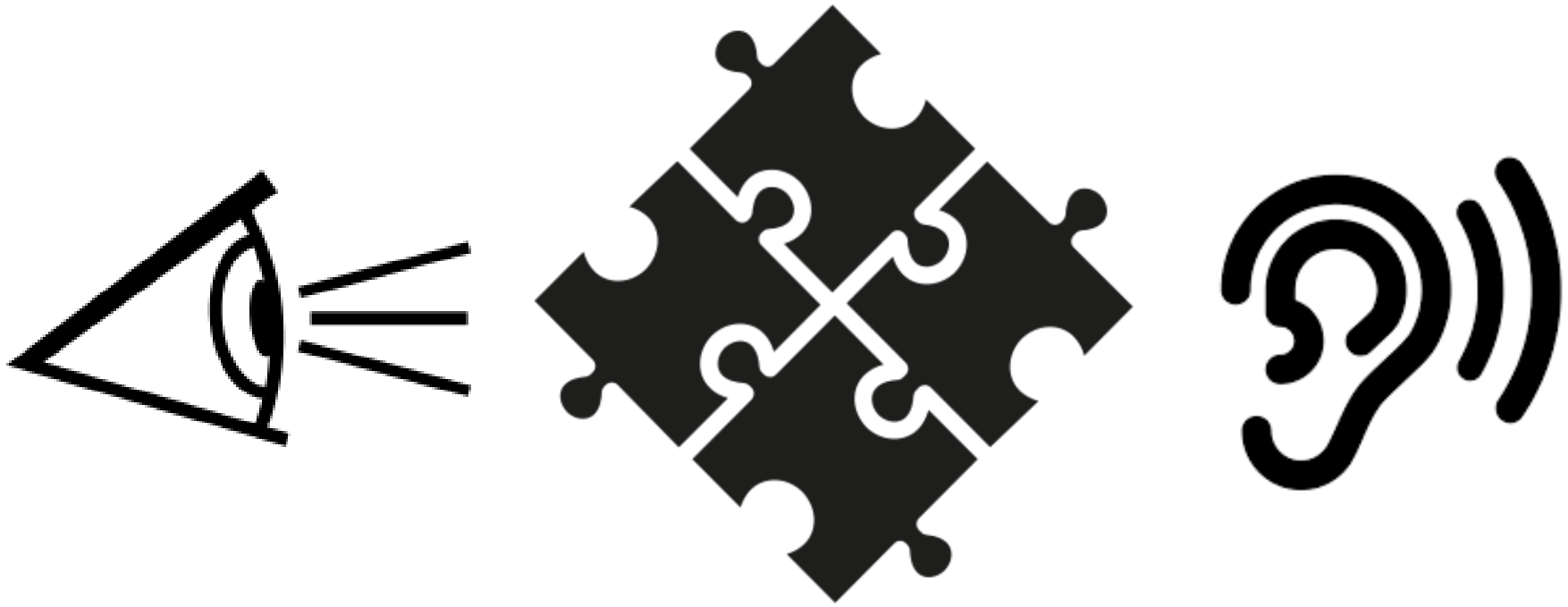
Augmented and virtual reality

Gaming

Film dubbing and Foley

Video conferencing

Immersive audio-visual (re-)generation



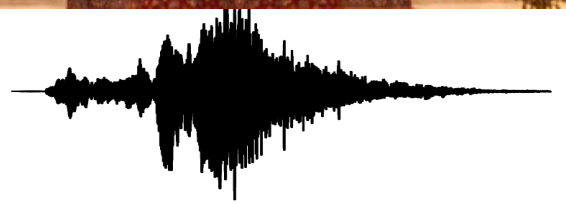
Critical to achieve audio-visual *coherence*

- Easier to understand
- Enhanced realism
- Smooth user experience

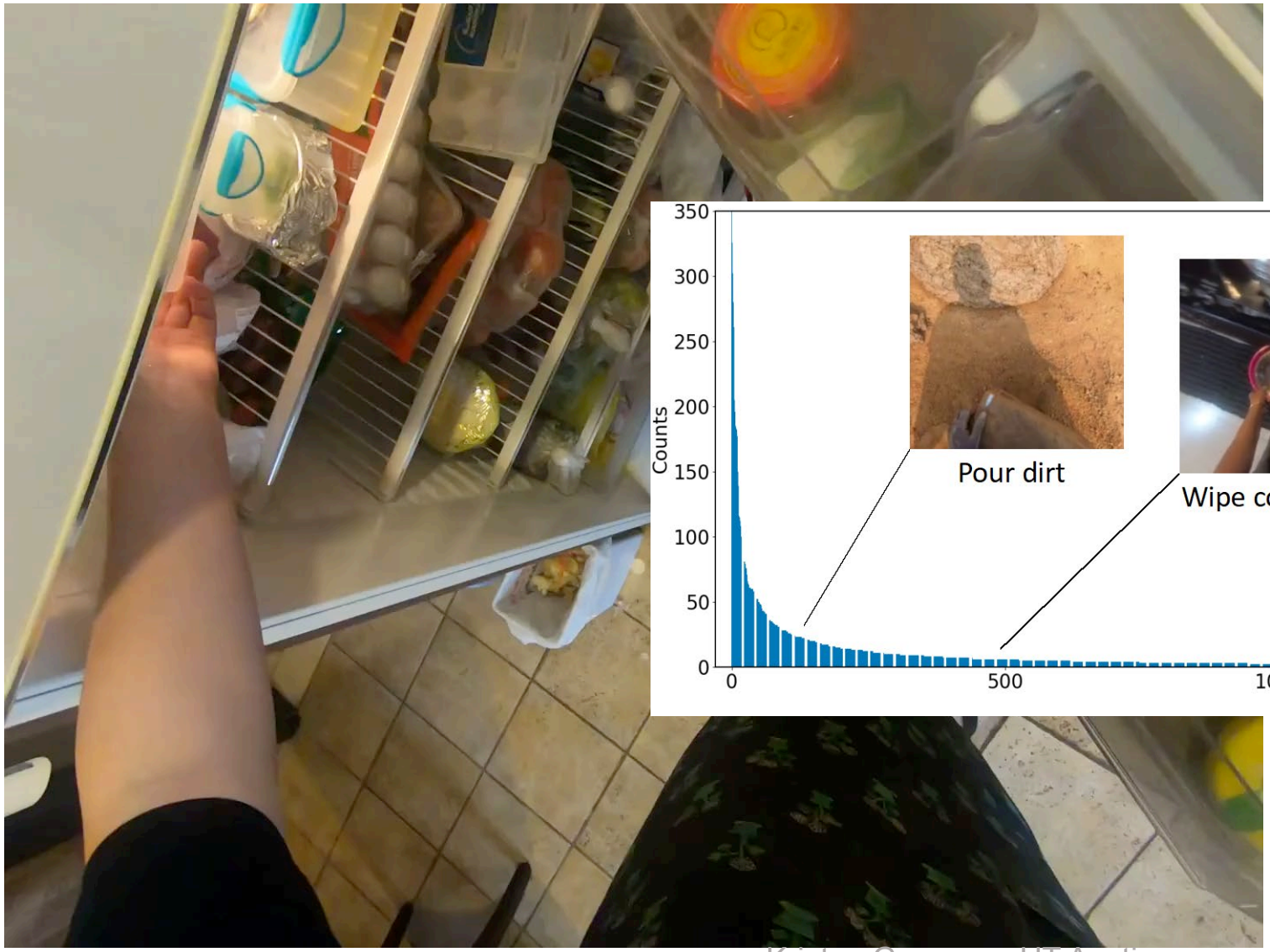
Two key visual associations

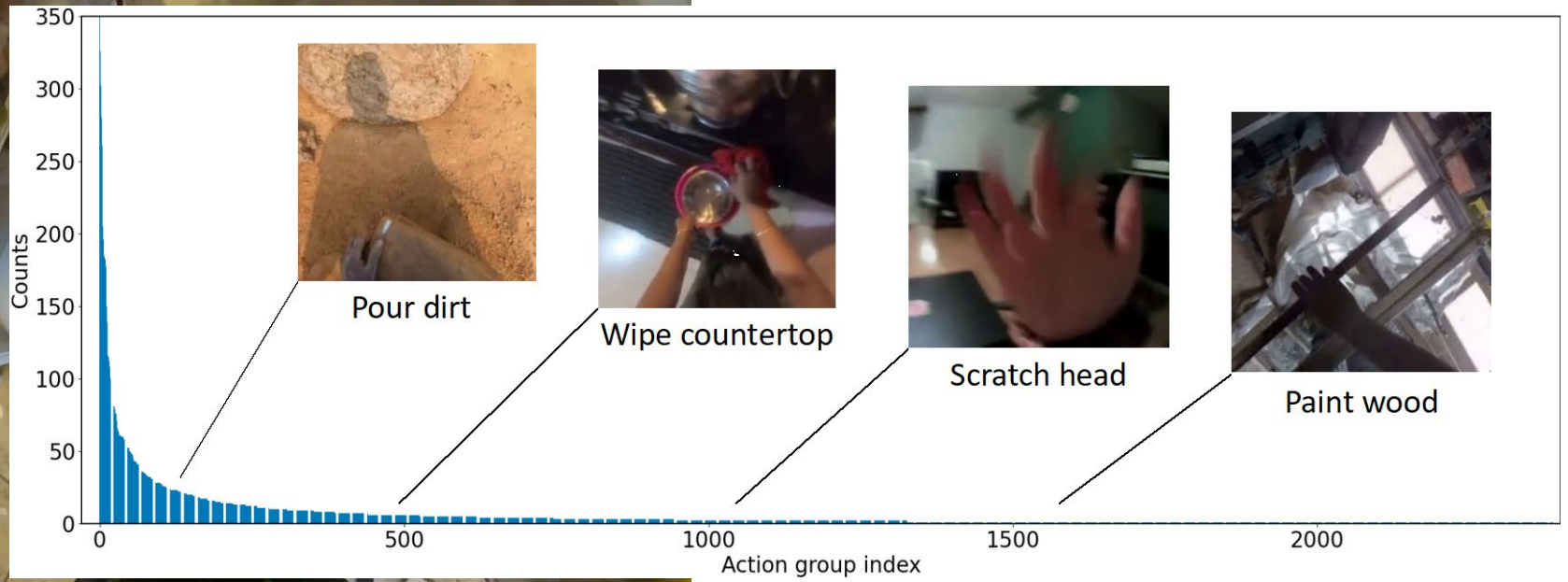
Actions

Scenes



Discovering sounding actions from in-the-wild video





Wide variety of **object interactions** → long tail of sounds

Discovering sounding actions

Goal: determine when sounds are directly caused by human action

Challenge: non-action sounds can be correlated with the visual signal

Changan Chen et al., SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos, CVPR 2024

Temporal segments from narrated video

Ego4D[1]/EPIC-Kitchens[2] have timestamped narrations describing atomic actions

C cuts the grass with sickle tool

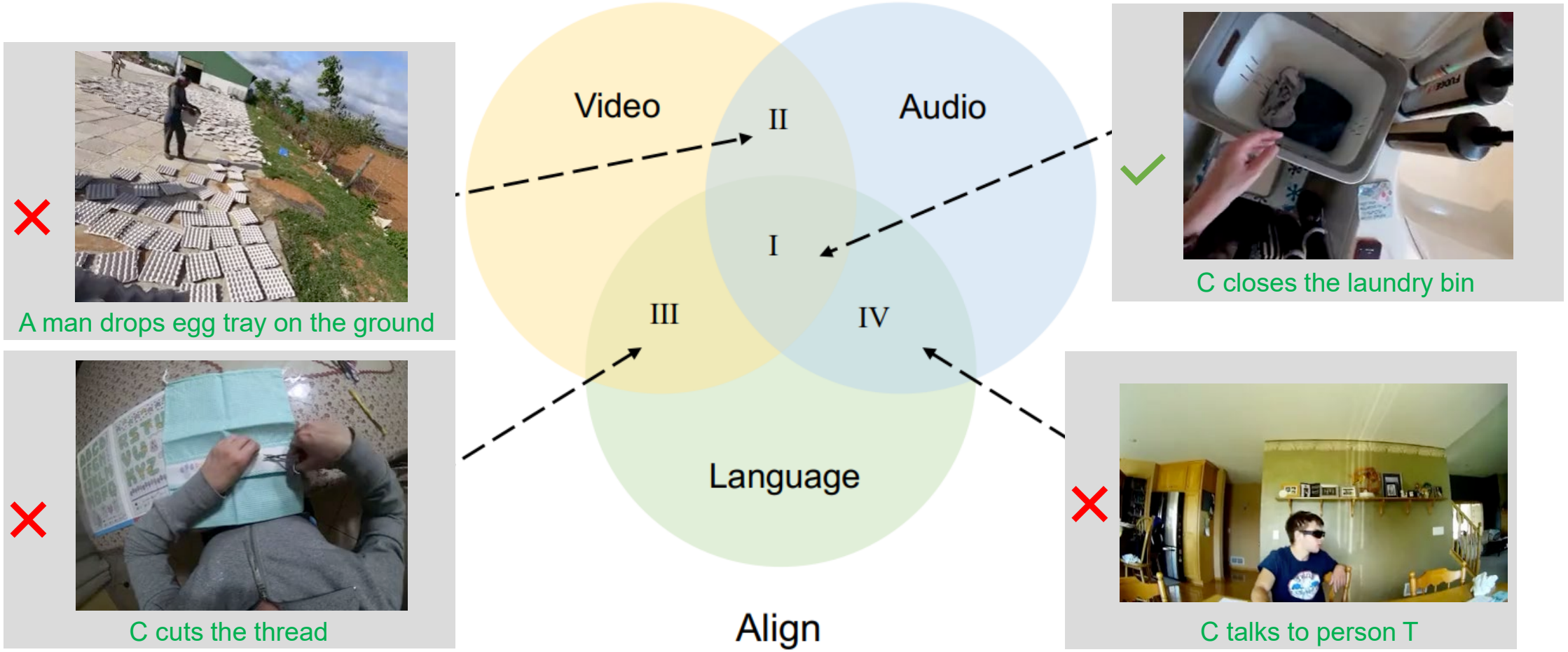
C separates coconut leave stalks

C moves a cutting machine with hands

[1]Ego4D, Grauman et al., CVPR 2022
[2]EPIC-KITCHENS, Damen et al., ECCV 2018

Discovering sounding actions from narrated video

Seek semantic agreement between video, audio and language

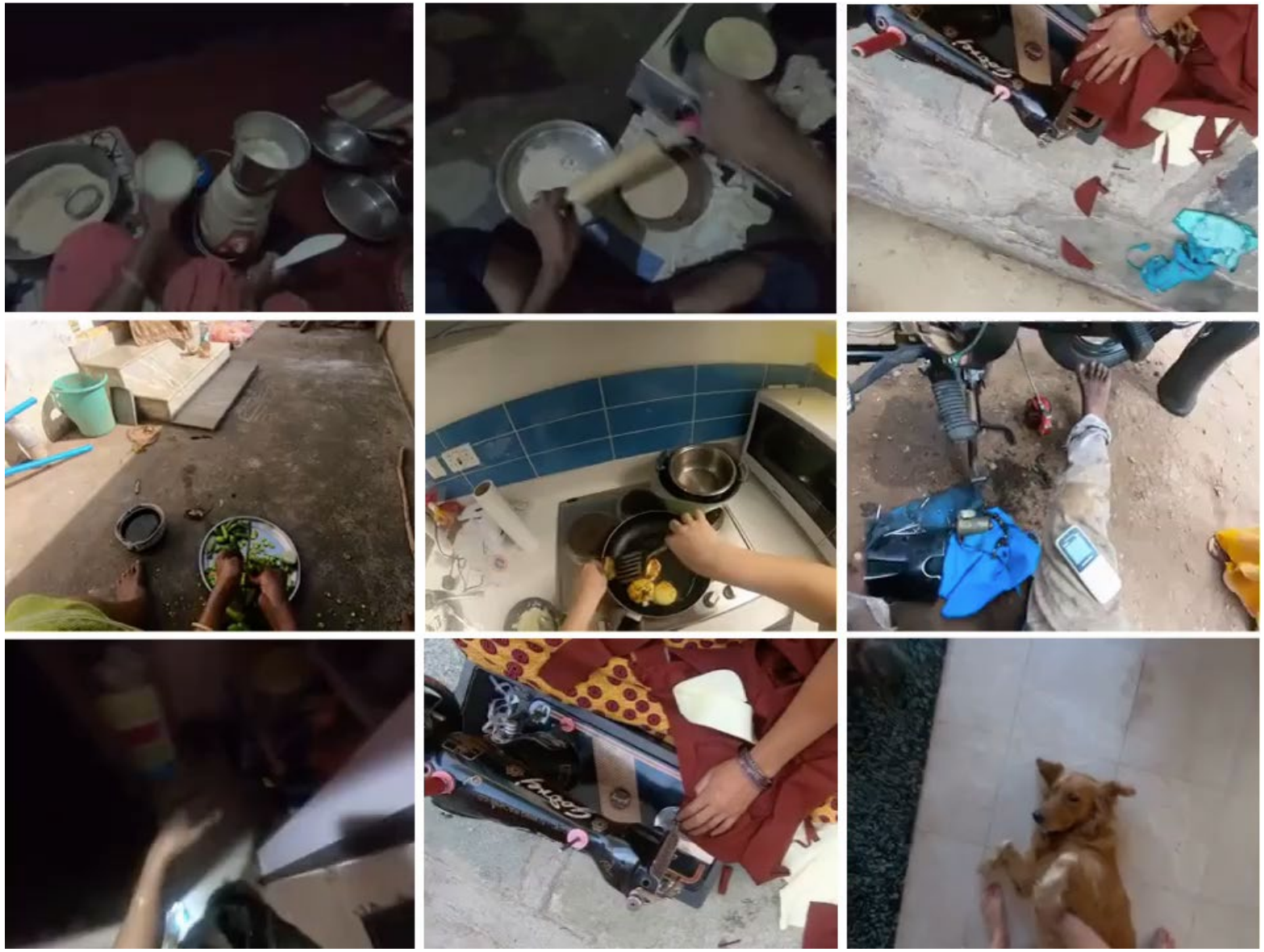


Novel “multimodal contrastive consensus” loss

Top-ranked sounding actions

Changan Chen et al., SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos, CVPR 2024

Bottom-ranked sounding actions

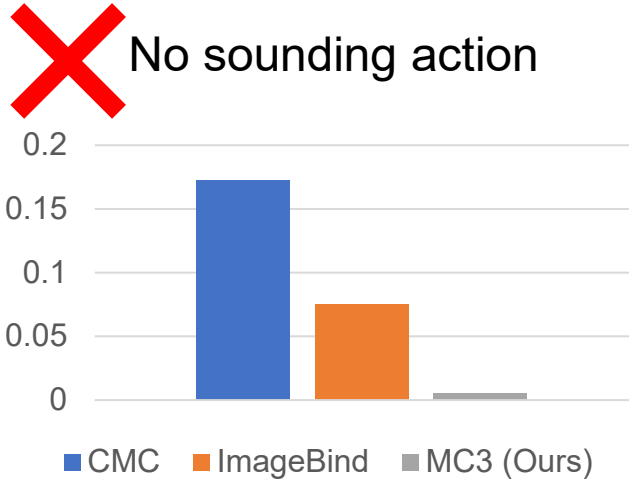


Changan Chen et al., SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos, CVPR 2024

Is this a sounding action?

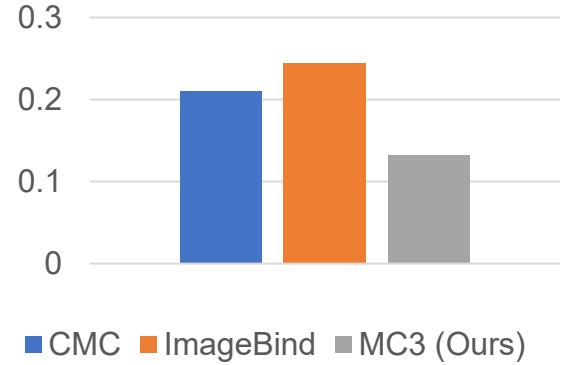
(Reference only) C drops the coconut on the tray

No sounding action



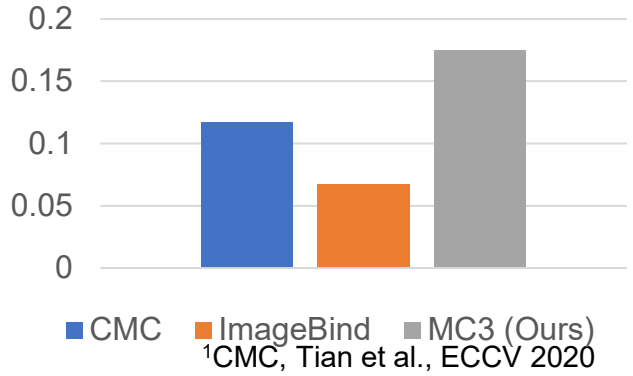
(Reference only) C lifts the left hand

No sounding action



(Reference only) C rinses her hand with water

Sounding action



Changan Chen et al., SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos, CVPR 2024

[1]CMC, Tian et al., ECCV 2020

[2]ImageBind: Girdhar et al., CVPR 2023

Sounding action discovery accuracy

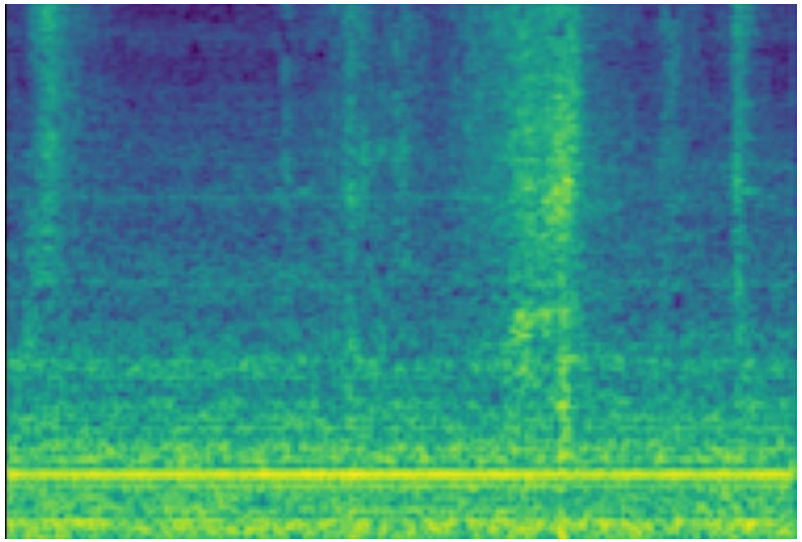
Results on Ego4D

				AV		AL		
	🔊	🎥	📄	ROC	PR	ROC	PR	
Random	✗	✗	✗	0.500	0.559	0.500	0.559	
CLAP [1]	✓	✗	✓	-	-	0.637	0.695	bi-modal
CM-ACC [2]	✓	✓	✗	0.540	0.590	-	-	bi-modal
CMC [3]	✓	✓	✓	0.550	0.601	0.635	0.693	trimodal joint
ImageBind [4]	✓	✓	✓	0.554	0.605	0.642	0.685	trimodal sequential
w/o $\mathcal{L}_{\text{consensus}}$	✓	✓	✓	0.563	0.615	0.635	0.694	
w/o $\mathcal{L}_{\text{contrastive}}$	✓	✓	✓	0.436	0.493	0.584	0.620	
w/o align-stage	✓	✓	✓	0.448	0.507	0.464	0.521	
MC3	✓	✓	✓	**0.598**	**0.666**	**0.658**	**0.715**	

[1]CLAP, Wu et al., ICASSP 2023
[2]CM-ACC: Ma et al., ICLR 2021
[3]CMC, Tian et al., ECCV 2020
[4]ImageBind: Girdhar et al., CVPR 2023

- Outperforming existing bi-modal or trimodal losses
- Both contrastive and consensus losses are important for learning

How to *generate* action sounds?



Given a silent video, can we generate plausible action sounds?

Challenge

In-the-wild videos have coupled ambient/background and action sounds

Offscreen machine sound

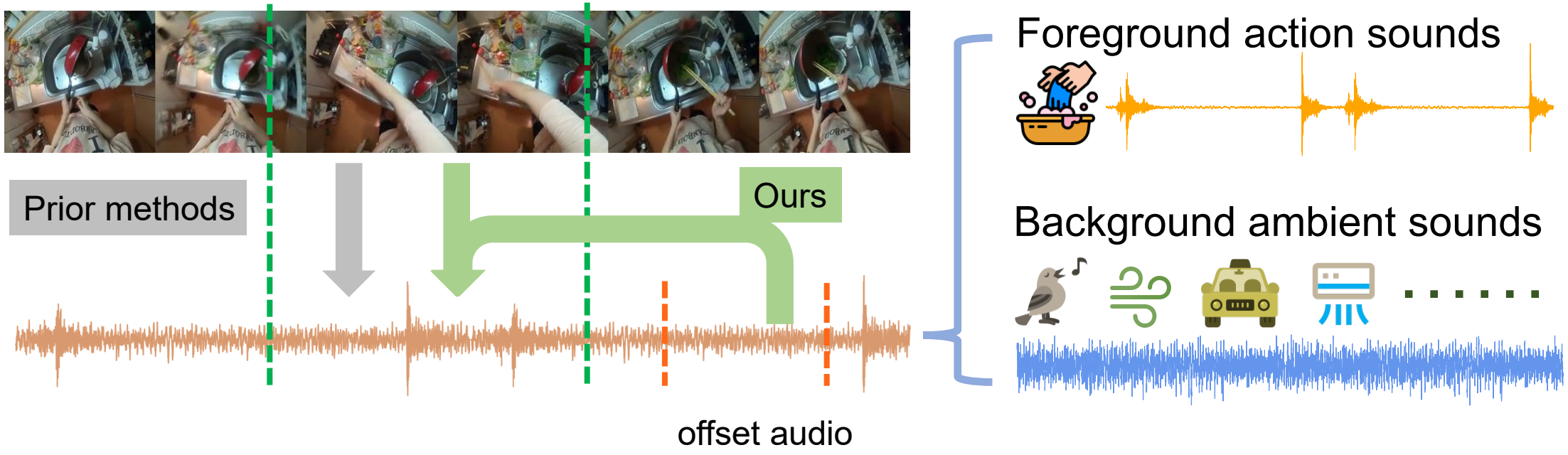
Birds chirping in the background

Low frequency buzzing sound

High risk of hallucination if we ignore this overlap in training,
yet we have no ground truth for ambient/action separated sounds.

Action2sound: ambient-aware sound generation

Insight: Ambient sound tends to stay similar throughout a long video

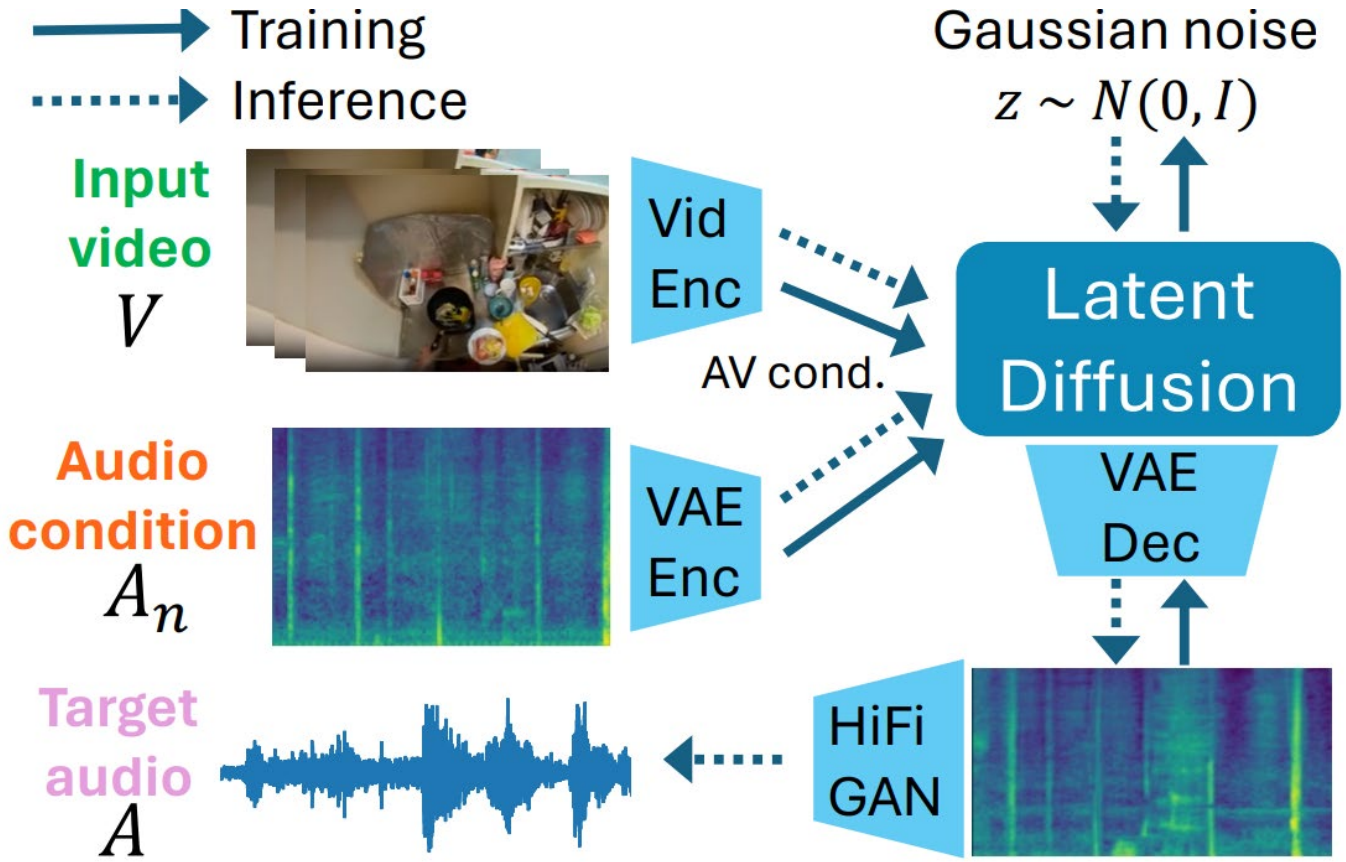


Train: Conditioning on temporally nearby audio encourages the model to focus on action sound generation, while learning to ignore the ambient sound.

Chen and Peng et al., Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos, ECCV 2024

Action2sound: ambient-aware sound generation

Audio-visual latent diffusion model



Train: Conditioning on temporally nearby audio encourages the model to focus on action sound generation, while learning to ignore the ambient sound.

HiFi-GAN, Kong et al. 2020

Chen and Peng et al., Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos, ECCV 2024

Action2sound: ambient-aware sound generation

Ours action-ambient

Ours action-focused

Diff-Foley

Original video

[1]Diff-Foley, Luo et al., NeurIPS 2023

Chen and Peng et al., Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos, ECCV 2024

Action2sound: ambient-aware sound generation

Ours action-ambient

Ours action-focused

Diff-Foley

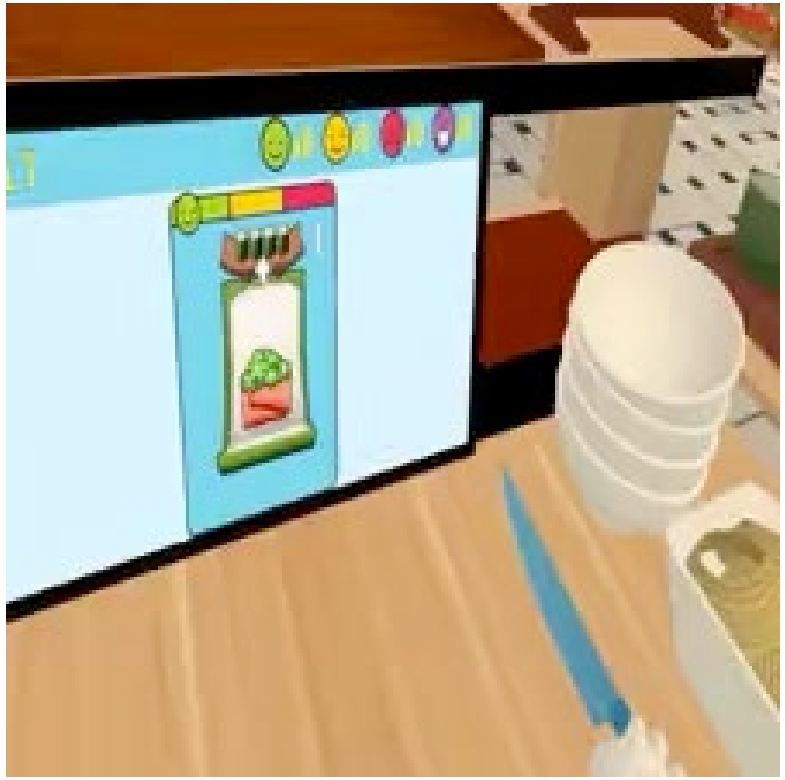
Original video

[1]Diff-Foley, Luo et al., NeurIPS 2023

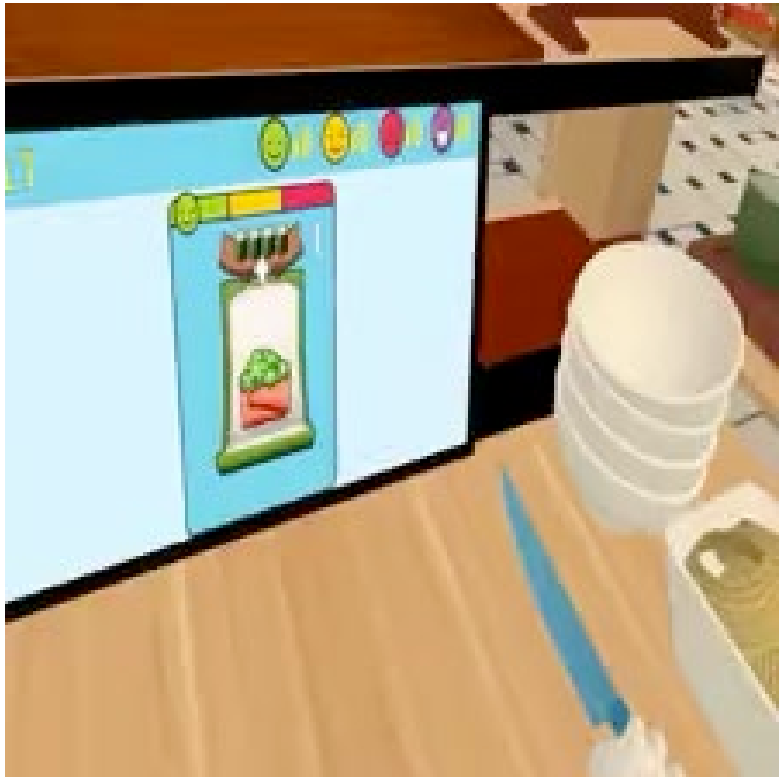
Chen and Peng et al., Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos, ECCV 2024

Action2sound: ambient-aware sound generation

Train on real video → test on games



Game clip



Ours (generated from visual frames alone)

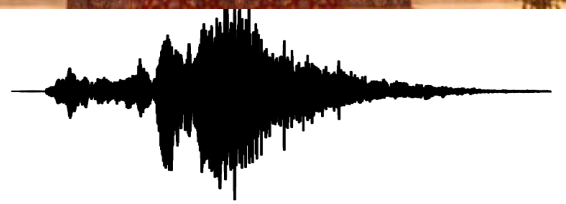
Video source: Clash Of Chefs

Chen and Peng et al., Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos, ECCV 2024

Two key visual associations

Actions

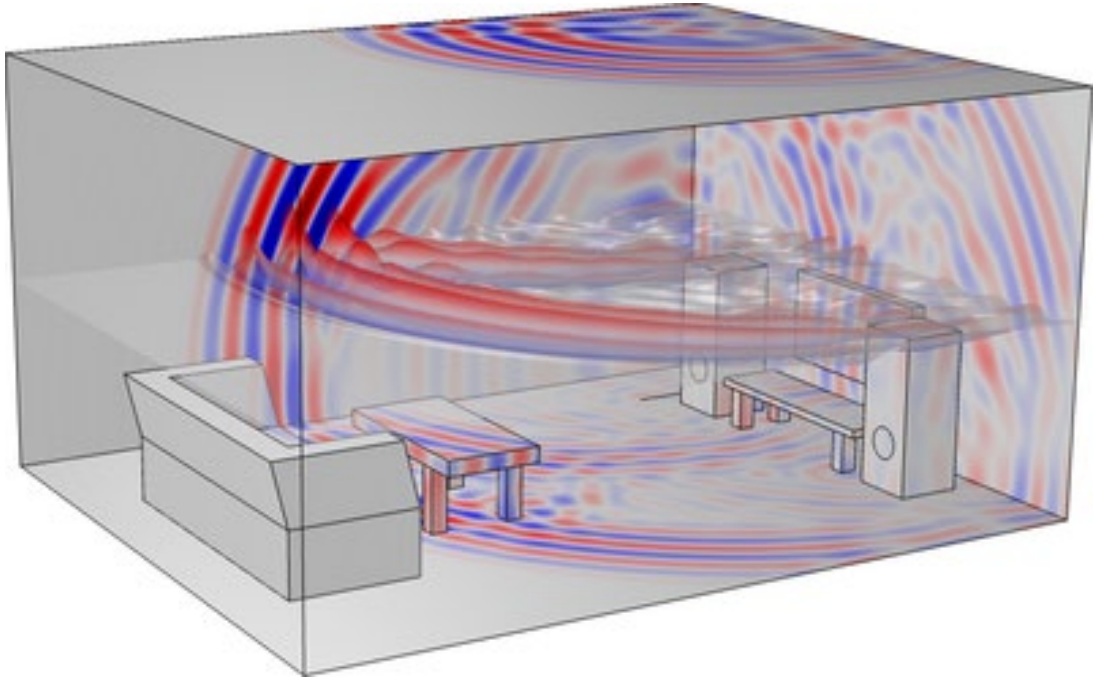
Scenes

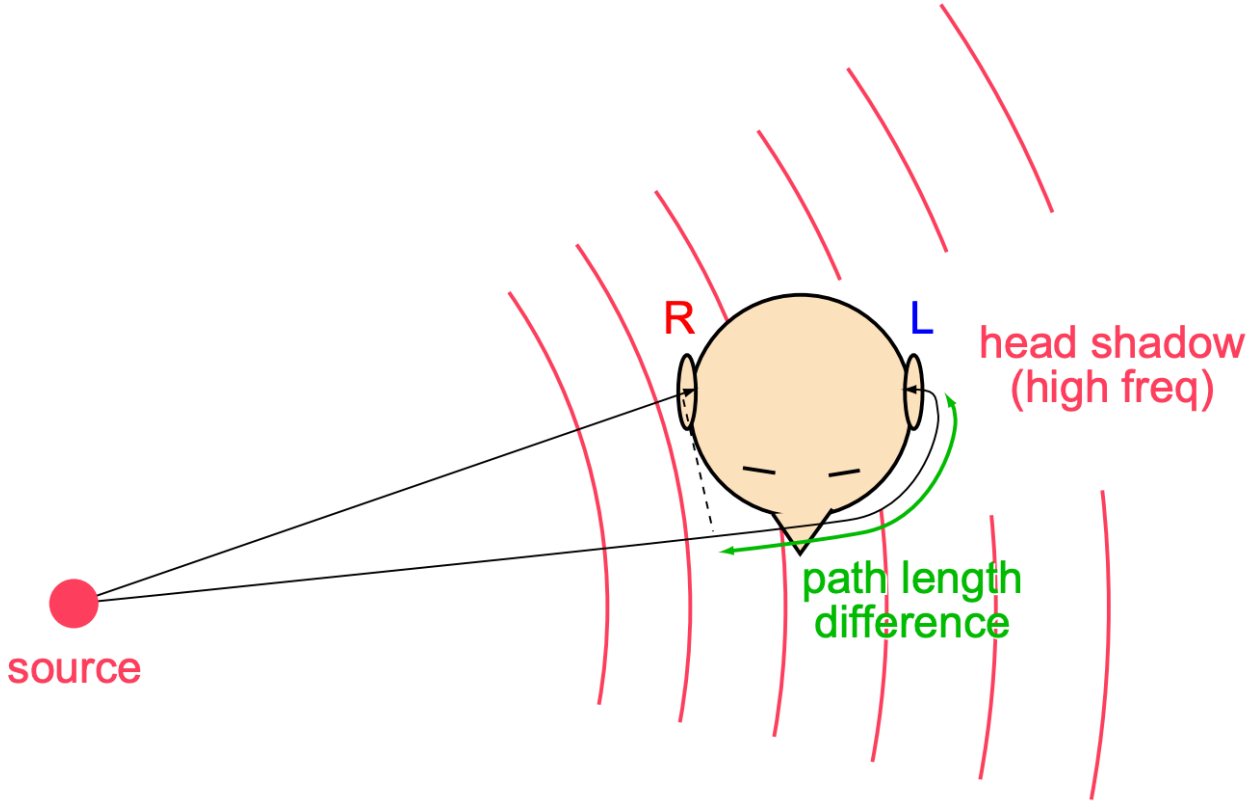


1 drum kit, 5 different spaces

Source: Shred Shed Studio

Spatial effects in audio





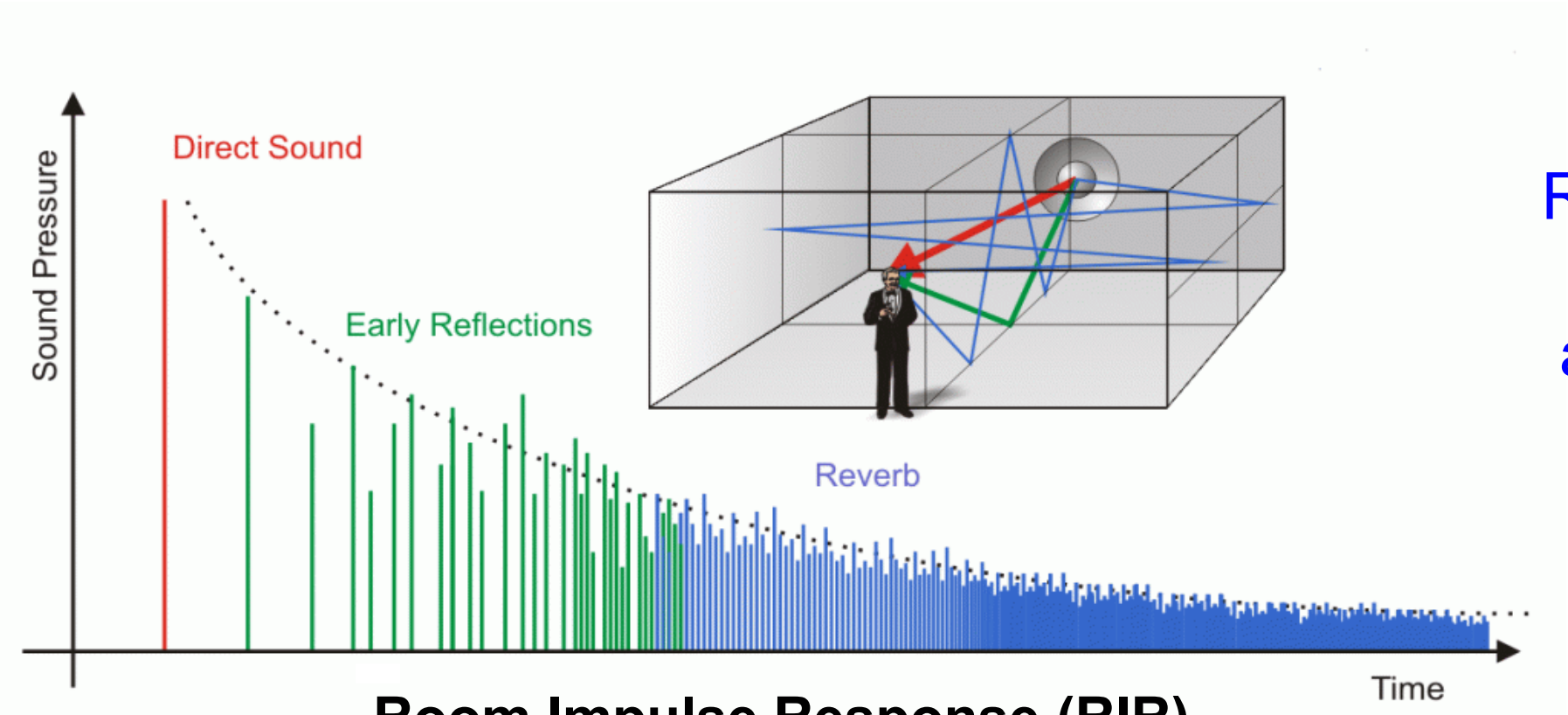
Factors from 3D environment:

- Geometry of the space
- Materials in the room
- Position of source and receiver

Agent's spatial hearing cues:

- Interaural time difference (ITD),
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)

Modelling scene acoustics via room impulse responses



Room Impulse Response (RIR)

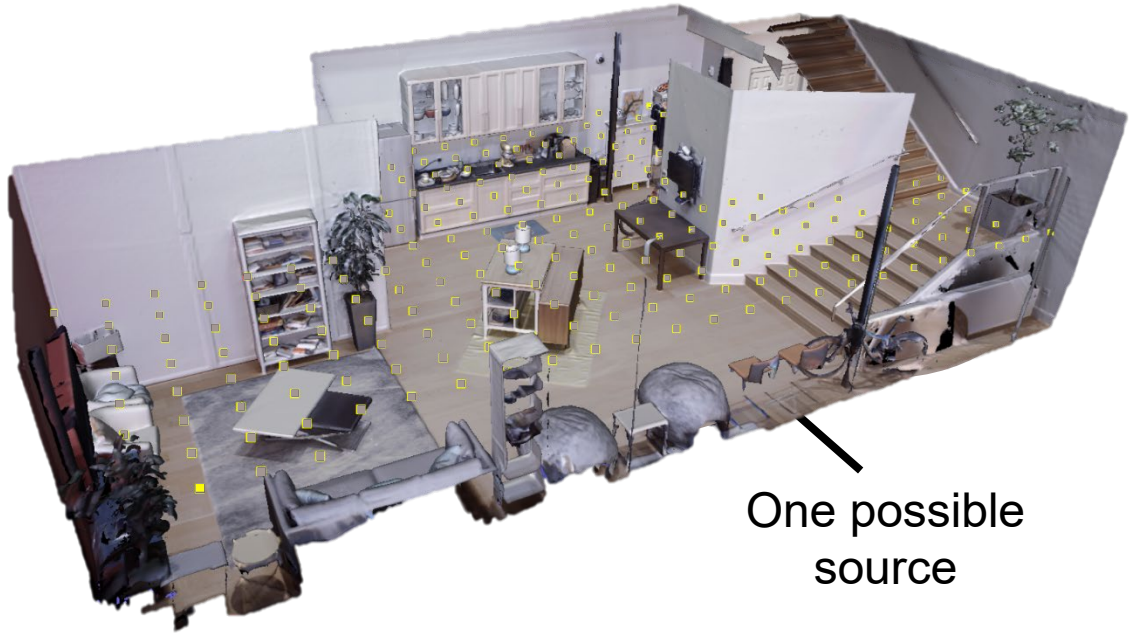
RIR is "ticket" to acoustically accurate audio generation!

Image source: *Creatac Beat Kaufmann*

SoundSpaces audio simulation platform

https://github.com/facebookresearch/sound-spaces

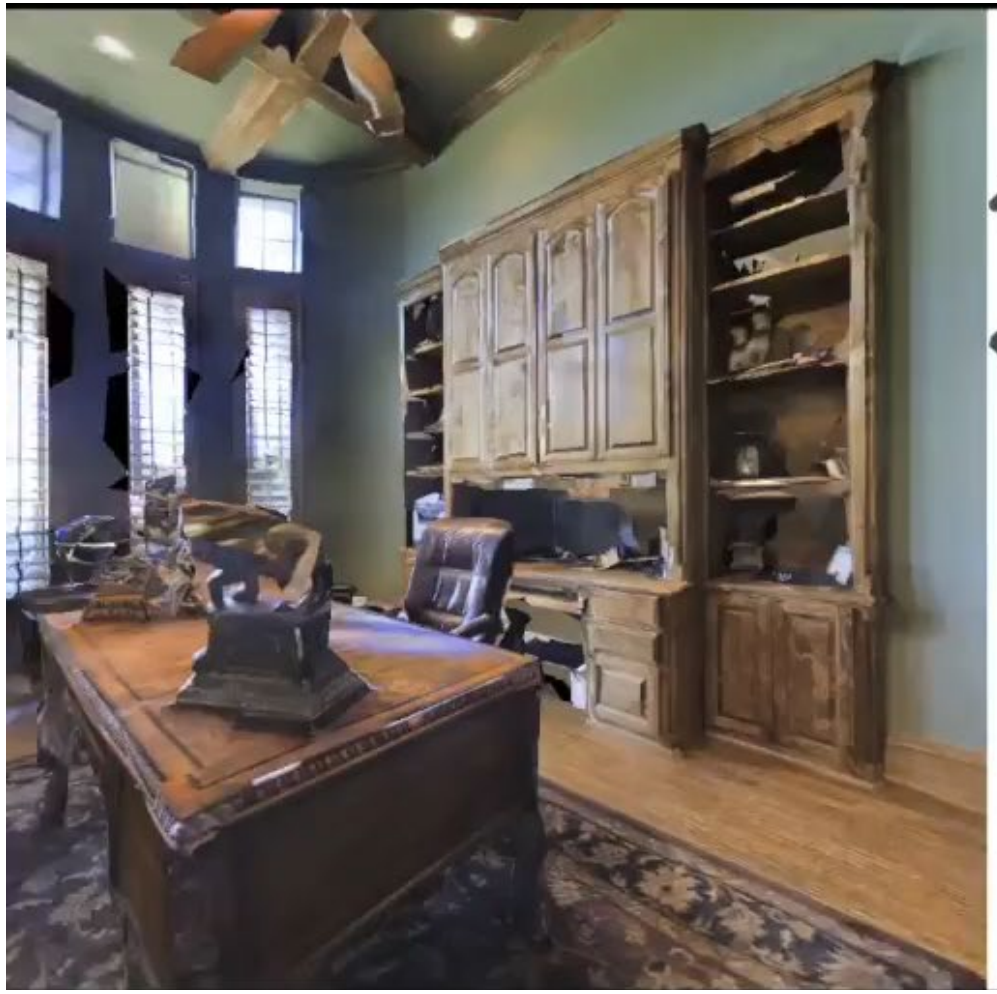
- Visually realistic real-world 3D environments (Matterport3D, Replica, Gibson, HM3D, your choice…)
- Acoustically realistic (geometry, materials, source location) binaural sound in real-time, for waveform of your choice
- Room impulse response (RIR) for any source *x* receiver location, on-the-fly
- Habitat-compatible



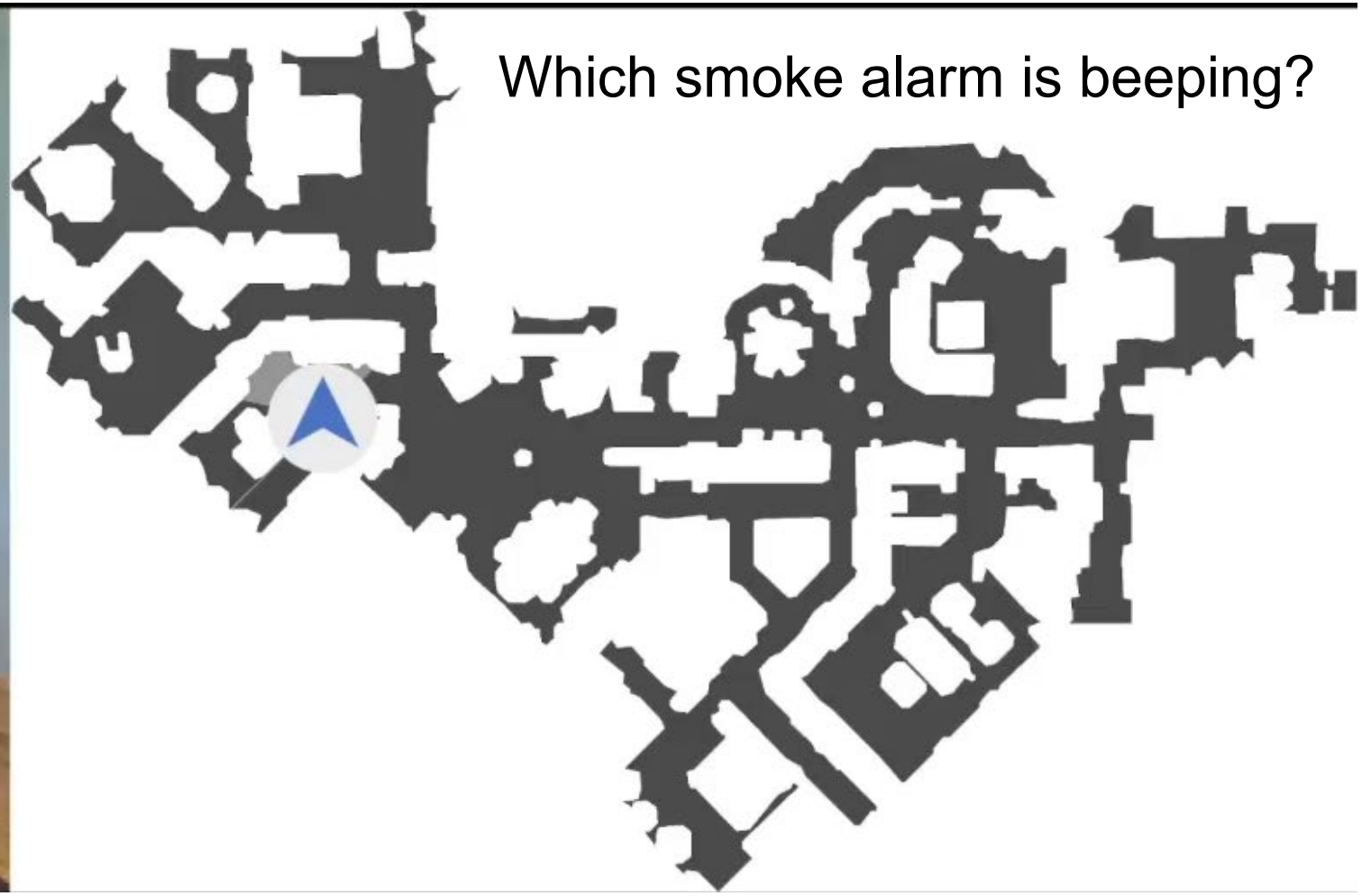
C. Chen*, U. Jain*, et al., SoundSpaces, ECCV 2020
C. Chen et al. SoundSpaces 2.0 NeurIPS 2022

SoundSpaces audio simulation

C. Chen*, U. Jain*, et al., SoundSpaces, ECCV 2020 & SoundSpaces 2.0, NeurIPS 2022



Agent view



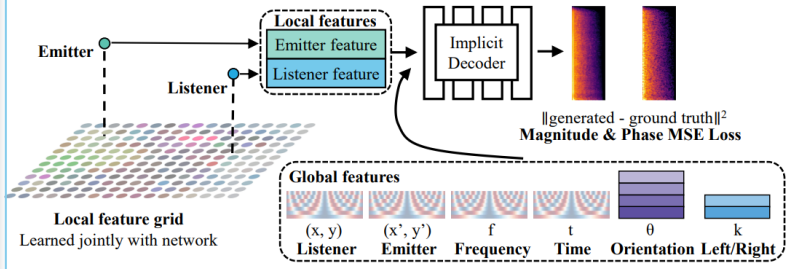
Top-down map (unknown to the agent)

Listen with headphones online for spatial sound experience http://vision.cs.utexas.edu/projects/audio_visual_navigation/

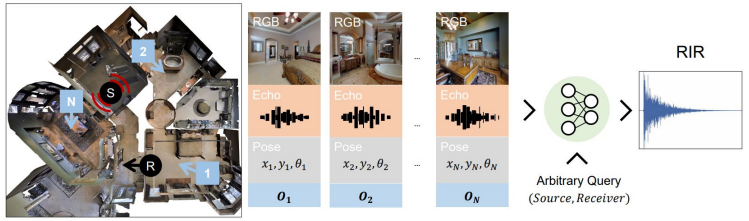
How to build an acoustic model for a new environment?

Prior work: building a scene acoustics model

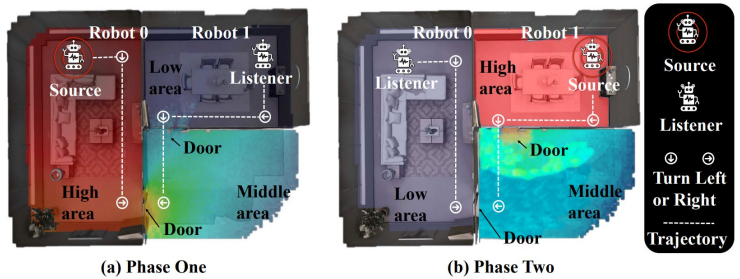
Neural Acoustic Fields (NAF)[1]
Densely samples space to build Neural Field acoustic representation



Few-ShotRIR[2]
Builds acoustic context from fixed number of randomly selected samples



Measuring Acoustics with Collaborative Multiple Agents[3]
Two agents navigate a space and build acoustic map using local reward

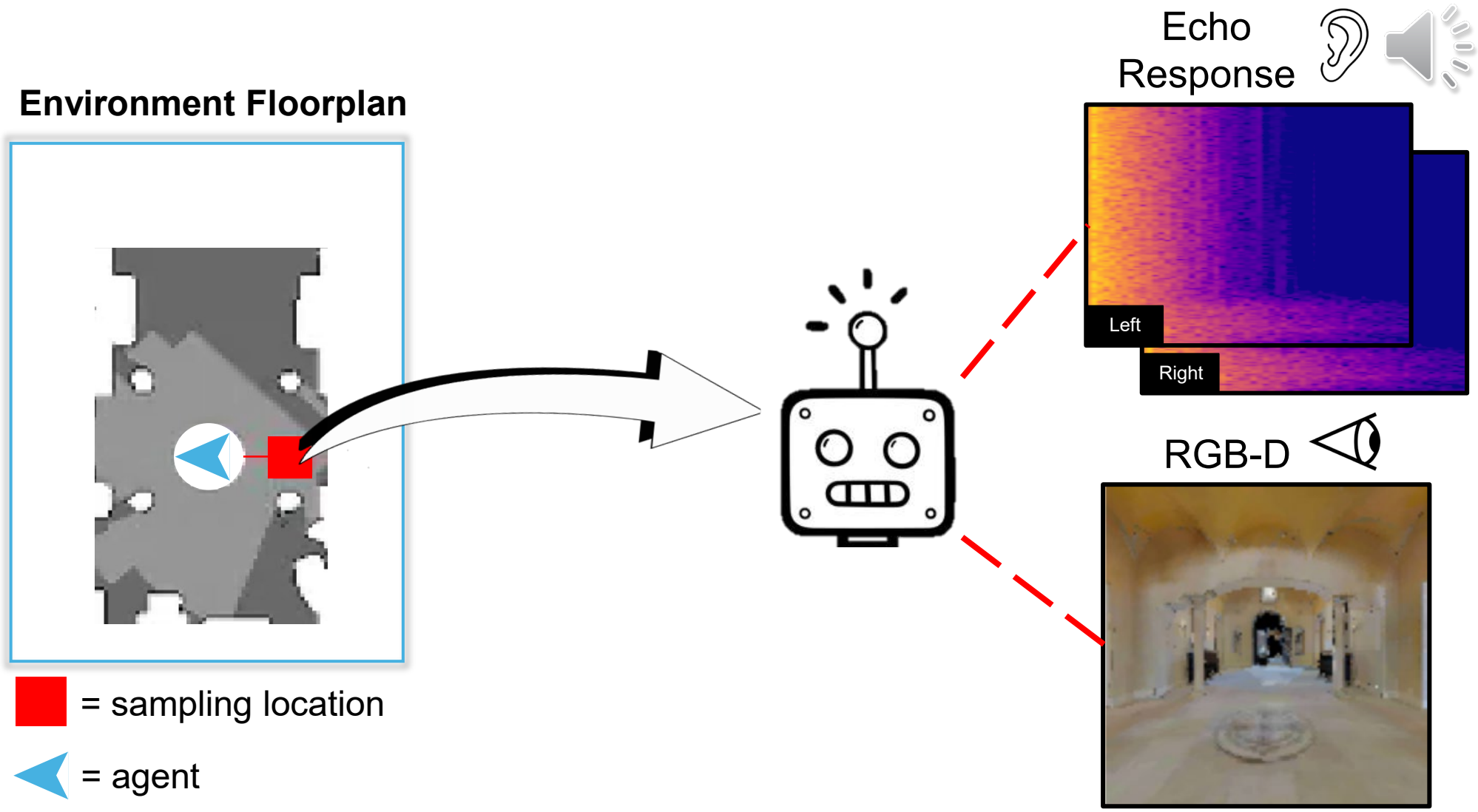


Require knowledge of room geometry – **often unrealistic** ✗

Dense sampling in the space – **expensive and time-consuming** ✗

[1] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan, "Learning Neural Acoustic Fields," NeurIPS 2022
[2] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman, "Few-Shot Audio-Visual Learning of Environment Acoustics," NeurIPS, 2022
[3] Y. Yu, C. Chen, L. Cao, F. Yang, and F. Sun, "Measuring Acoustics with Collaborative Multiple Agents," IJCAI, 2023

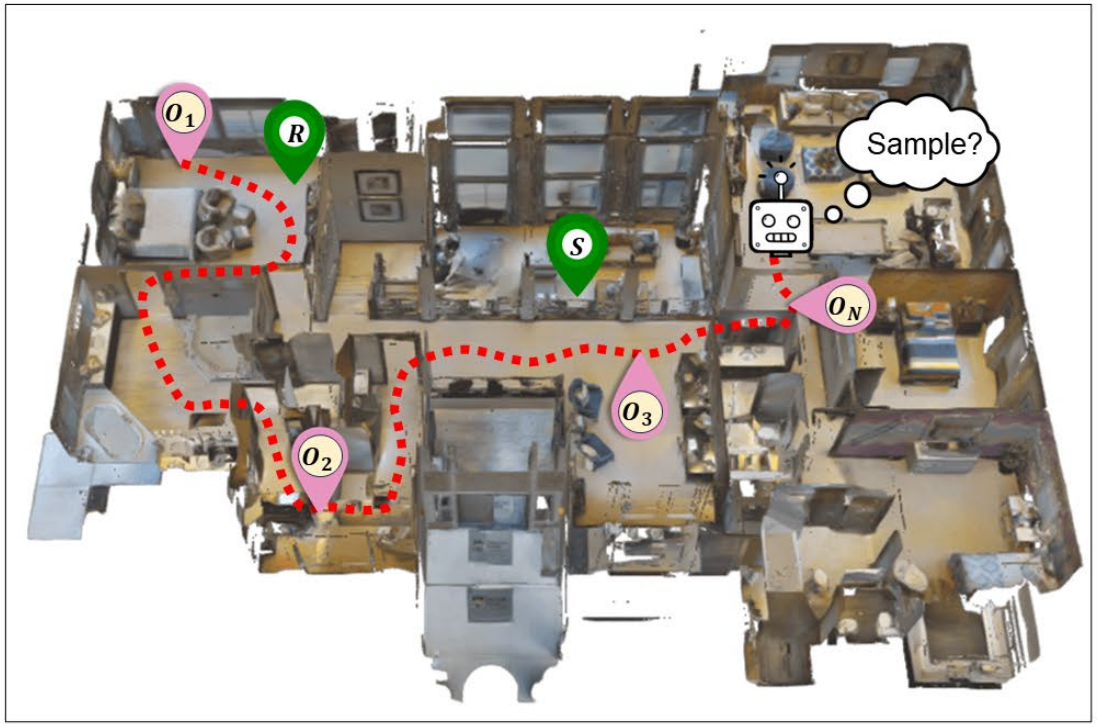
“Sampling” an audio-visual observation

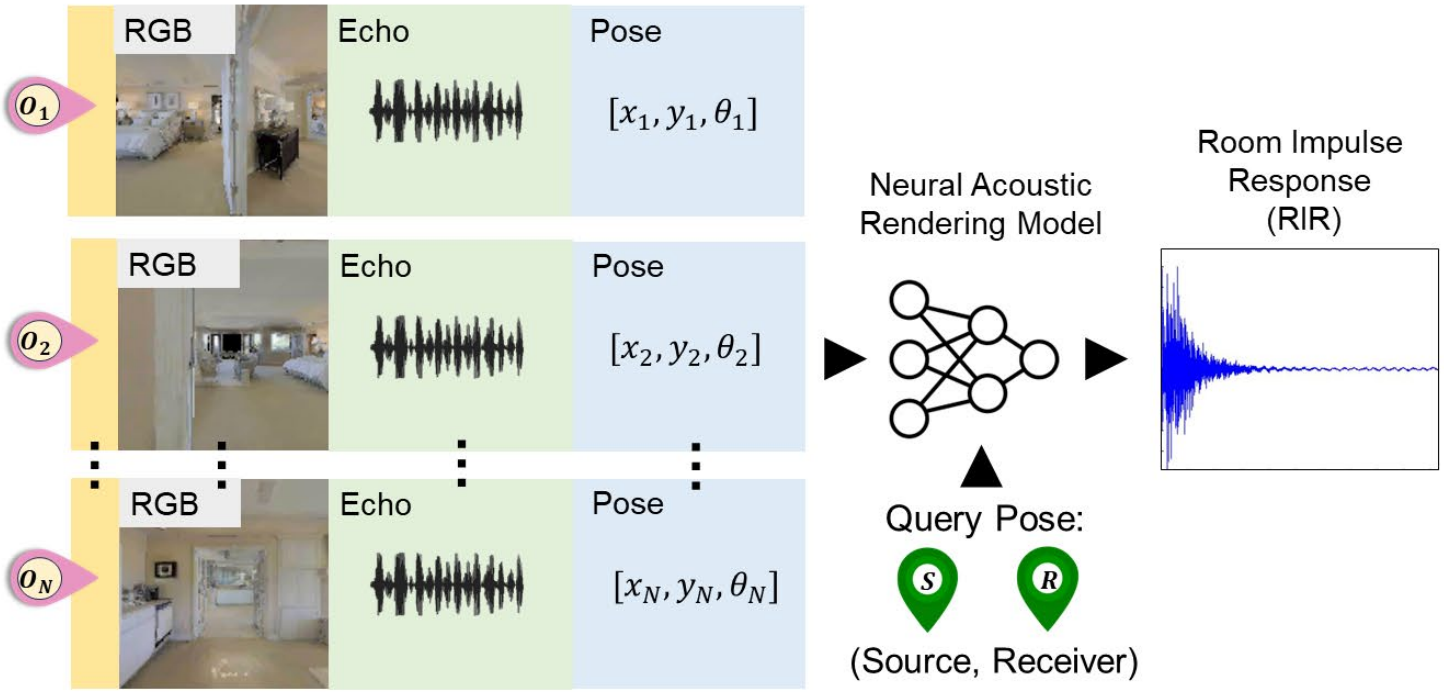


Somayazulu et al. IROS 2024

Goal: Active acoustic sampling

Agent must **actively select** audio-visual observations $\{o_i\}$ as it navigates an **unmapped** scene, constructing an acoustic model of the environment.

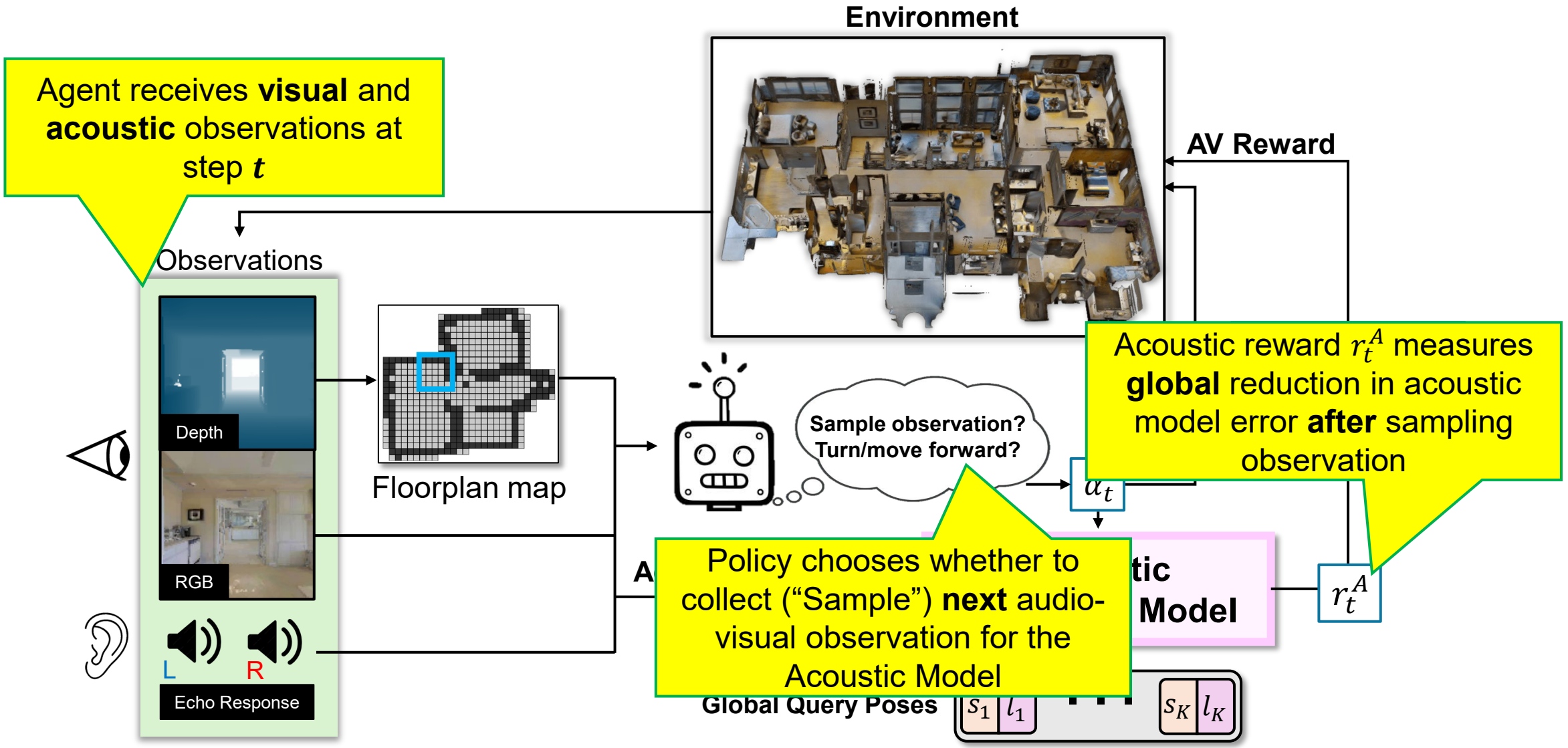




Agent has **no prior knowledge of the scene's floorplan or geometry**, and must complete exploration within a budget of T timesteps and N samples

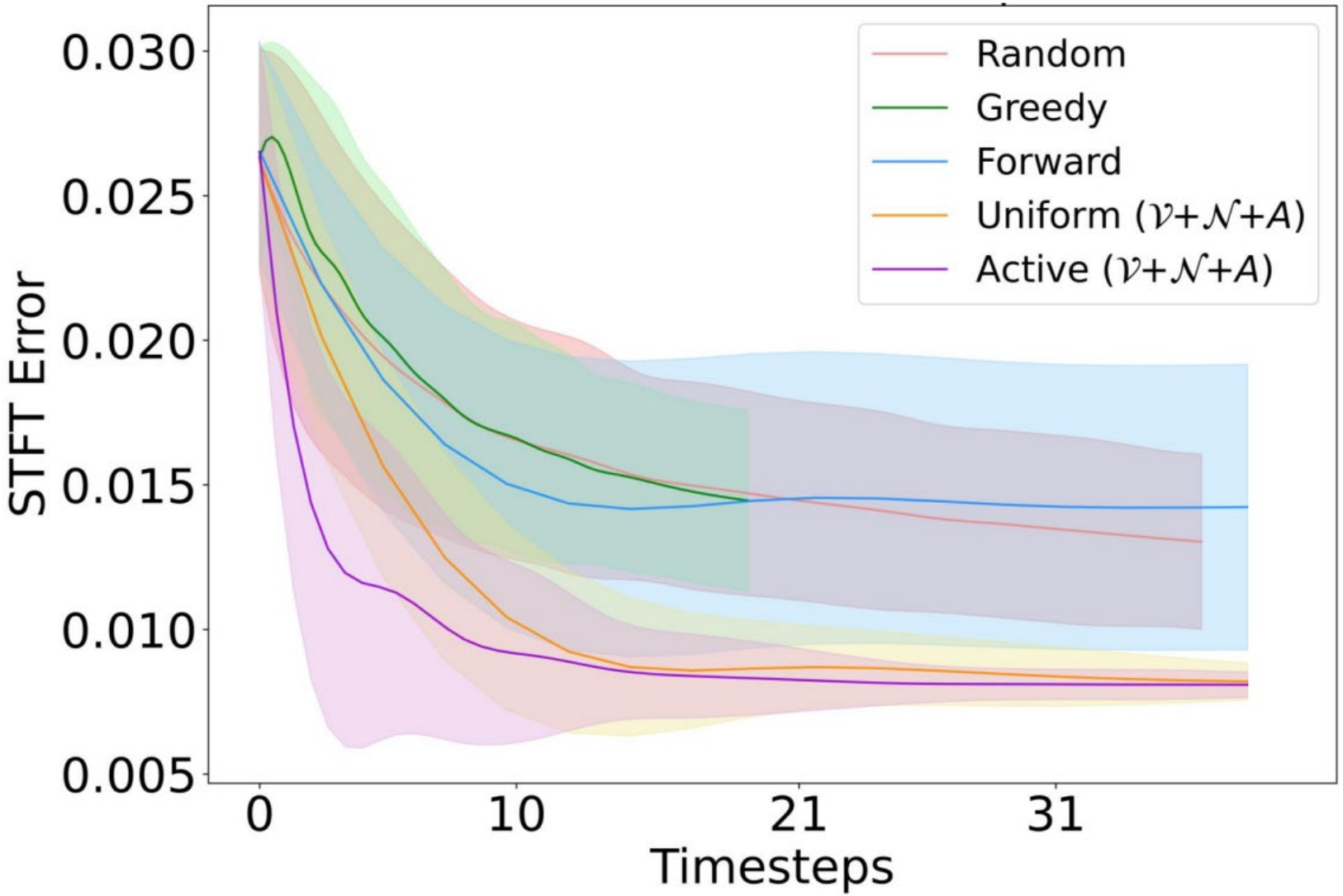
Somayazulu et al. IROS 2024

ActiveRIR: an RL approach



Somayazulu et al. IROS 2024

ActiveRIR results



ActiveRIR (purple) rapidly minimizes acoustic model error using fewer steps

Somayazulu et al. IROS 2024

ActiveRIR acoustic exploration

Passive sampling agent

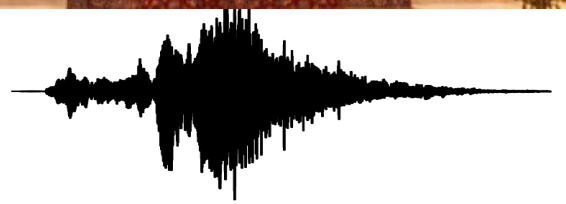
Agent “wastes” context on **spatially redundant observations** ✕

Agent explores only a **small region of scene** ✕

Two key visual associations

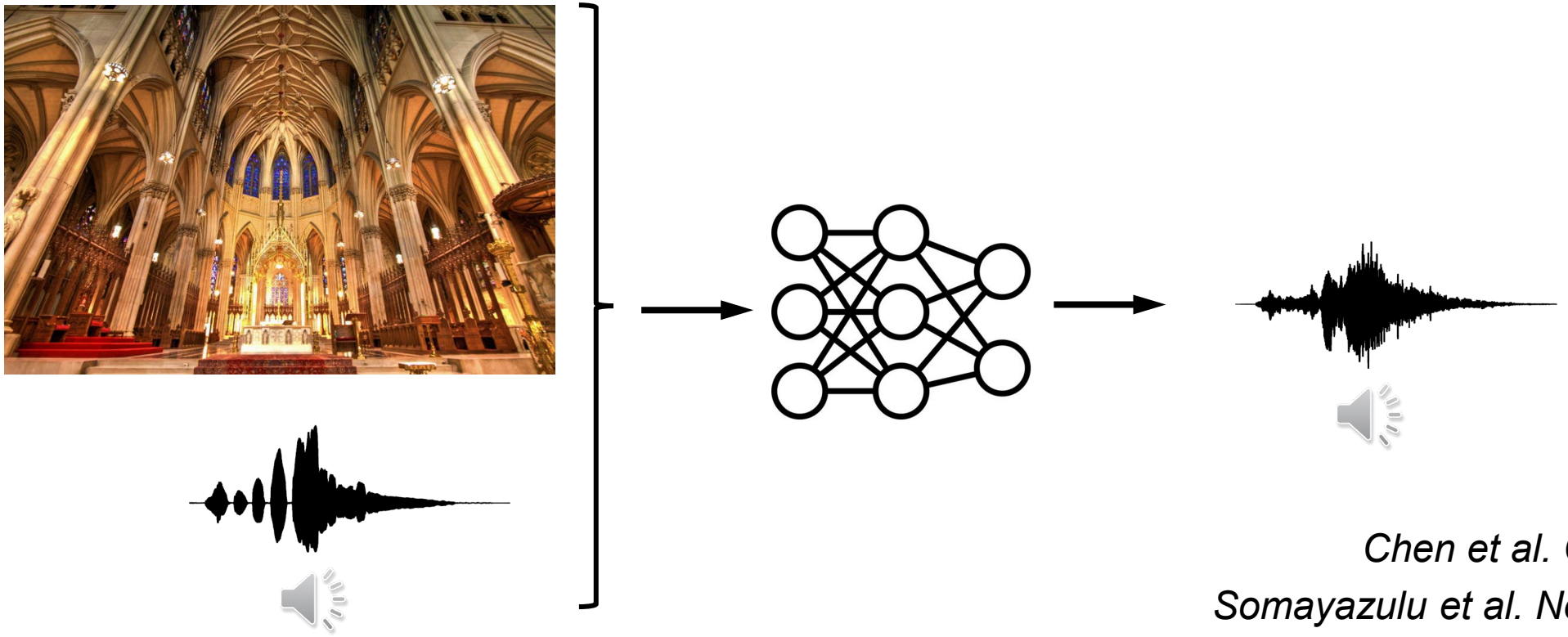
Actions

Scenes



Our idea: Visual acoustic matching

Goal: Transform the sound recorded in one space to sound like it was instead recorded in a target visual scene.

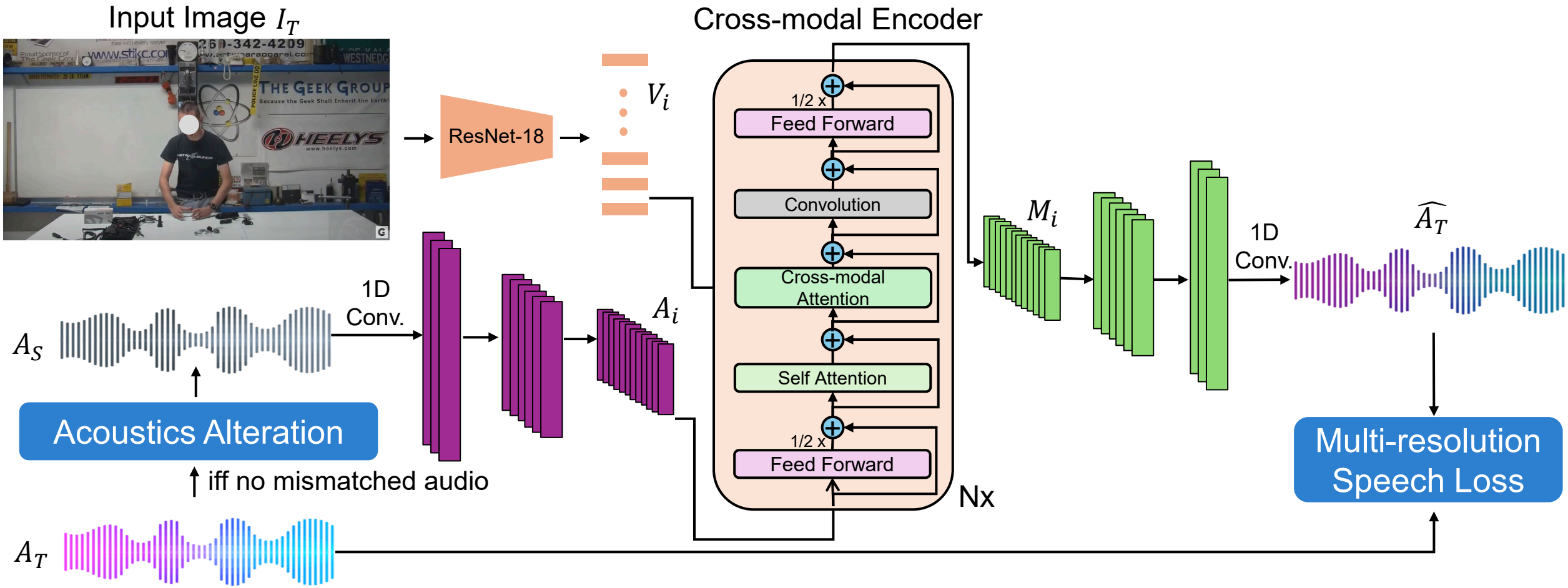


Chen et al. CVPR 2022

Somayazulu et al. NeurIPS 2023

Prior work in acoustic matching – audio-based and typically require reference audio in target environment *[Steinmetz et al. 2021; Valimaki et al. 2016; Su et al. 2020; Koo et al. 2021; Sarroff et al. 2020]*

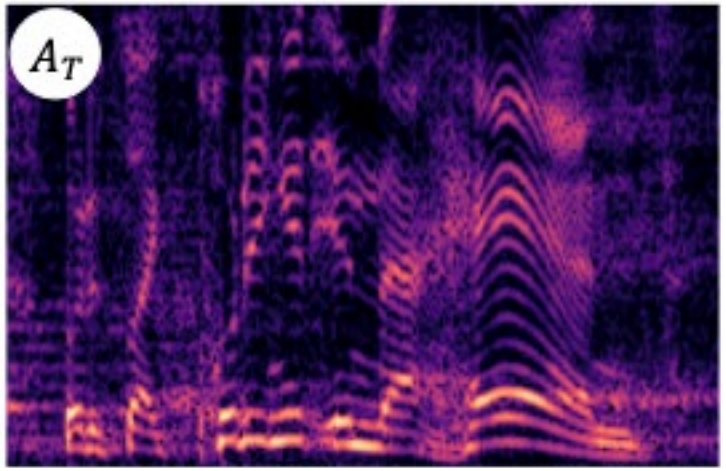
Audio-Visual Transformer for Audio Generation (AViTAR)



Chen et al. CVPR 2022

Acoustic alteration for self-supervised training

Audio in target environment (GT)



“Source” audio preserves content but has mismatched acoustics

Allows training directly on Web videos!

Chen et al. CVPR 2022

Somayazulu et al. NeurIPS 2023

Visual acoustic matching results

	Office	Garage	Auditorium
Input			
AViTAR			
Reverb time	0.34s	0.40s	0.58s

AViTAR reasons about the image content and learns to inject more reverberation into the speech as the environment gets larger.

AV Speech dataset [Ephrat et al.]

Chen et al. CVPR 2022

Visual acoustic matching results

Augmented/mixed reality

Video conference

Chen et al. CVPR 2022

Visual acoustic matching results

	SoundSpaces-Speech						*Acoustic AVSpeech*			
	Seen			*Unseen*			*Seen*		*Unseen*	
	STFT	RTE (s)	MOSE	STFT	RTE (s)	MOSE	RTE (s)	MOSE	RTE (s)	MOSE
Input audio	1.192	0.331	0.617	1.206	0.356	0.611	0.387	0.658	0.392	0.634
Blind Reverberator [61]	1.338	0.044	0.312	-	-	-	-	-	-	-
Image2Reverb [52]	2.538	0.293	0.508	2.318	0.317	0.518	-	-	-	-
AV U-Net [20]	**0.638**	0.095	0.353	**0.658**	0.118	0.367	0.156	0.570	0.188	0.540
AViTAR w/o visual	0.862	0.140	0.217	0.902	0.186	0.236	0.194	0.504	0.207	0.478
AViTAR	0.665	**0.034**	**0.161**	0.822	**0.062**	**0.195**	**0.144**	**0.481**	**0.183**	**0.453**

AV-Speech dataset: Ephrat et al. SIGGRAPH 2018
Blind Reverberator: Valimaki et al. More than 50 years of artificial reverberation, 2016
Image2Reverb: Singh et al. ICCV 2021
AV U-Net: Gao & Grauman CVPR 2019

+ User perception studies (not shown)

Chen et al. CVPR 2022

Summary

Kristen Grauman
grauman@cs.utexas.edu

Generating visually-coherent audio

- Discovering sounding actions from video via multimodal consensus (CVPR'24)
- Action2sound to disentangle ambient and action sounds (ECCV'24)
- ActiveRIR for embodied environment acoustics learning (IROS'24)
- Self-supervised visual acoustic matching (CVPR'22, NeurIPS'23)

Changan Chen

Arjun Somayazulu

Sagnik Majumder

Ziad Al-Halah