4D Audio-Visual Perception: Simulating, Synthesizing and Navigating with Sounds in Spaces

Changan Chen
changan.io
UT Austin
12/14/2023

TEXAS
The University of Texas at Austin

Human perception is multisensory

We often use *vision*, *audio*, *touch*, *smell* to sense the world


The status quo of audio-visual learning

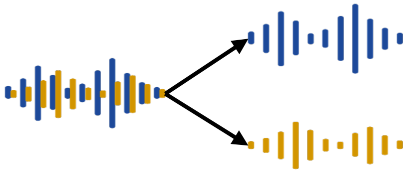
Object-centric:

the semantic correspondence between sight and sound of objects

Classification

Drum Piano

Separation



Localization

Source: Drumeo

1 drum kit 5 different spaces

Source: Shred Shed Studio

Autonomous agents

Home assistance robot

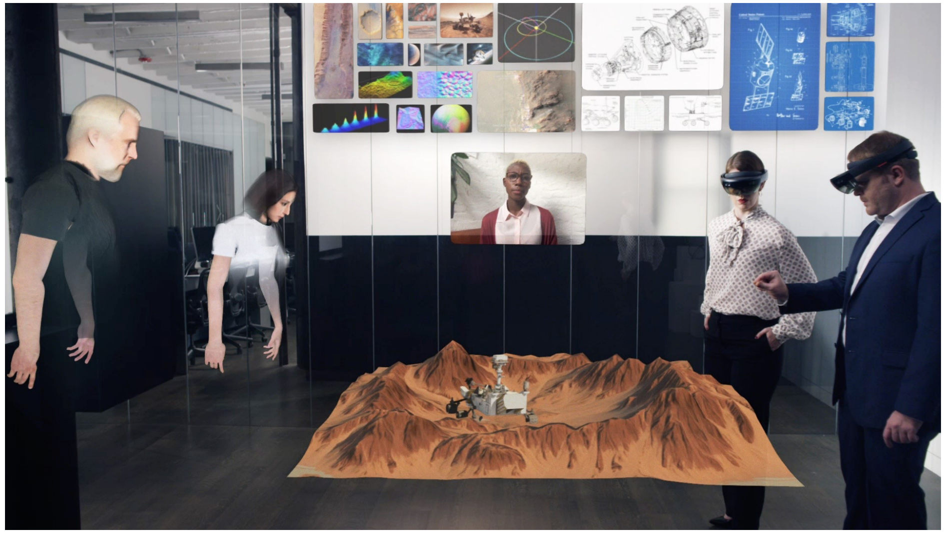
Rescue robot

Robots that can navigate and localize sounding objects by reasoning the spatial, semantic, acoustic information in the audio and visual observation

Augmented reality and virtual reality

Enhanced hearing

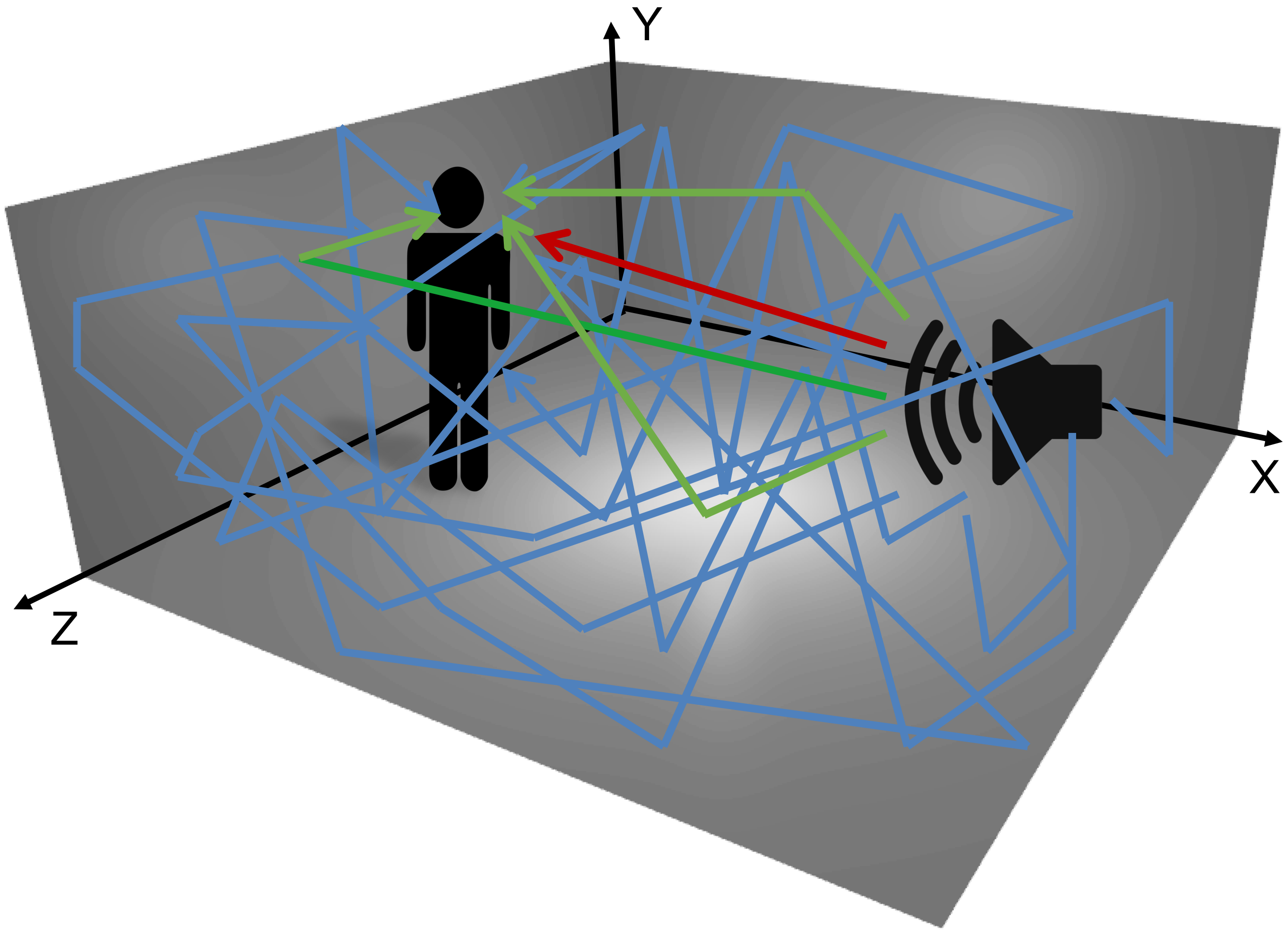
Immersive experience

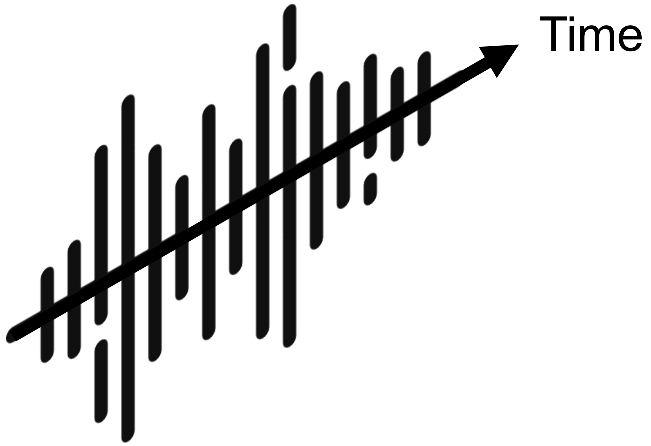


AR/VR systems that can augment the hearing ability of the device wearer as well as create immersive experiences for users

4D audio-visual perception

My research: learning the correspondence between sight and sound in spaces



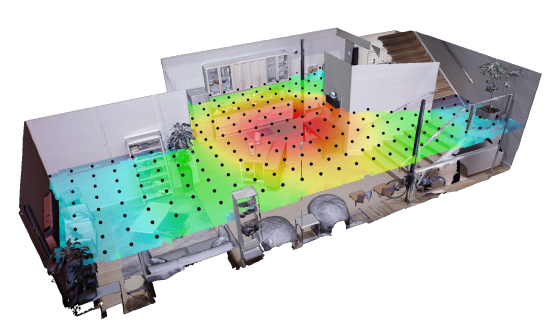


4D audio-visual perception

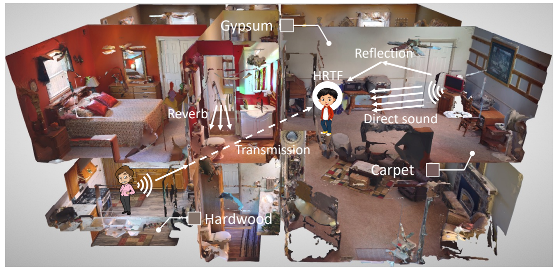
My research: learning the correspondence between sight and sound in spaces

Simulating sounds in spaces

SoundSpaces [ECCV20]

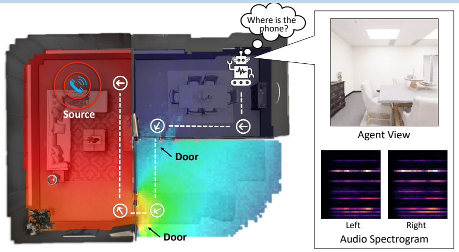


SoundSpaces 2.0 [NeurIPS22]

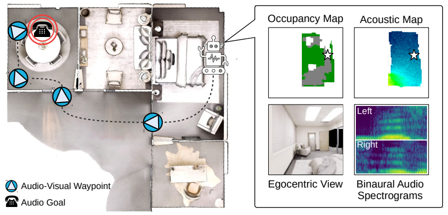


Navigating with sounds in spaces

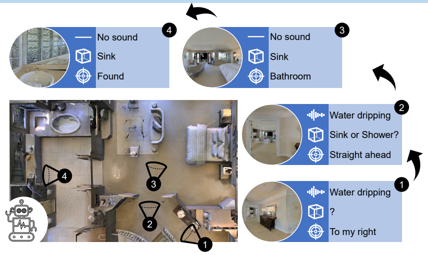
Audio-visual navigation SoundSpaces [ECCV20]



Efficient & hierarchical AV nav [ICLR21]

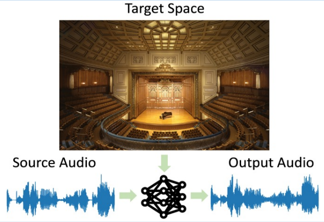


Semantic audio-visual navigation [CVPR21]

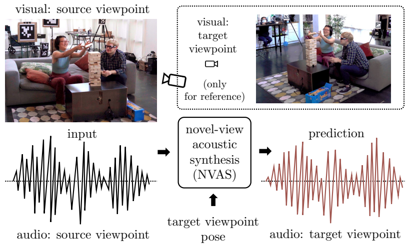


Synthesizing sounds in spaces

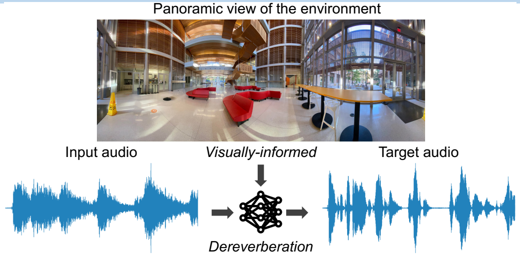
Visual acoustic matching [CVPR22]



Novel-view acoustic synthesis [CVPR23]



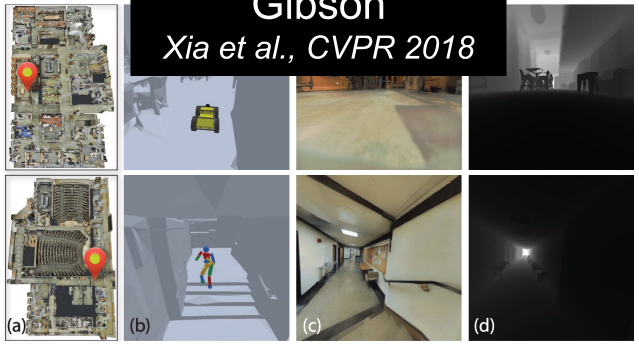
Audio-visual dereverberation [ICASSP23]



Simulating embodiment in 3D scenes

Datasets

Gibson
Xia et al., CVPR 2018



Matterport3D
Chang et al., 3DV 2017

Replica
Straub et al., arXiV 2019

Simulators

Habitat
Savva et al., ICCV 2019

iGibson
Xia et al., ICRA 2020

AI2THOR
Kovle et al., arXiV 2017

Advantages: Large-scale training, fast experimentation, consistent benchmarking and replicable research

Sim2Real

Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance, Kadian et al., IRAL 2020
Sim-to-Real Transfer for Vision-and-Language Navigation, Anderson et al., CoRL 2020
RoboThor: An Open Simulation-to-Real Embodied AI Platform, Deitke et al., CVPR 2020

Enabling embodied agents and tasks

Source: Gibson

Today's embodied agents (robots) are deaf

- We want robots that can see, hear and react in the environment

Vision-Only

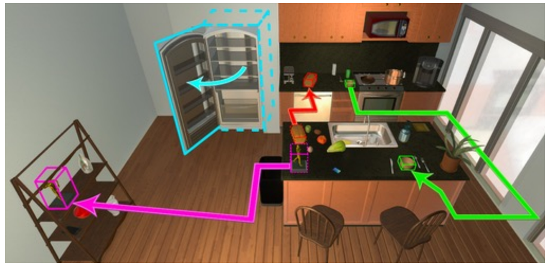
Gupta et al., 2017
Zhu et al., 2017
Sava et al., 2019
…

Vision-Language

Anderson et al., 2018
Wang et al., 2018
Wang et al., 2019
…

Vision-Interaction

Zhu et al., 2017
Gordon et al., 2018
Wortsman et all, 2019
…



Vision-Audio

Chen and Jain et al., 2020
(this work)

- No existing simulation supports audio-visual rendering
- No existing formulation for audio-visual navigation

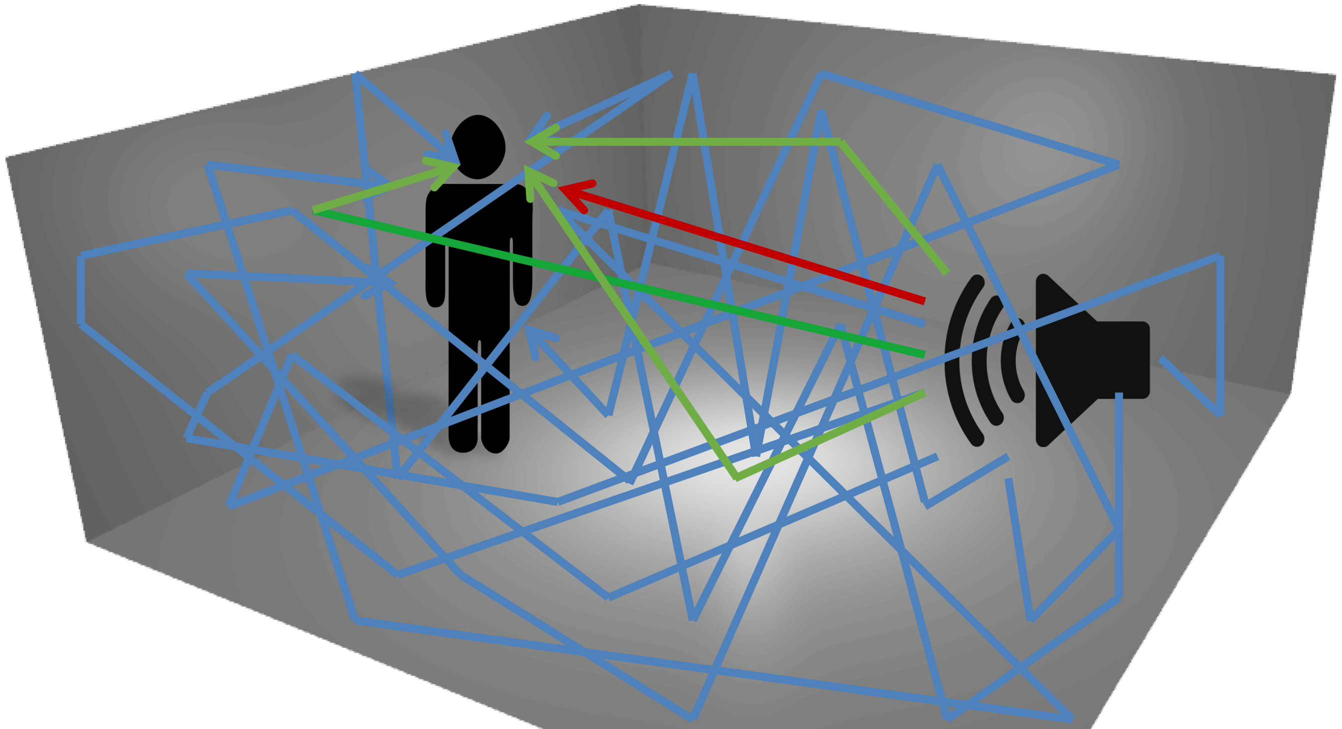
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

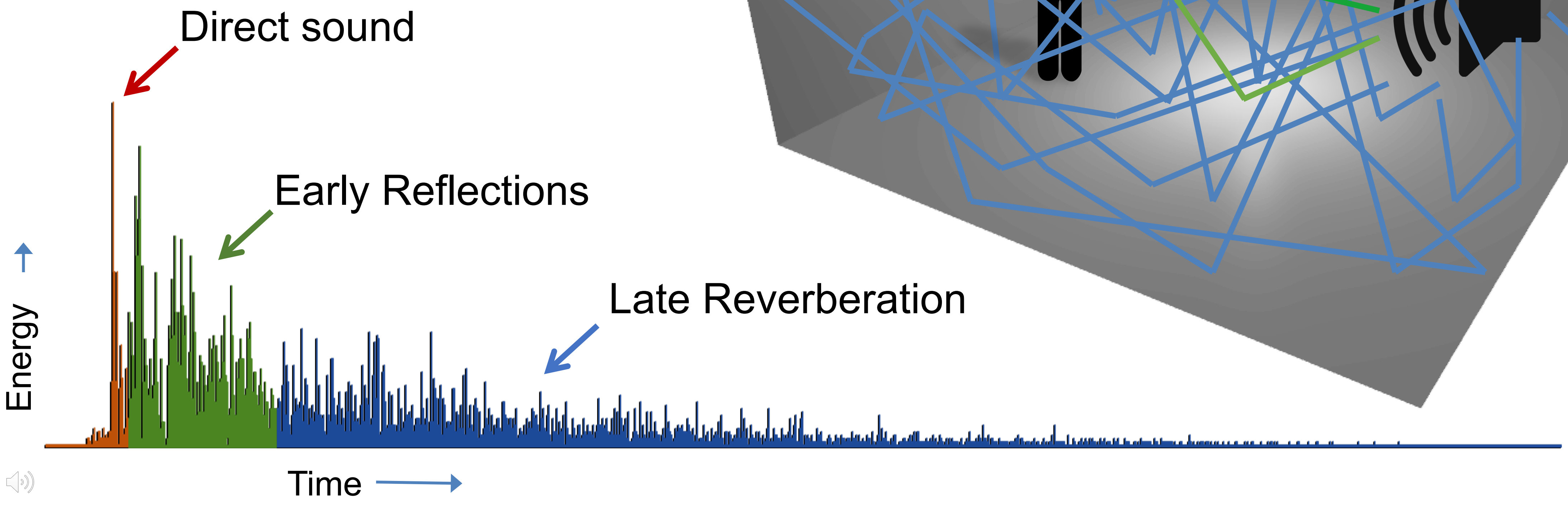
SoundSpaces demo

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Background: acoustic simulation

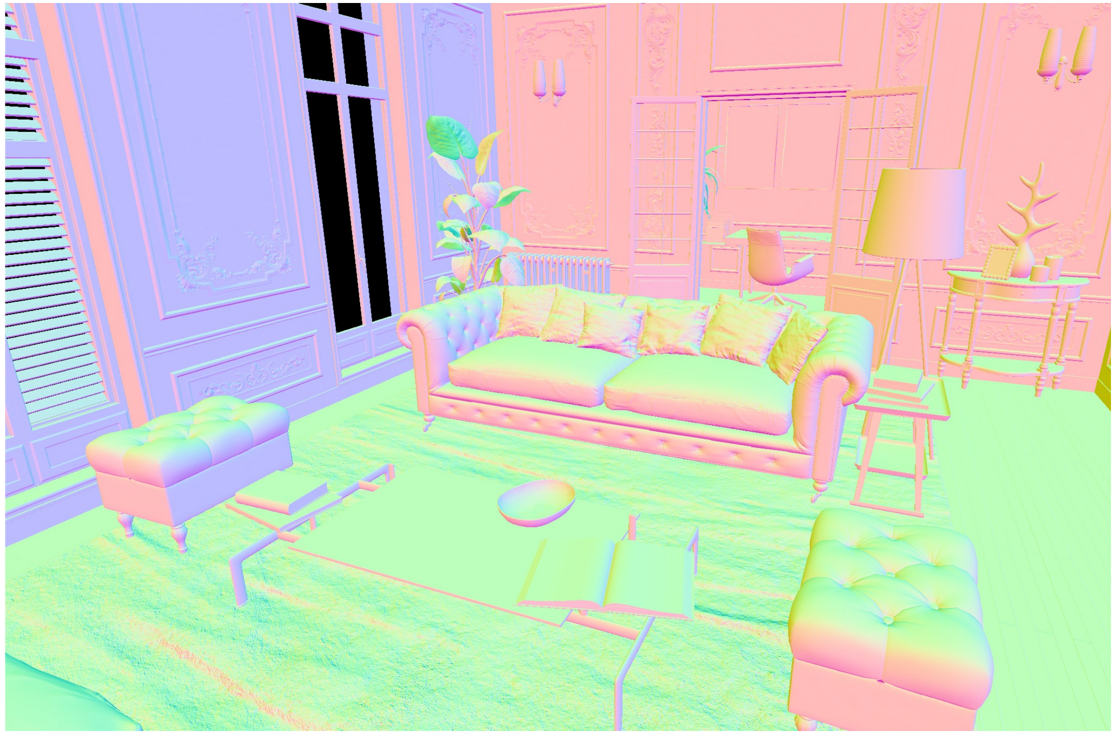
Goal: simulate a perceptually-valid approximation of the room impulse response (RIR)



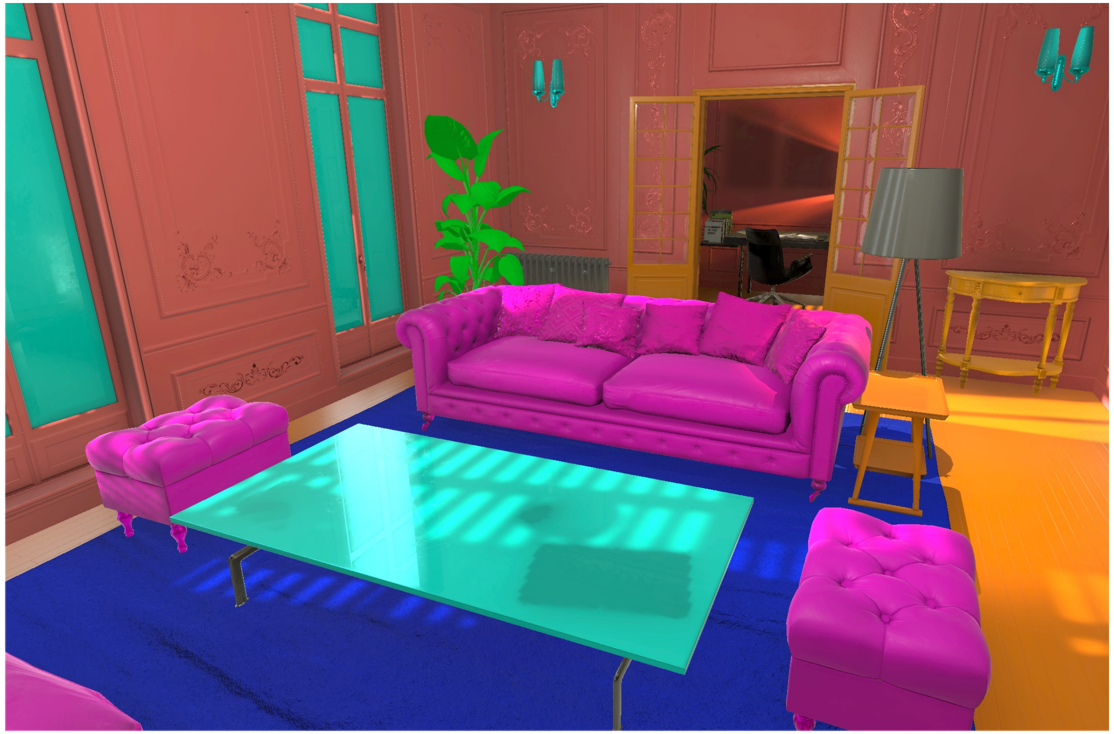


Physics-based audio rendering

3D Geometry



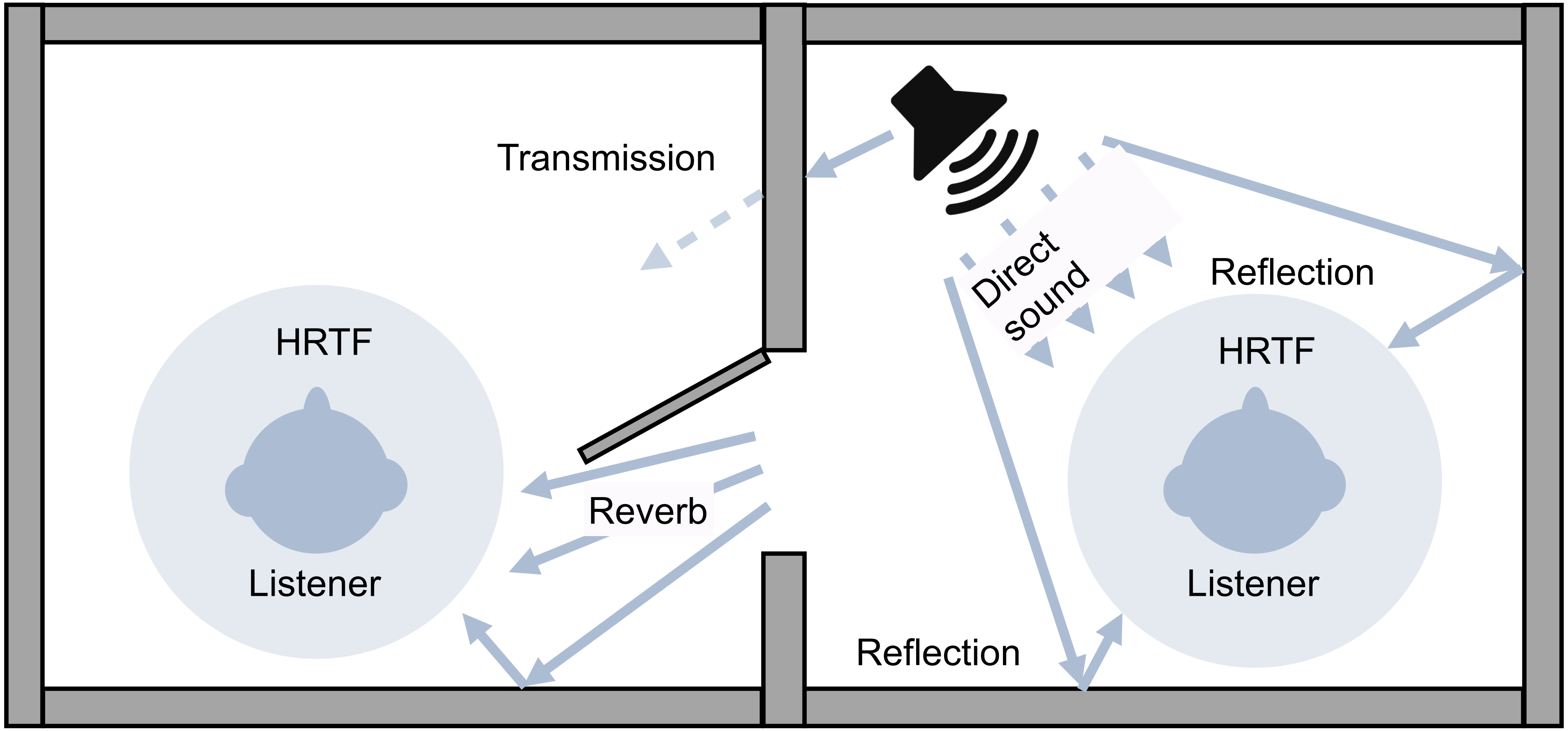
Material Properties



Simulate the sound received by the listener from a source location

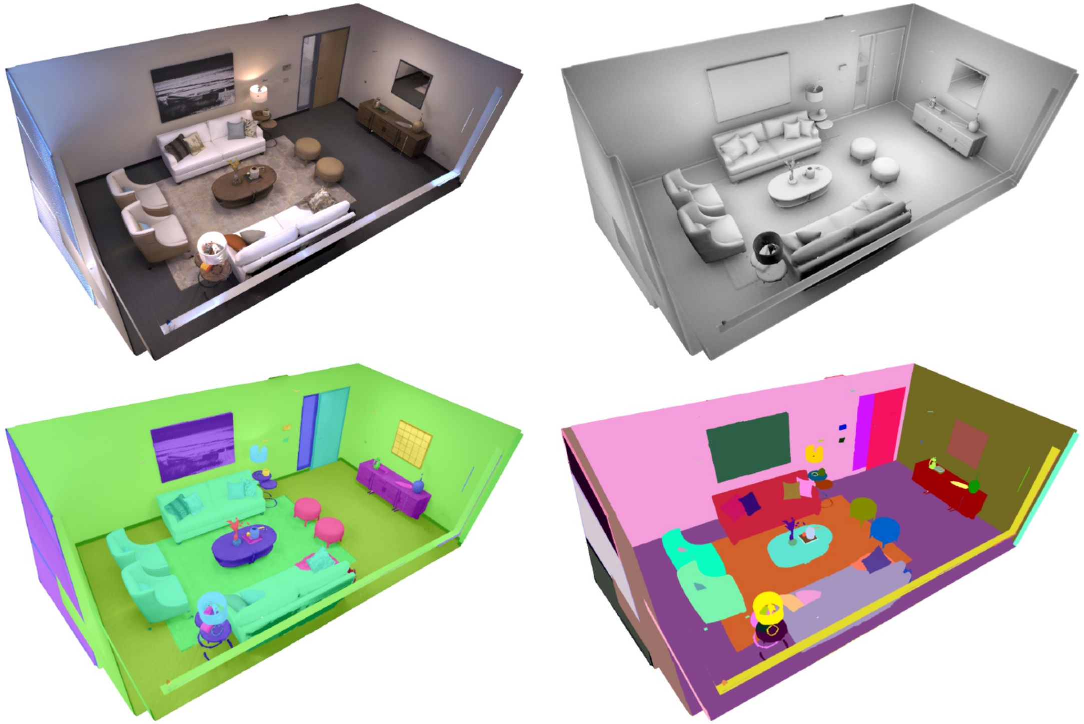
Sound propagation system

3D spatial audio for reflections and reverb with realistic acoustics based on bidirectional ray tracing

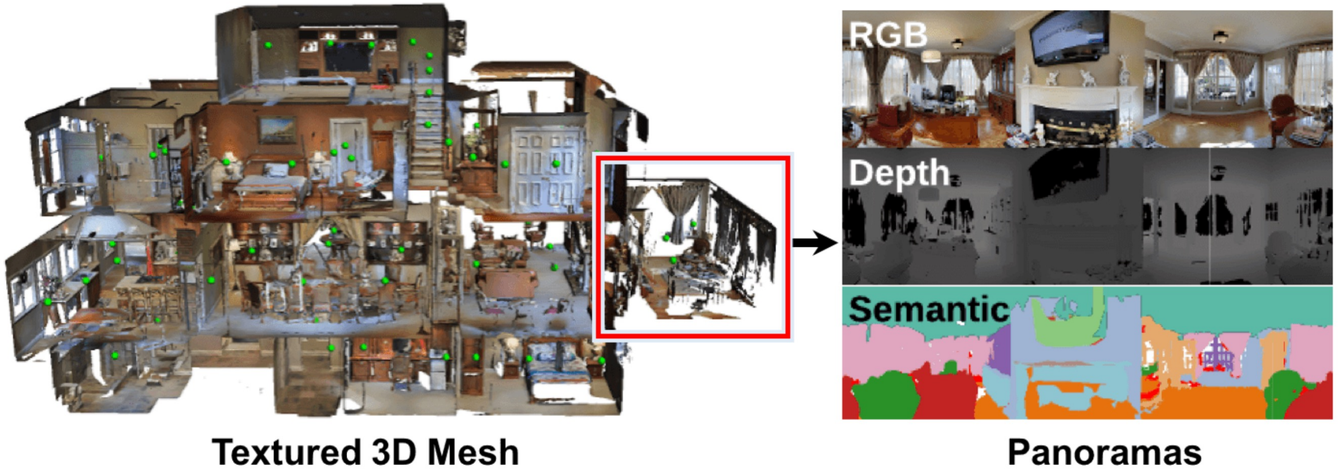


Real-scan environments

Replica[1] dataset



Matterport3D[2] dataset

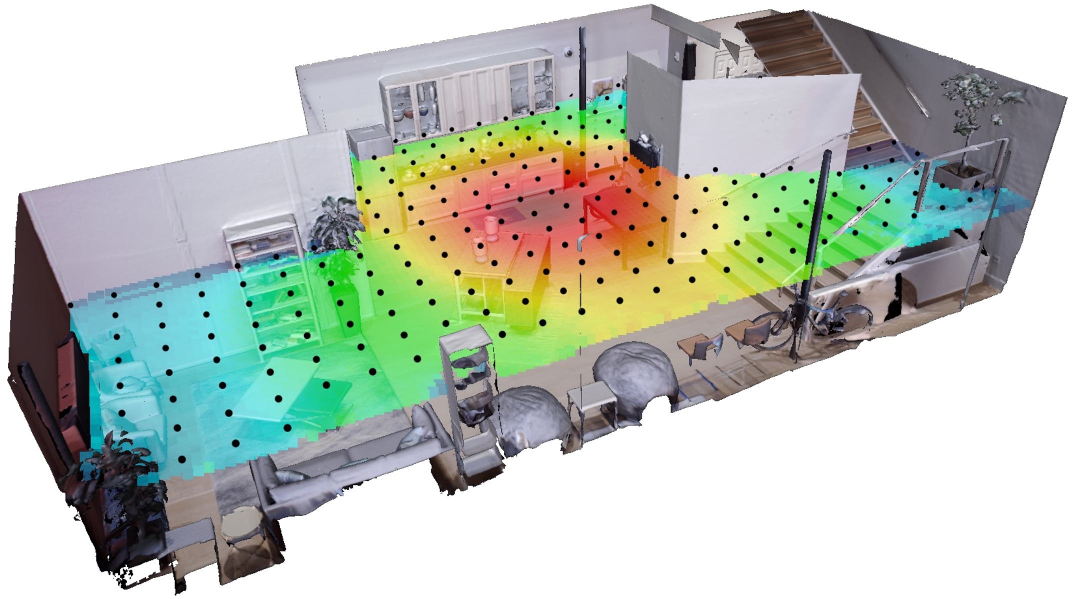


[1]The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019
[2]Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

SoundSpaces: our audio simulator

SoundSpaces produces realistic audio rendering based on the room geometry, materials, and sound source location by **precomputing** the room impulse response function (RIR)

Users can insert any sound of their choice at runtime. The received sound is obtained by convolving the RIR with the source sound.



	# Scenes	Avg. Area	# RIRs
Replica	18	47.24 m^2	0.9M
Matterport3D	85	517.34 m^2	16.7M

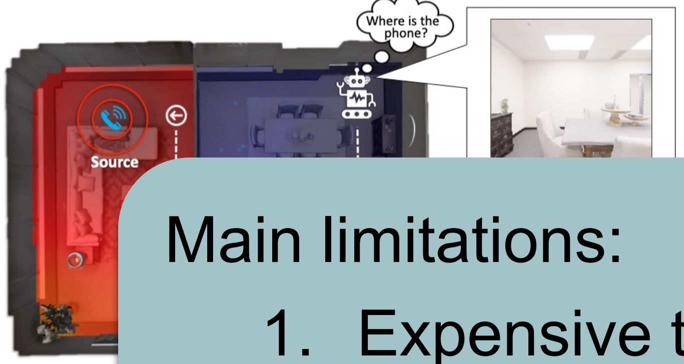
Table: Summary of dataset statistics

Visit soundspaces.org for more information!

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

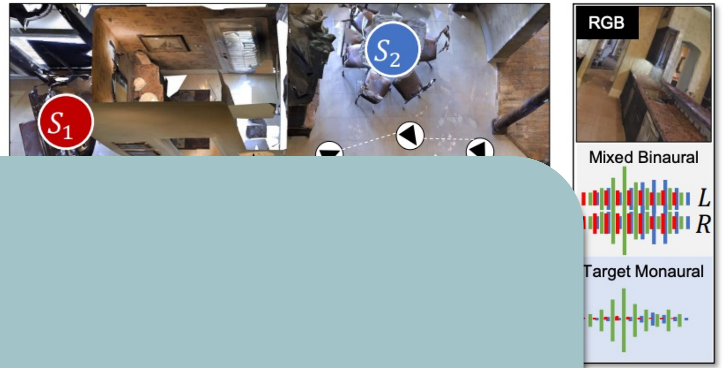
Enabling audio-visual embodied AI and beyond

Audio-Visual Navigation
Chen et al., ECCV 2020



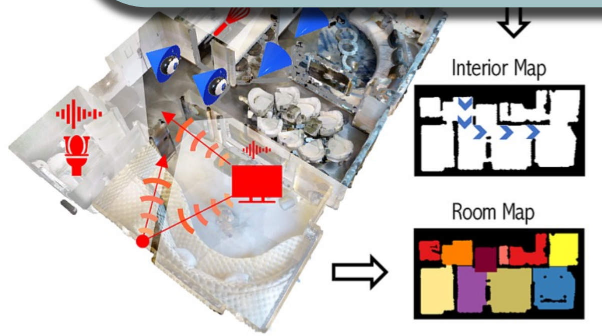
Echolocation Learning
Gao et al., ECCV 2020

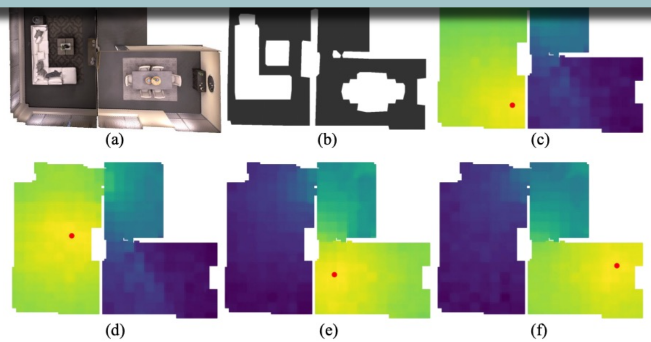
Audio-Visual Separation
Majumder et al., ICCV 2021

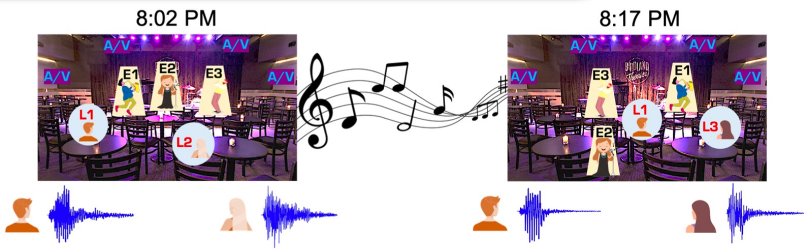


Main limitations:

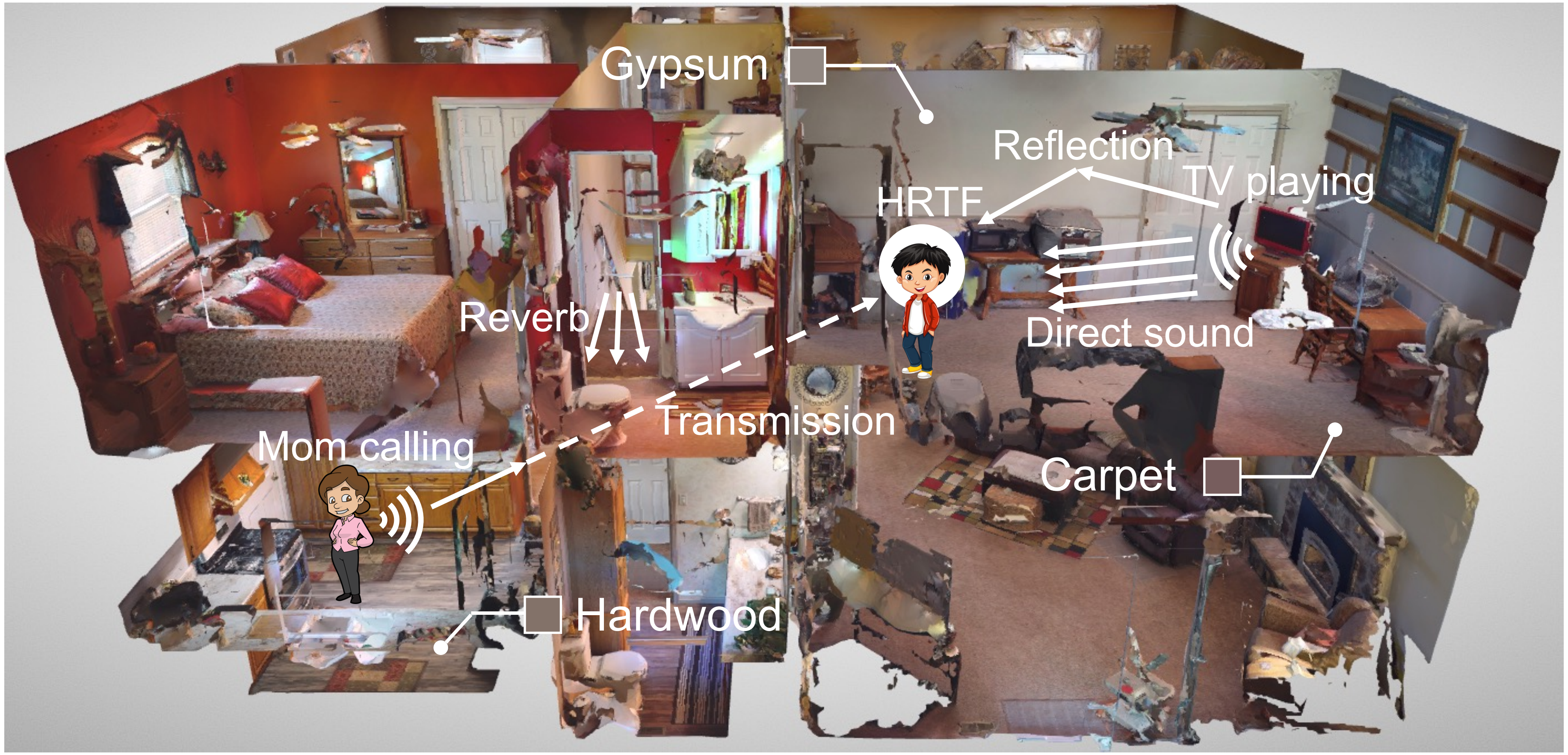
1. Expensive to store millions of IRs
2. Does not generalize to new locations or environments
3. Microphones are not configurable

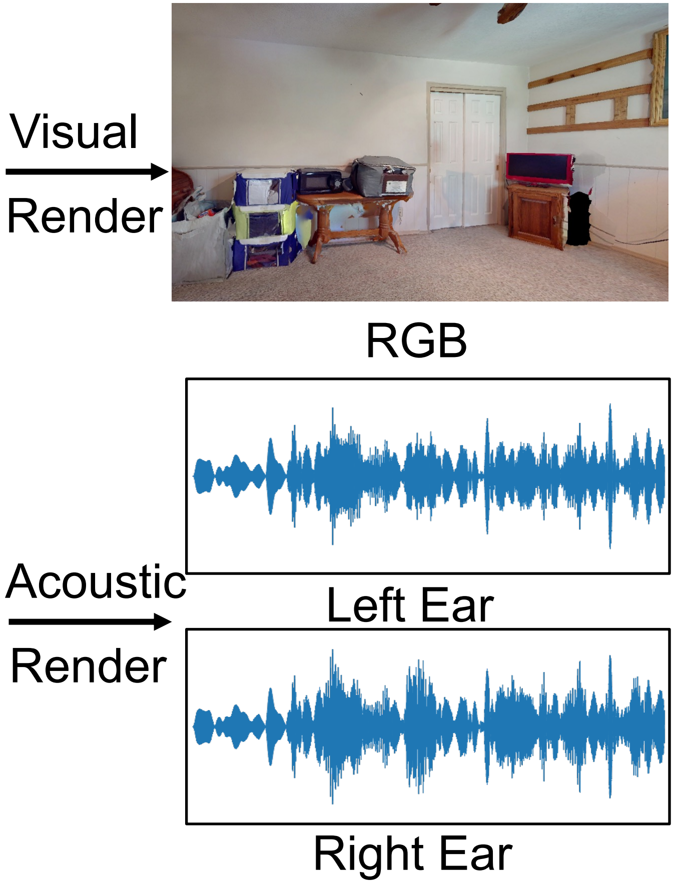






SoundSpaces 2.0: A fast, continuous, configurable and generalizable audio-visual simulation platform

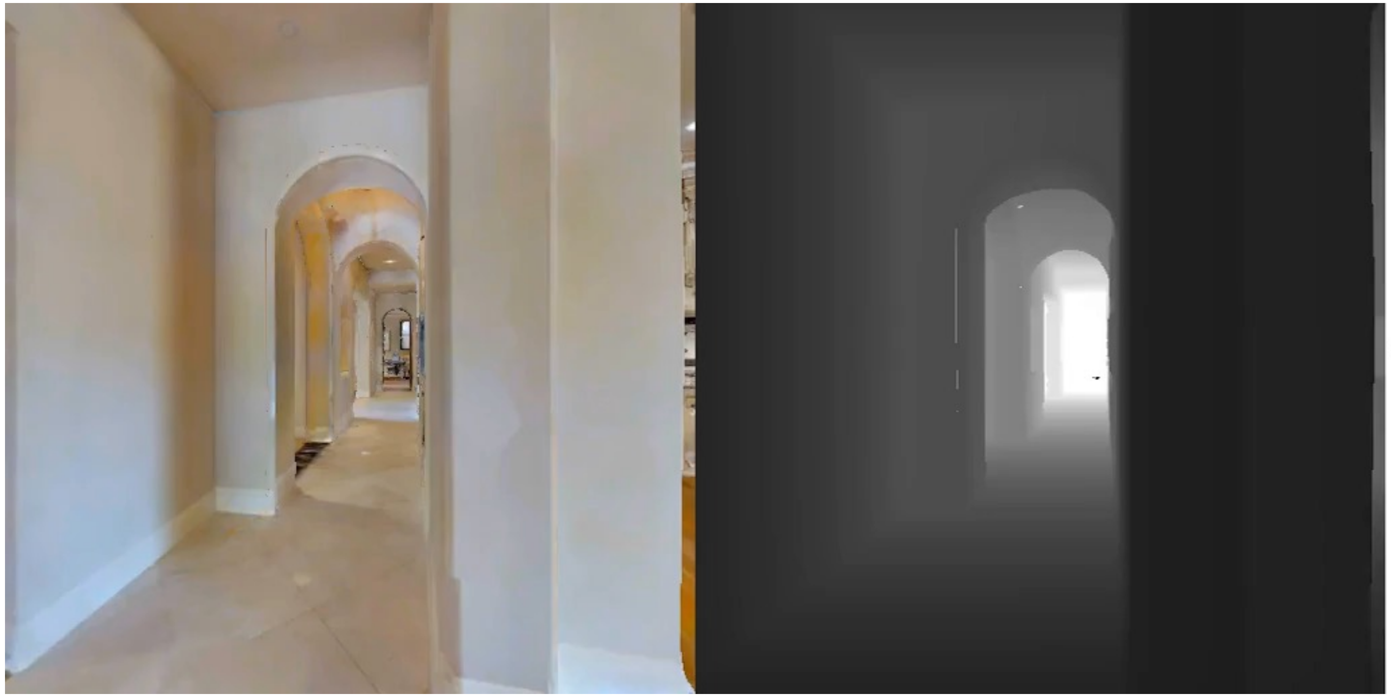




Changan Chen et al., SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022

Continuous rendering

We offer both spatial and acoustic continuity.



Navigating to someone speaking

Changan Chen et al., SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022

Configurable simulation

Users can change all these parameters!

Simulation parameters

- Frequency bands
- Direct sound
- Indirect sound
- Transmission
- Diffraction
- Number of rays
- Number of threads
- Sample rate
- …

Microphone types

- Mono
- Binaural
- Stereo
- Quad
- Surround_5_1
- Surround_7_1
- Ambisonics
- Your mic array
- …

Material properties

- Absorption coefficients
- Scattering coefficients
- Transmission coefficients
- Damping coefficients
- Frequency band specs
- Instance level config
- …

Changan Chen et al., SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022

Generalizable simulation

We support arbitrary scene datasets.

Gibson
Xia et al., CVPR 2018

Matterport3D
Chang et al., 3DV 2017


Replica
Straub et al., arXiV 2019


HM3D
Ramakrishnan et al., NeurIPS 2021

Ego4D
Grauman et al., CVPR 2022

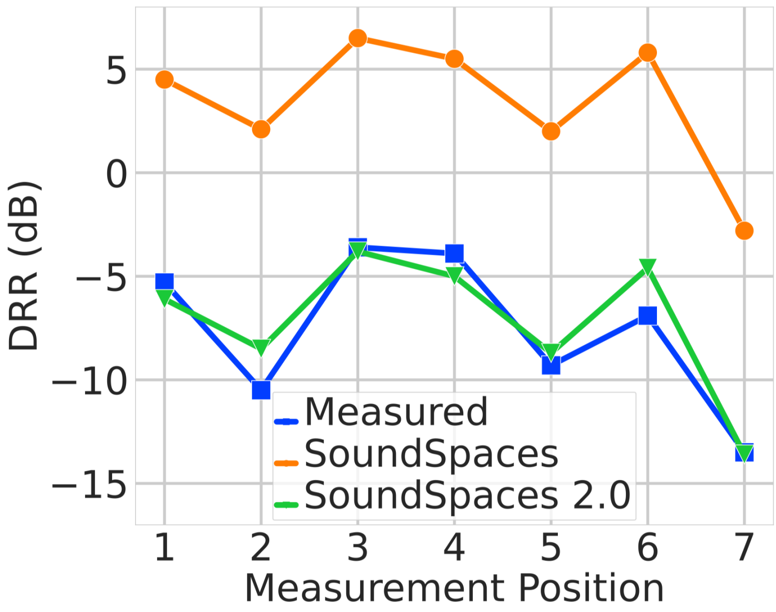

Your own environment!

Changan Chen et al., SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022

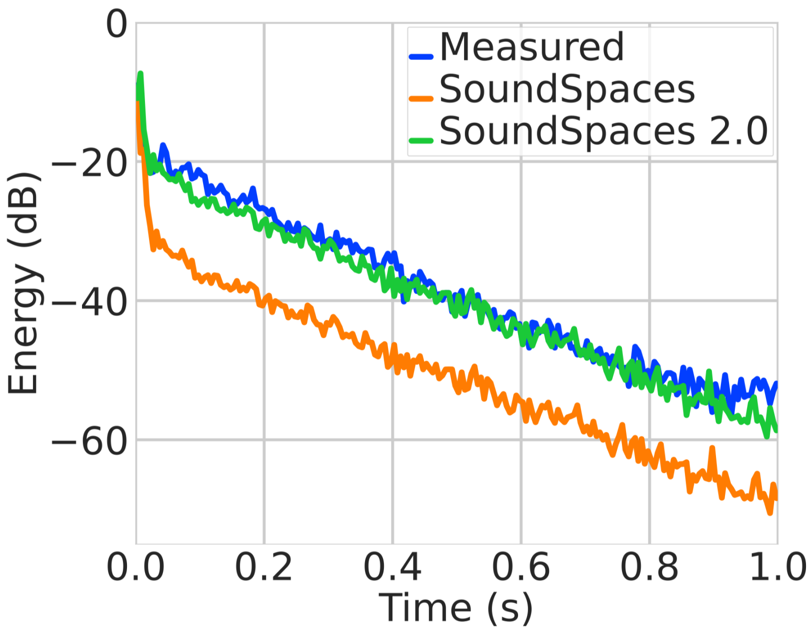
Validating simulation with real IRs

We collect acoustic measurements of the apartment in Replica dataset and compare to IRs rendered in SoundSpaces

SoundSpaces 2.0 has a better match of direct-to-reverberant ratio with real



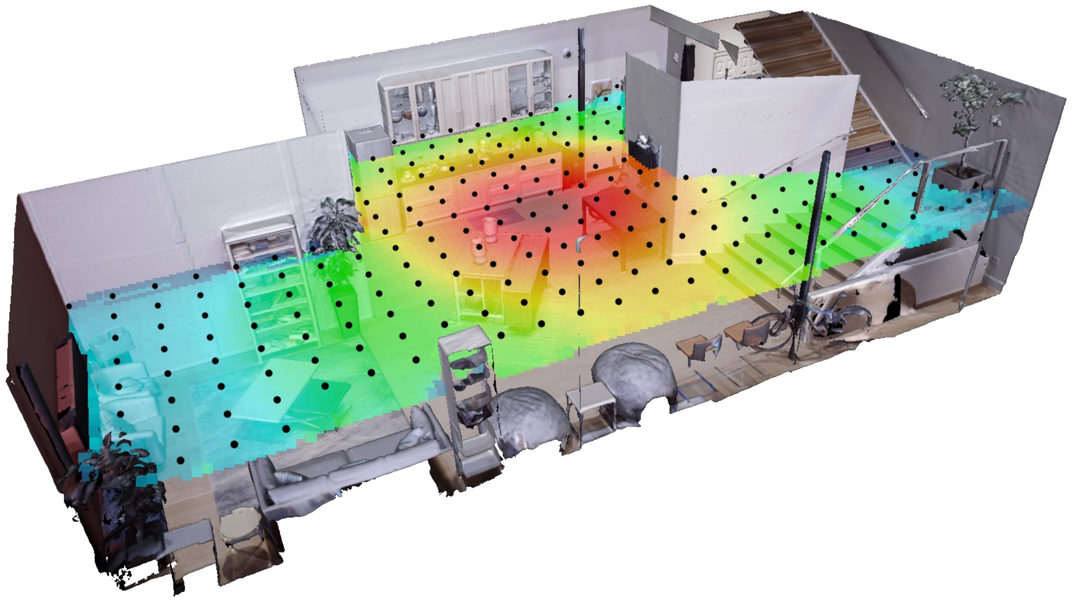
DRR comparison

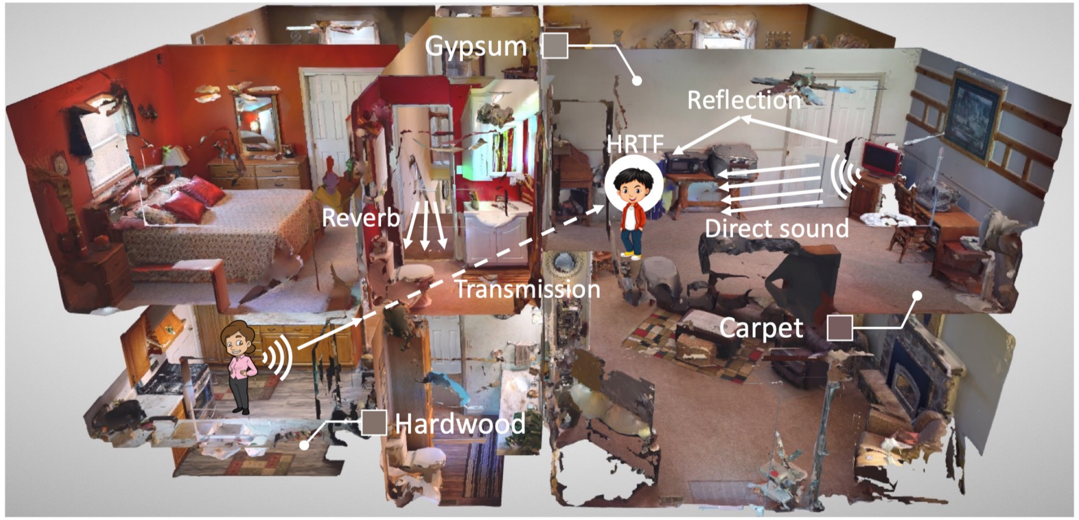


Early decay comparison

Changan Chen et al., SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022

Main differences





SoundSpaces 1.0

- 500 fps+
- Discrete and unconfigurable

SoundSpaces 2.0

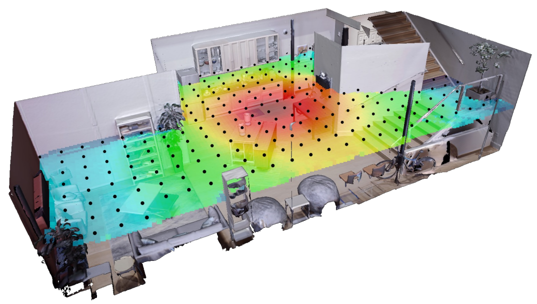
- 30 fps+
- Continuous and configurable

4D audio-visual perception

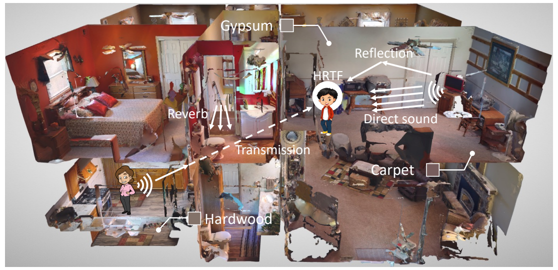
My research: learning the correspondence between sight and sound in spaces

Simulating sounds in spaces

SoundSpaces [ECCV20]

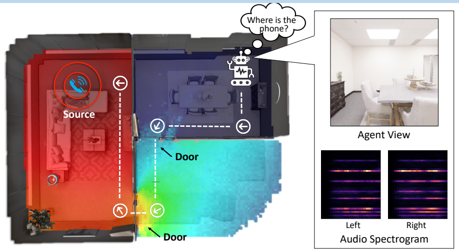


SoundSpaces 2.0 [NeurIPS22]

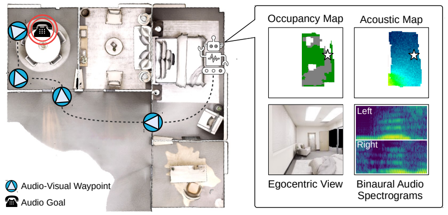


Navigating with sounds in spaces

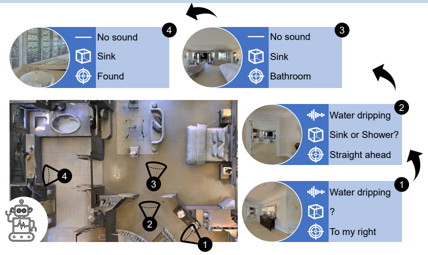
Audio-visual navigation SoundSpaces [ECCV20]



Efficient & hierarchical AV nav [ICLR21]

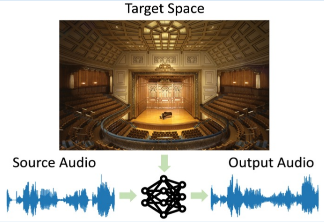


Semantic audio-visual navigation [CVPR21]

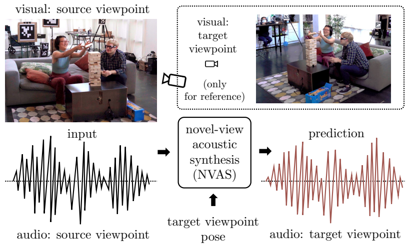


Synthesizing sounds in spaces

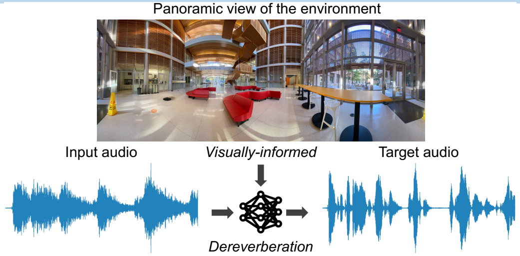
Visual acoustic matching [CVPR22]



Novel-view acoustic synthesis [CVPR23]

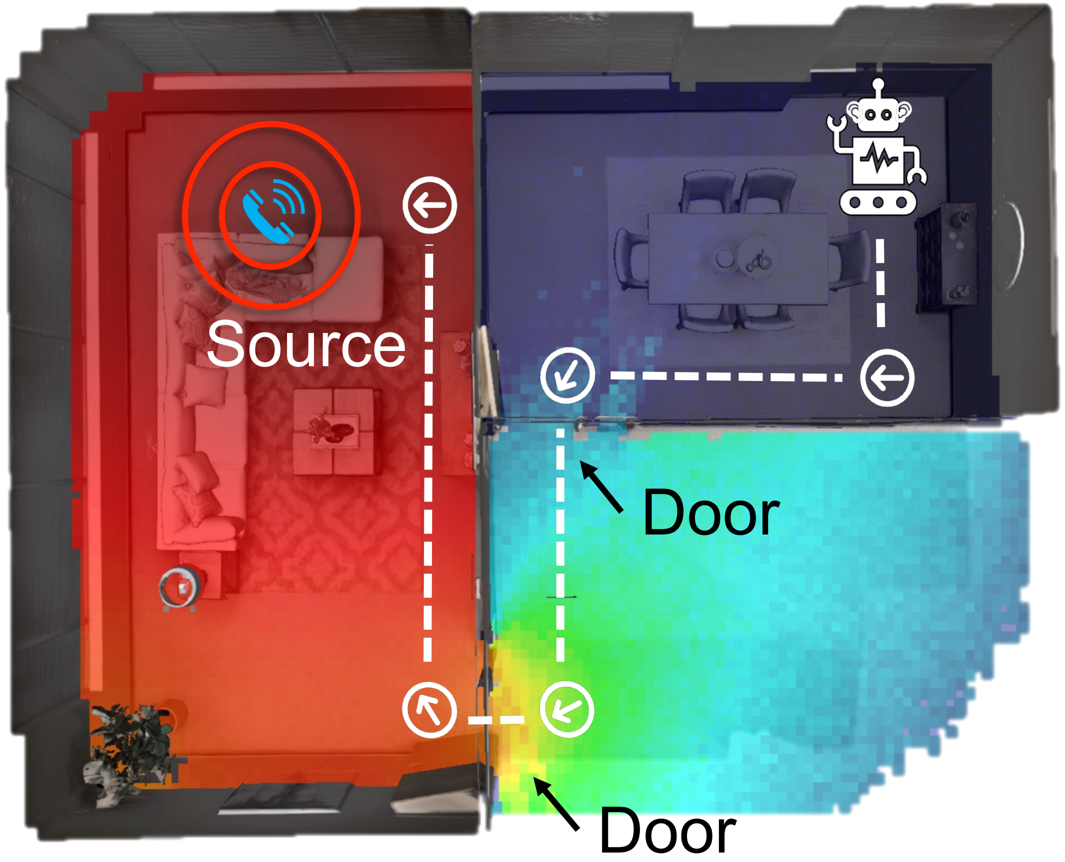


Audio-visual dereverberation [ICASSP23]



Audio-visual navigation in 3D environments

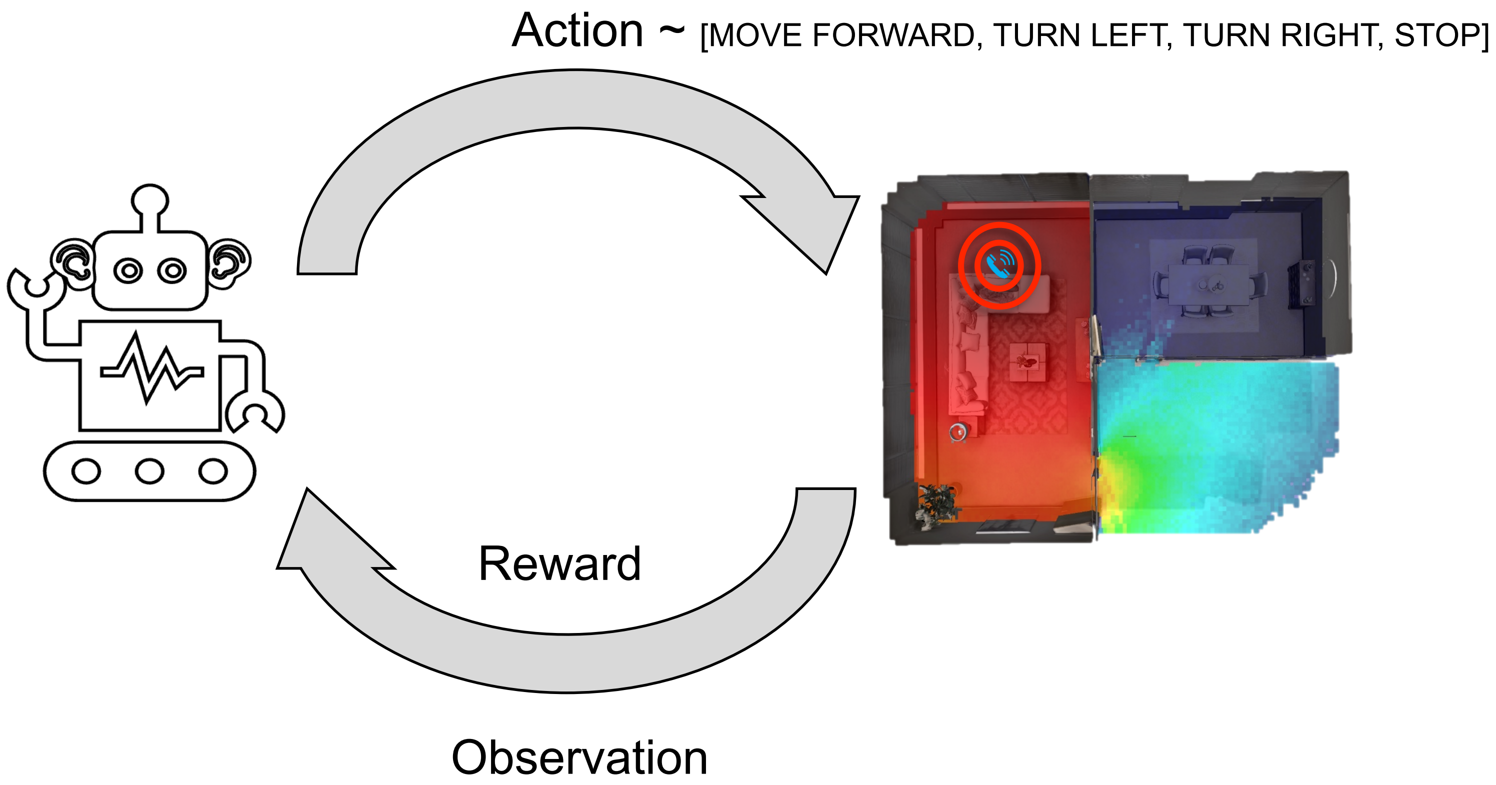
An agent navigates to a sounding object with vision and audio perception



C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

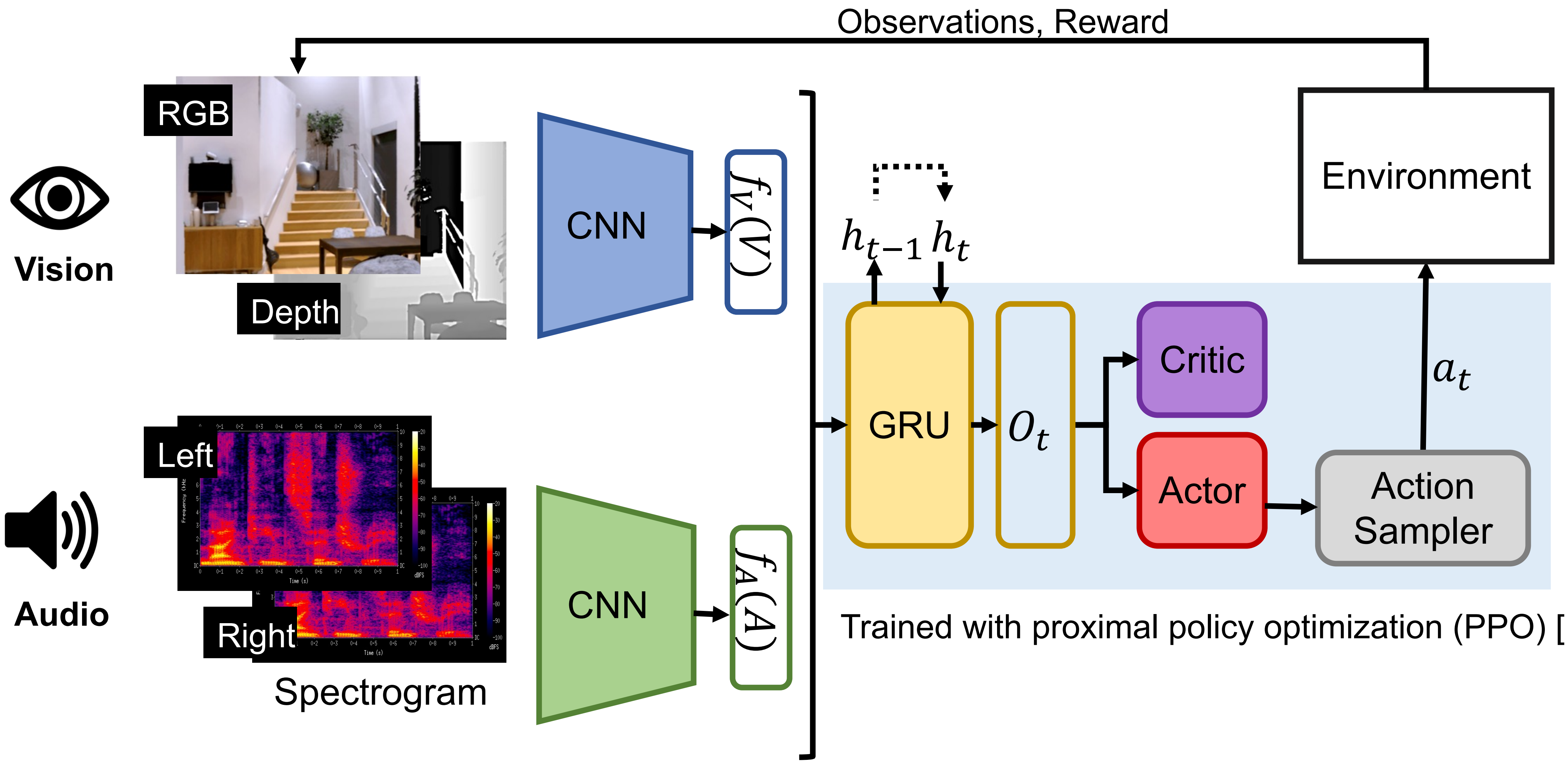
Learning with deep reinforcement learning

- Learn to navigate in simulation via trials and errors
- Rewarded +1 for getting close and +10 for reaching the goal



C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Navigation policy

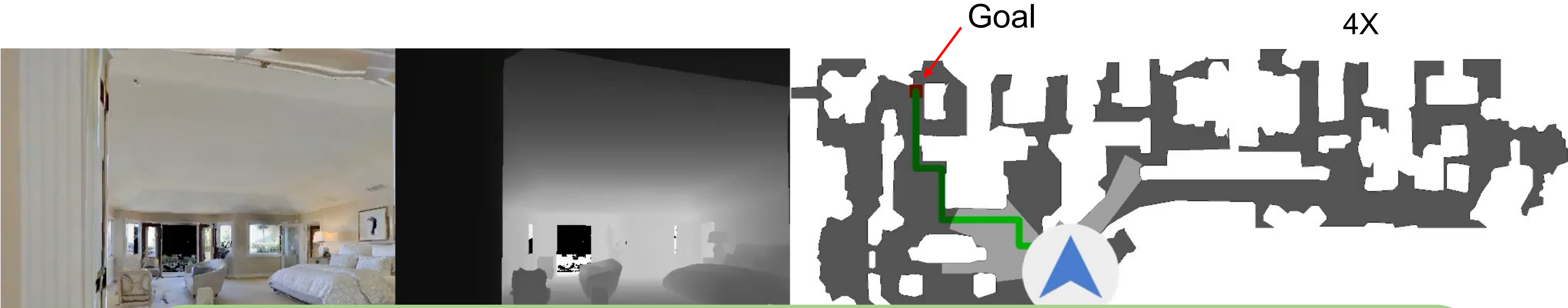


Trained with proximal policy optimization (PPO) [1]

[1] Proximal Policy Optimization Algorithms, John Schulman et al., arxiv 2017

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Navigation example



Key messages:

1. Embodied agent can locate sounds by seeing and hearing
2. A blind agent can also navigate by only using binaural cues

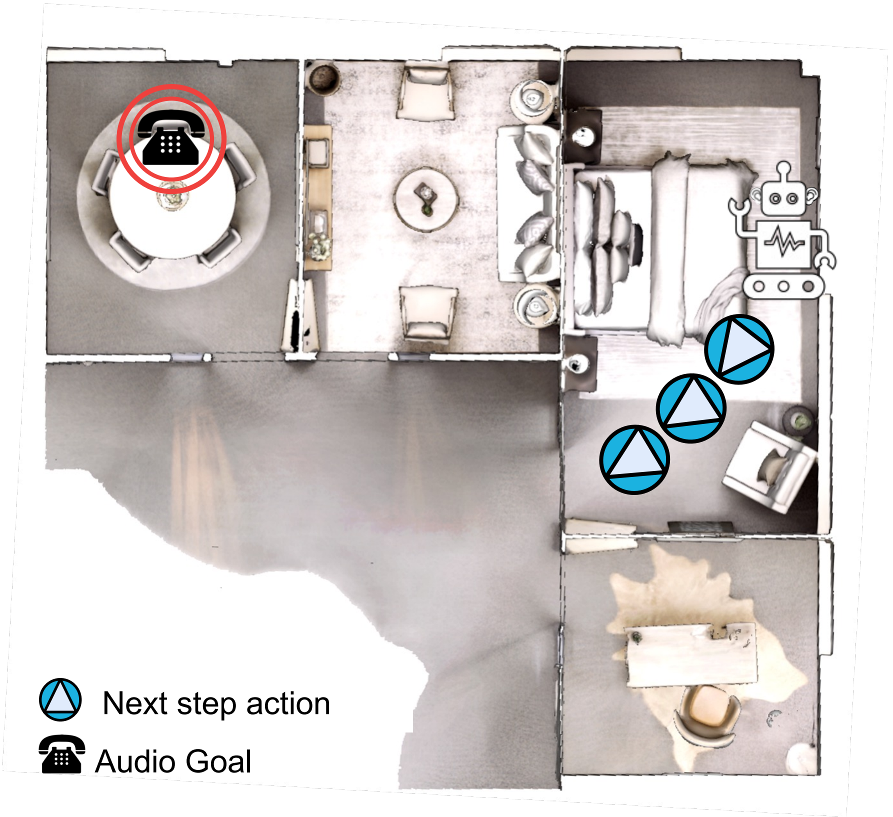
Agent Goal Start Shortest path Agent path Seen/Unseen area Occupied area Red Frame: Collision

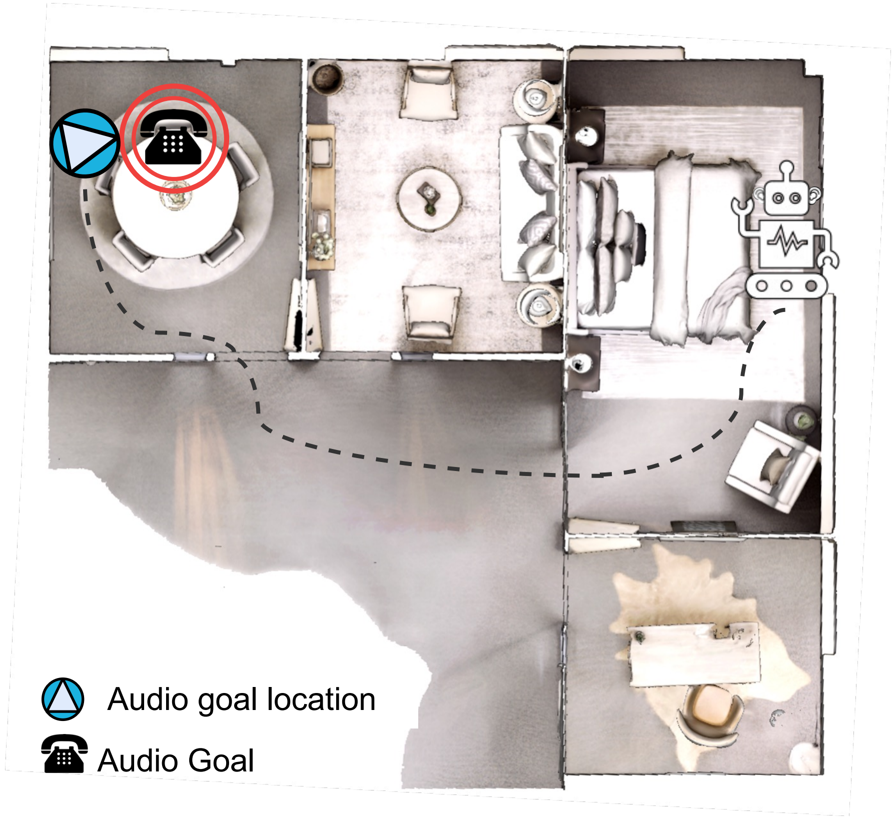
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Limitations of the navigation policy

Existing models learn to act at fixed granularities of action motion

- Chen et al.[1]: learn to generate primitive actions step-by-step
- Gan et al.[2]: predict target locations and navigate with geometric planner

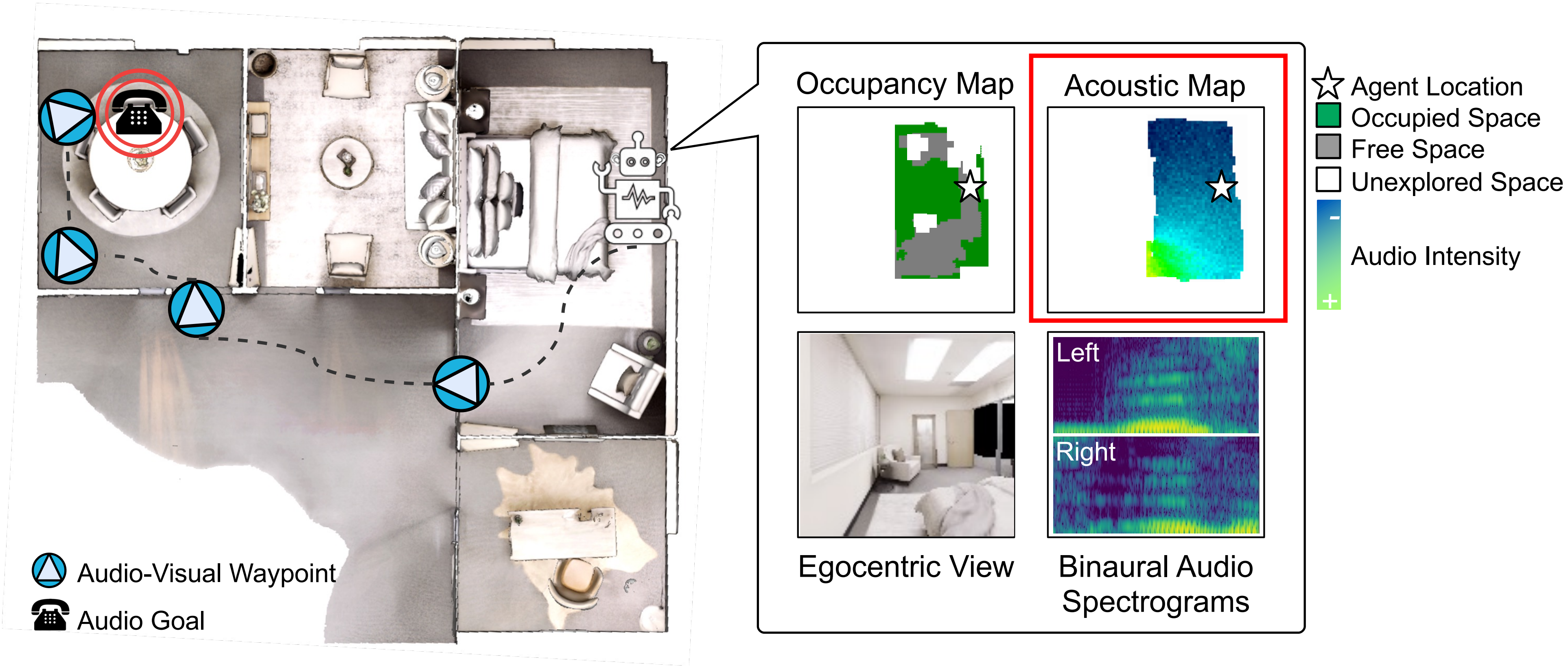




[1]SoundSpaces: Audio-Visual Navigation in 3D Environments, Chen et al., ECCV, 2020
[2]Look, Listen, and Act: Towards Audio-Visual Embodied, Gan et al., ICRA, 2020

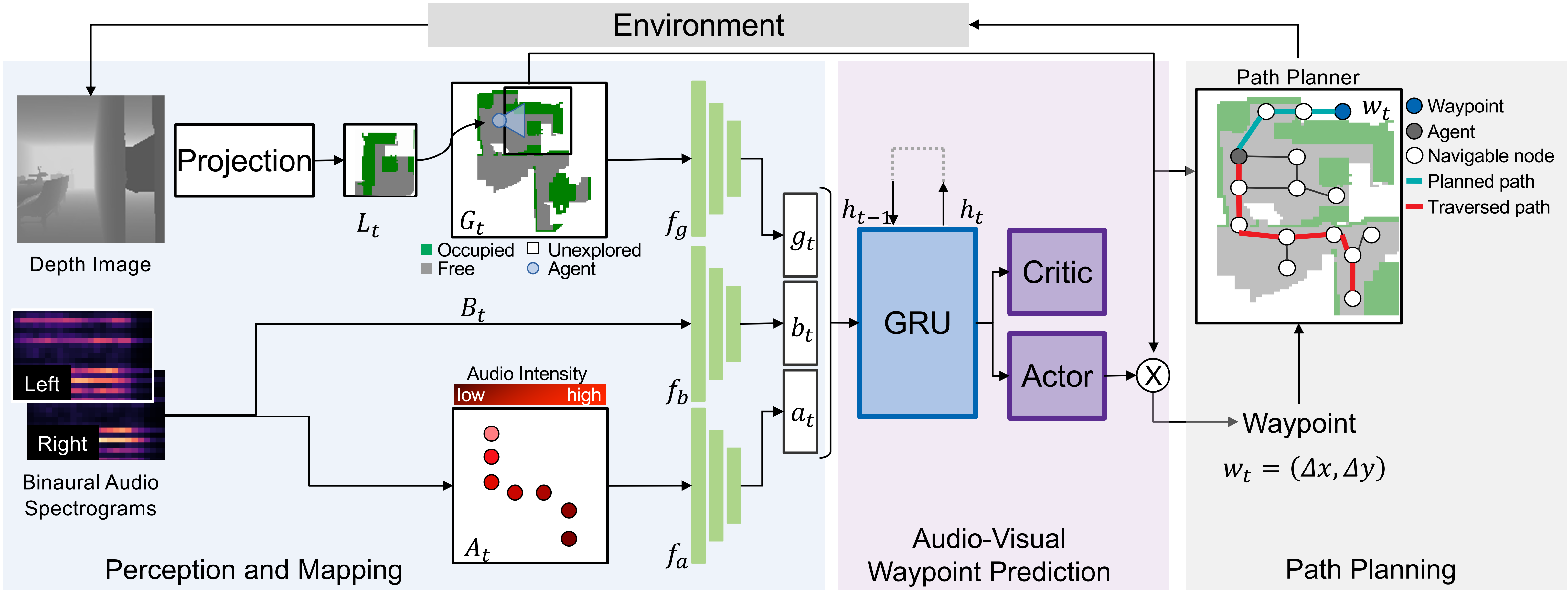
Learning to set waypoints for AV navigation

- Infer audio-visual subgoals with RL end-to-end at varying granularities
- Acoustic memory to help infer goal locations and decide stop actions



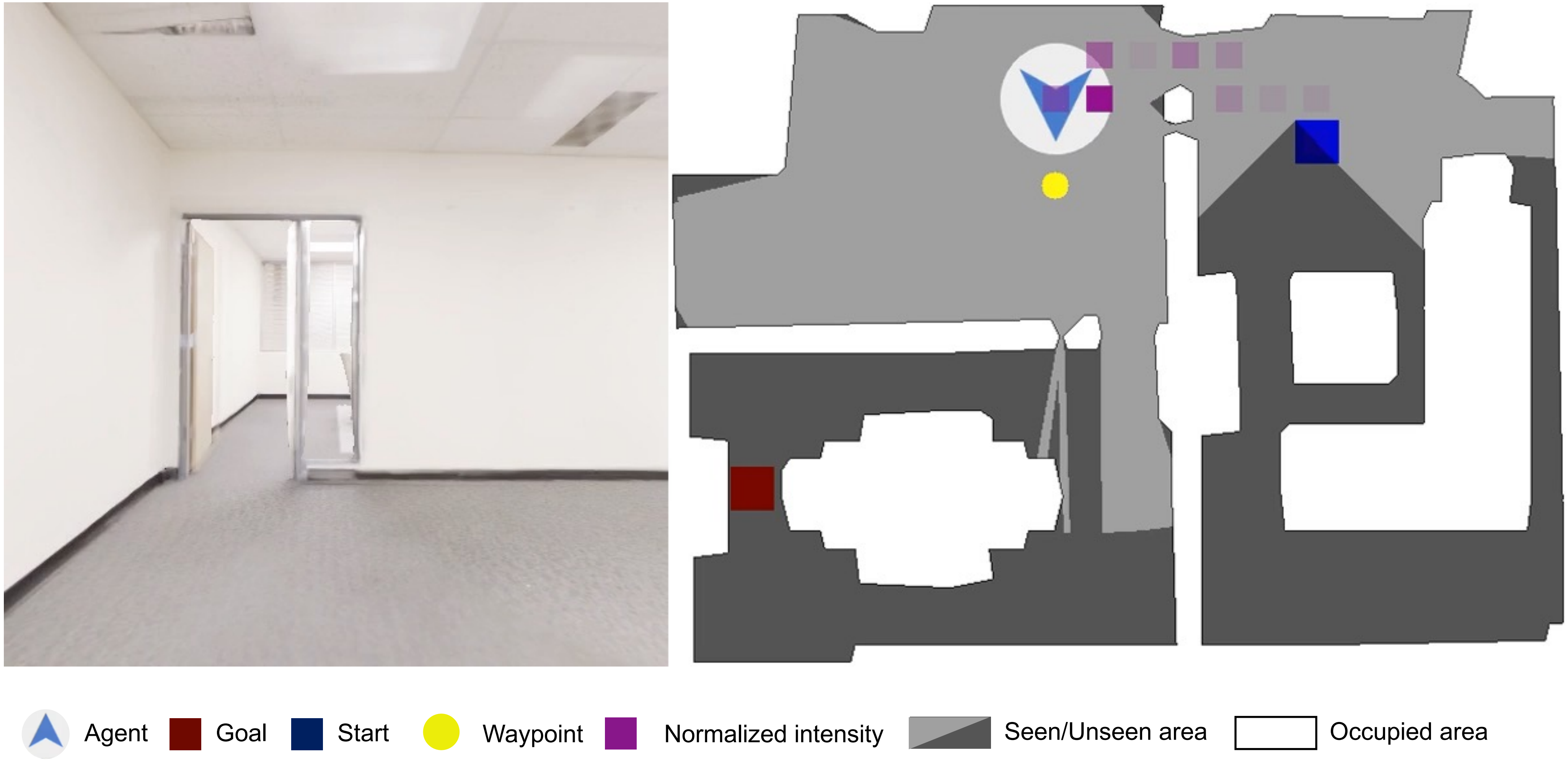
Changan Chen et al., Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021

Audio-visual waypoints navigation model (AV-WAN)



Changan Chen et al., Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021

Waypoint selection and acoustic memory

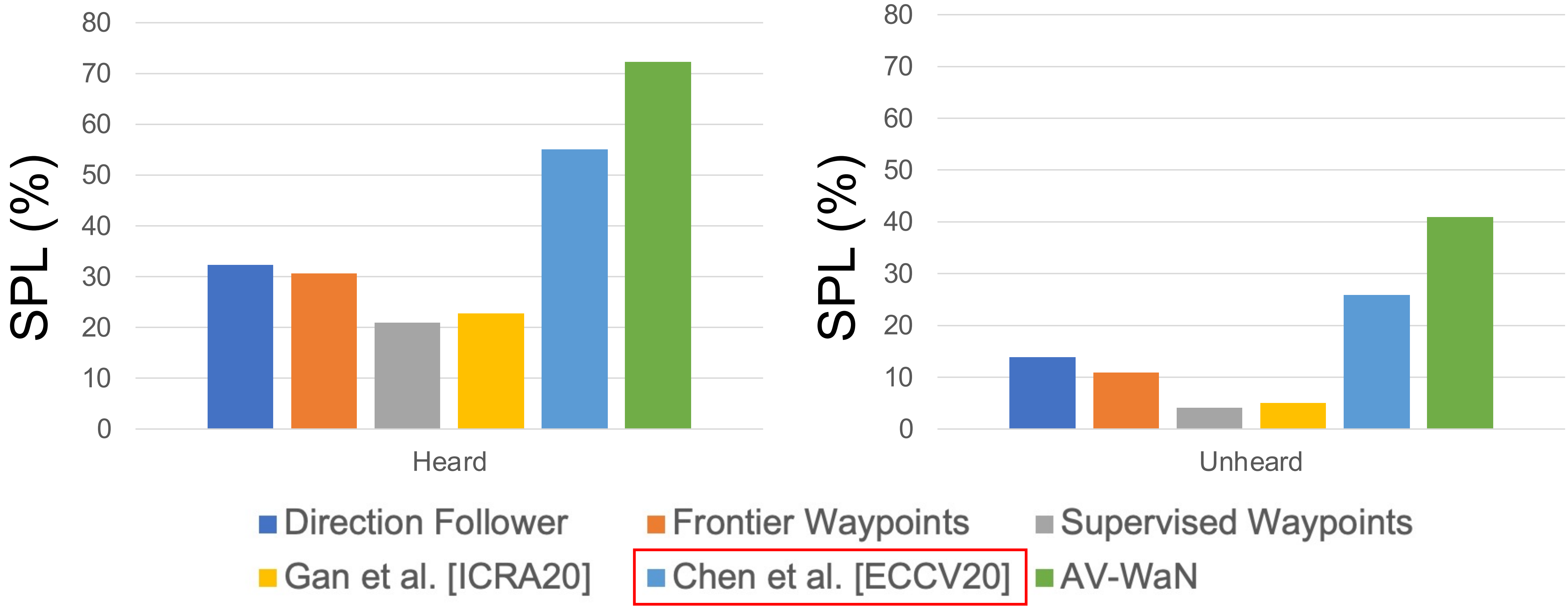


Our model dynamically selects waypoints and builds an acoustic memory as it moves.

Changan Chen et al., Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021

Navigation results

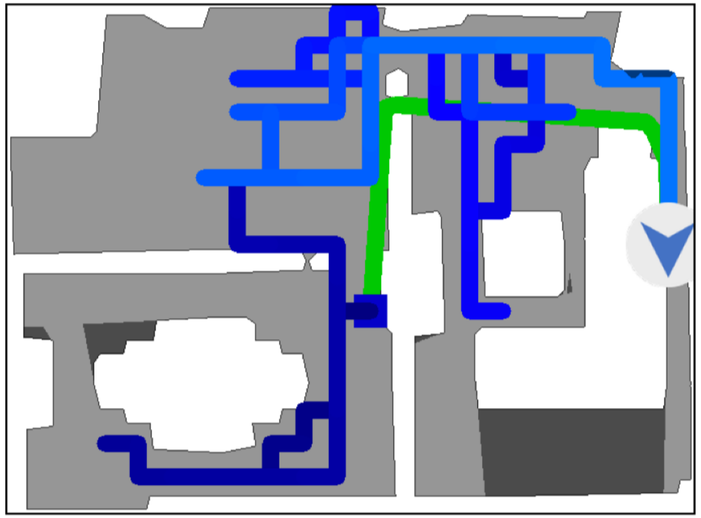
- Strongly outperforms all baselines and existing methods
- Generalizing to unheard sounds and unseen environments



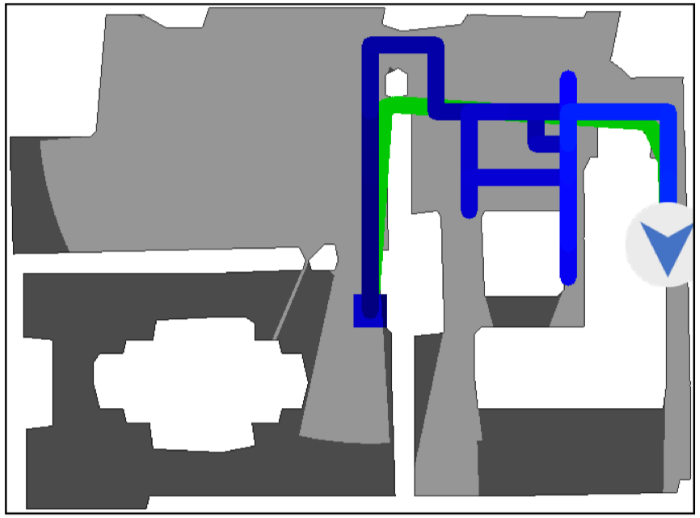
Changan Chen et al., Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021

Navigation trajectories

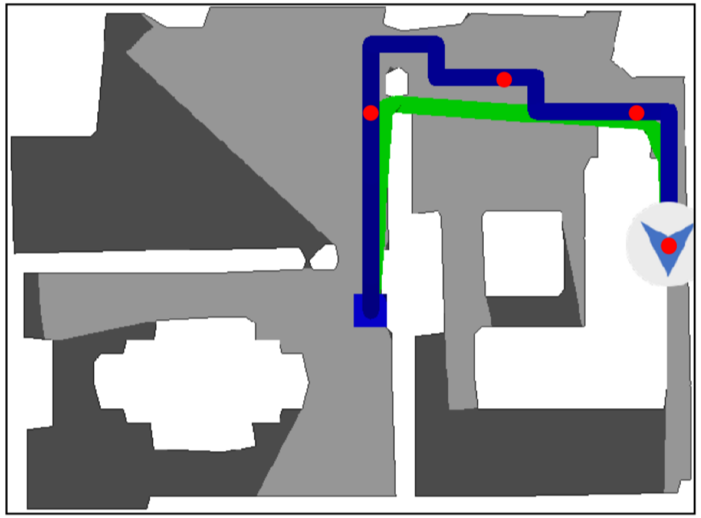
- Gan et al. [ICRA 20]: is prone to errors and often leads the agent to backtrack
- Chen et al. [ECCV20]: oscillates around obstacles
- AV-WaN (Ours): reaches the goal most efficiently



Gan et al. [ICRA20]



Chen et al. [ECCV20]



AV-WaN (Ours)

Agent Start Waypoint Shortest path Agent path Seen/Unseen area Occupied area

Changan Chen et al., Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021

Limitations of the AudioGoal task

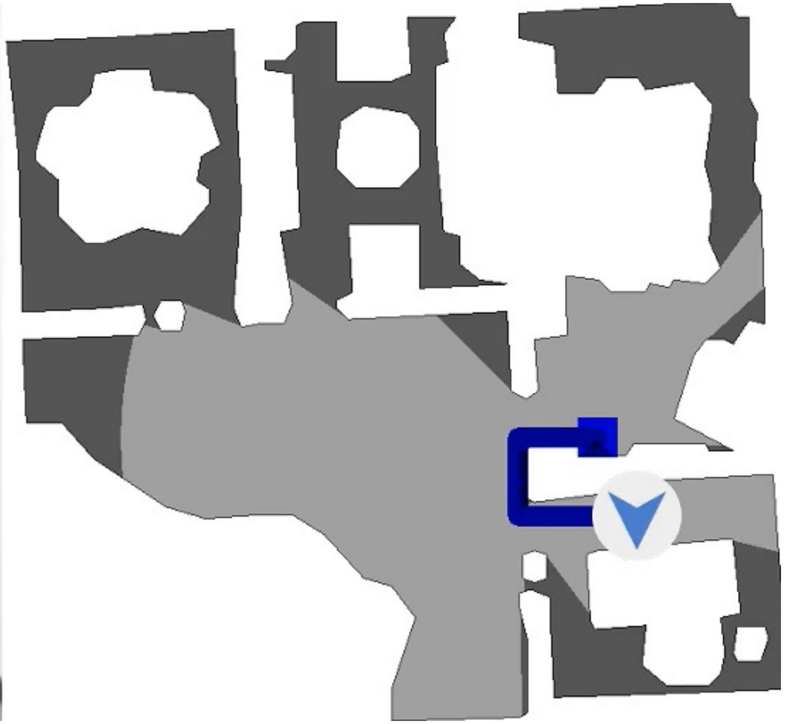
AudioGoal task (Chen et al. ECCV 2020, Gan et al. ICRA 2020):

- The sound is constant and periodic (it covers the whole episode)
- The goal has no visual embodiment

Agent's egocentric view

Top-down map

Telephone
not present!

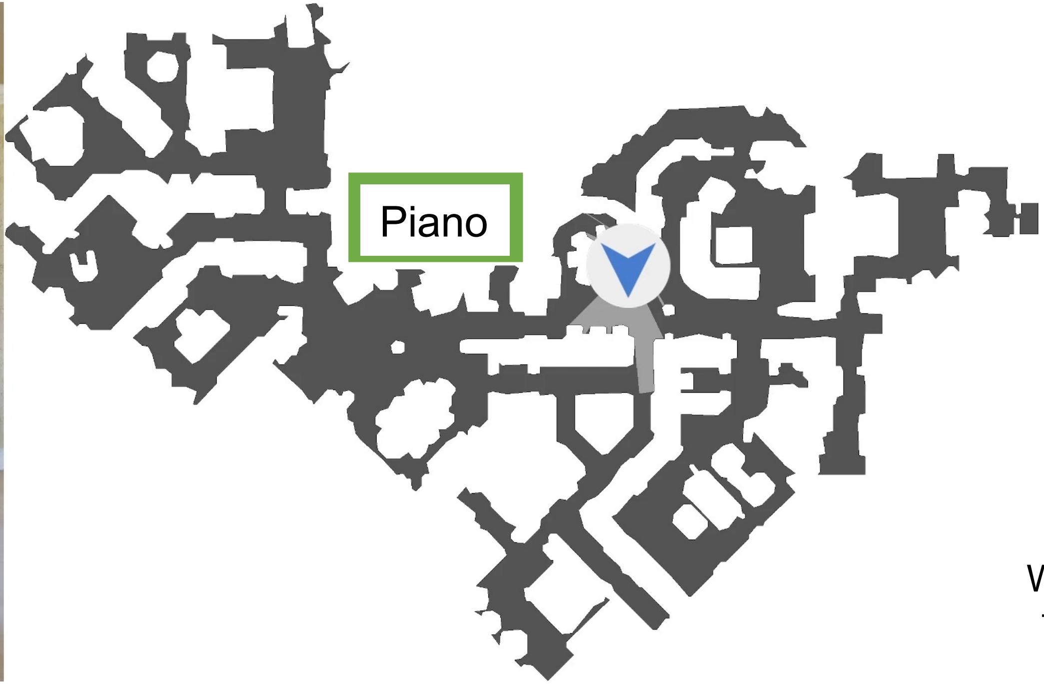


The agent searches for the ringing telephone in an unfamiliar environment

Semantic AudioGoal

Agent's egocentric view

Top-down map



Wear headphones
for spatial sound

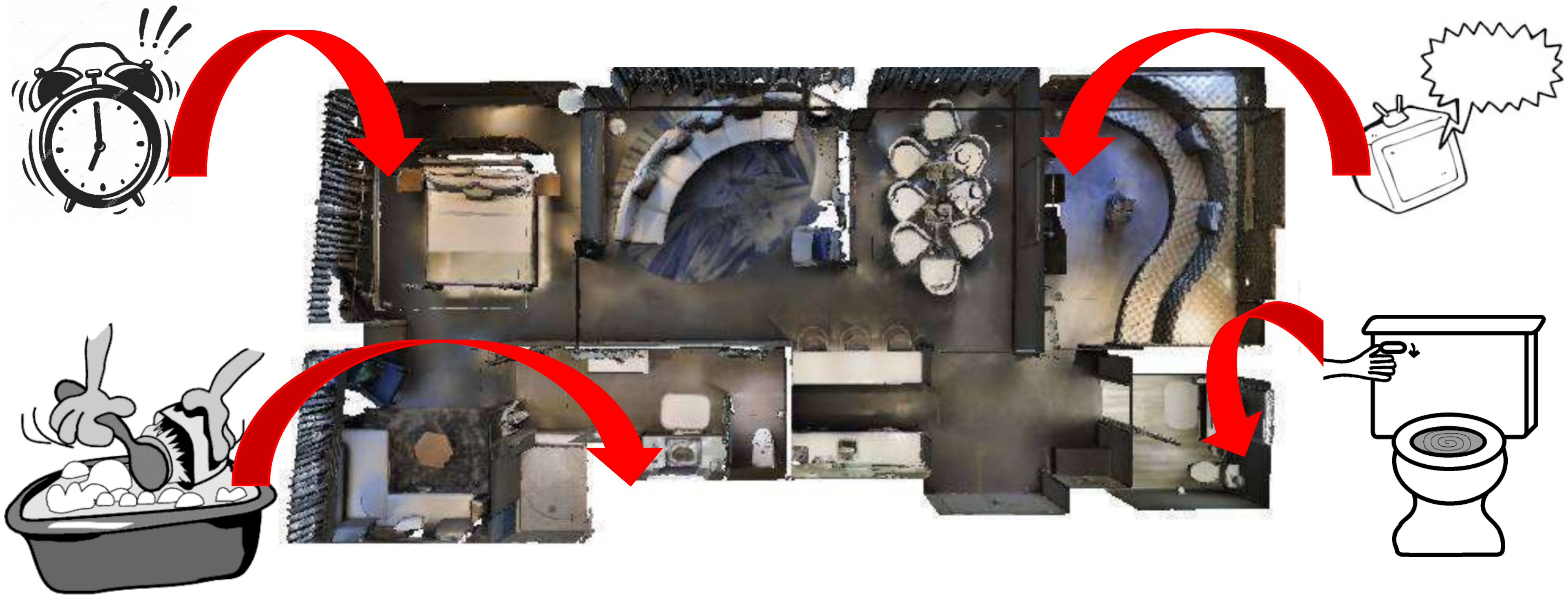
The agent must continue navigating even after the sound stops

Our proposed semantic AudioGoal task:
- The sound is associated with a semantically meaningful object
- The sound is not periodic and has variable length

Changan Chen et al., Semantic Audio-Visual Navigation, CVPR 2021

Semantic AudioGoal dataset

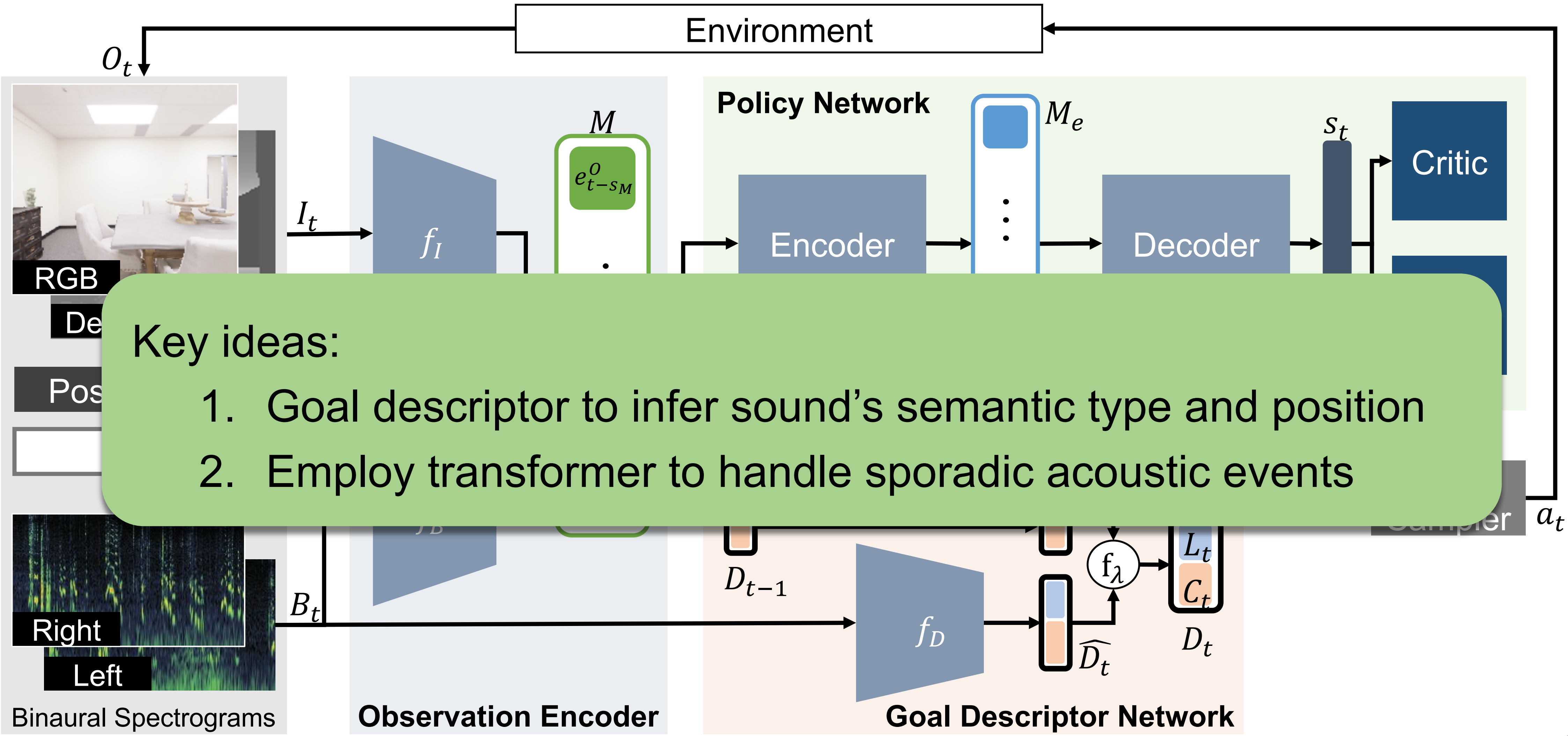
- Augment an existing simulator SoundSpaces[1] with semantic sounds
- 21 object categories in Matterport3D[2]: chair, TV, cabinet, sink etc.
- Object-emitted sounds and object-related sounds



[1]Changan Chen et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020
[2]Angle Chang et al., Matterport3D: Learning from RGB-D Data in Indoor Environments, 3DV 2017

Changan Chen et al., Semantic Audio-Visual Navigation, CVPR 2021

Semantic Audio-Visual Navigation (SAVi)



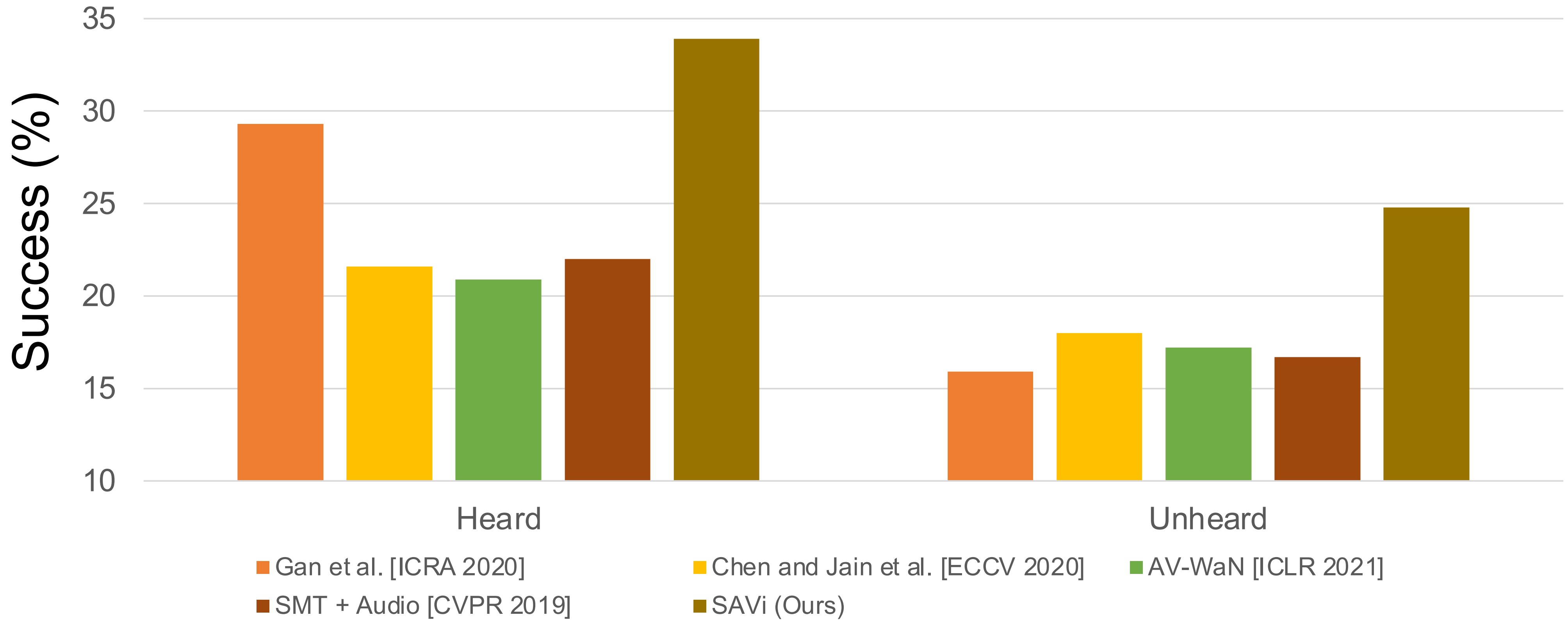
Key ideas:

1. Goal descriptor to infer sound's semantic type and position
2. Employ transformer to handle sporadic acoustic events

Changan Chen et al., Semantic Audio-Visual Navigation, CVPR 2021

Navigation results

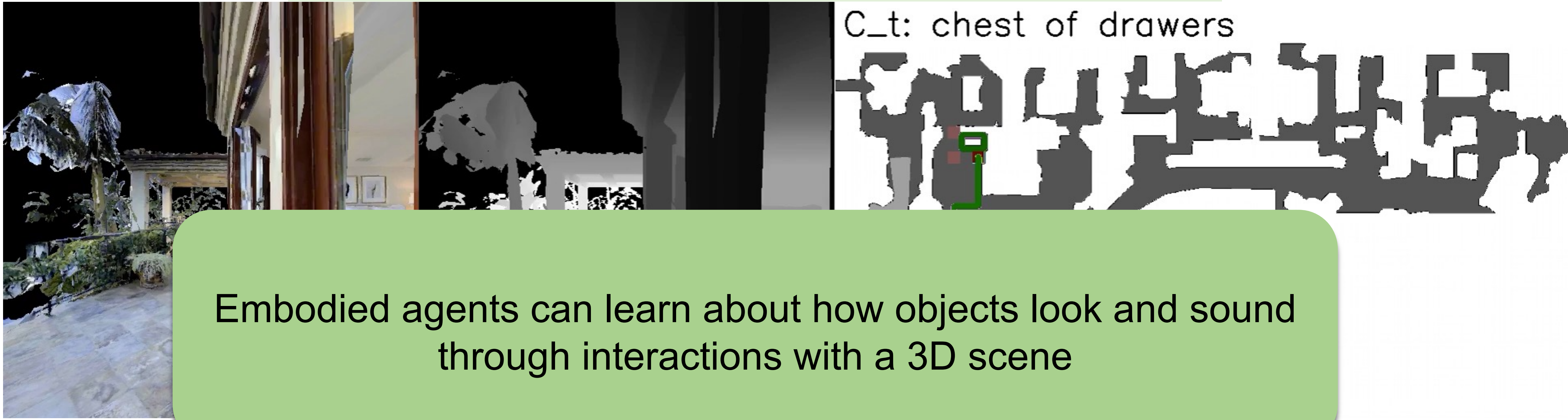
- SAVi strongly outperforms all existing methods
- Generalizing to unheard sounds



Changan Chen et al., Semantic Audio-Visual Navigation, CVPR 2021

Navigation example

Object: Chest of drawers **Sound:** Opening and closing a drawer



Embodied agents can learn about how objects look and sound through interactions with a 3D scene

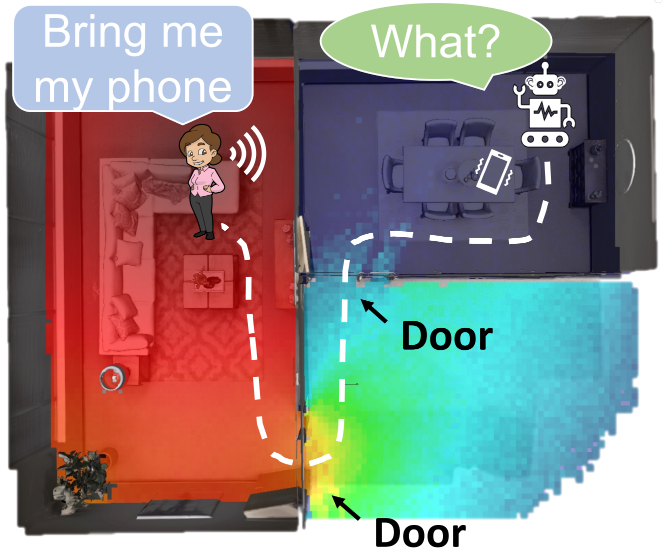
Agent | Object | Viewpoint | Start | Path w/ sound | Path w/o sound | Shortest path | Seen/Unseen | Occupied | L_t

The agent identifies it's drawer sliding sound and locates the target object with vision after the sound stops.

Changan Chen et al., Semantic Audio-Visual Navigation, CVPR 2021

Beyond navigation: recognition and synthesis

- Recognizing human speech in spaces is challenging due to reverberation
- Synthesizing sounds that are consistent with visual observations
- Requires studying perception separately from decision-making



Robotics

Home assistance

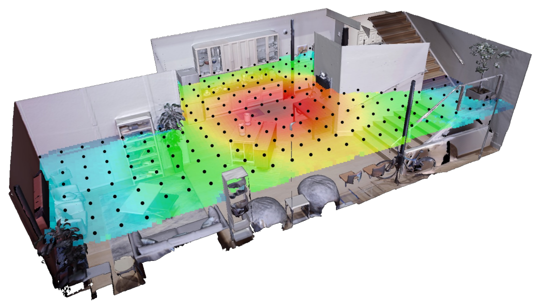
AR/VR

4D audio-visual perception

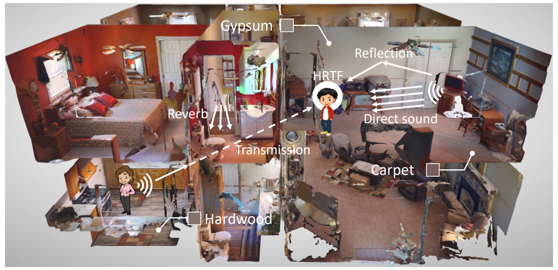
My research: learning the correspondence between sight and sound in spaces

Simulating sounds in spaces

SoundSpaces [ECCV20]

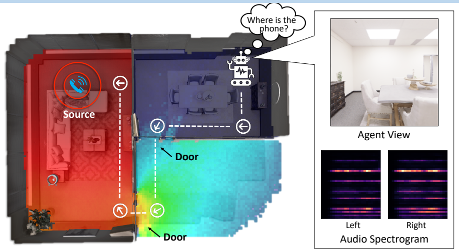


SoundSpaces 2.0 [NeurIPS22]

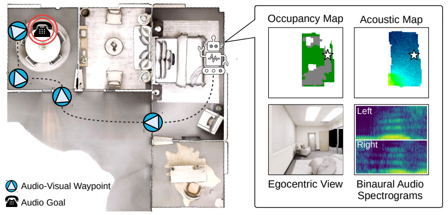


Navigating with sounds in spaces

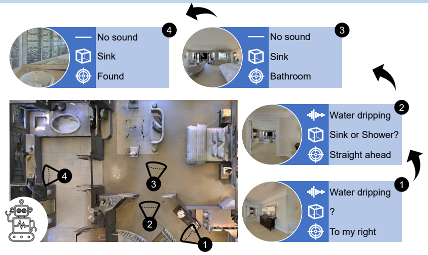
Audio-visual navigation SoundSpaces [ECCV20]



Efficient & hierarchical AV nav [ICLR21]

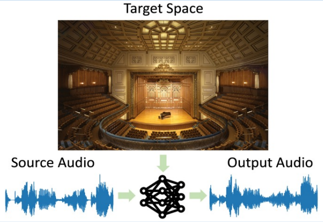


Semantic audio-visual navigation [CVPR21]

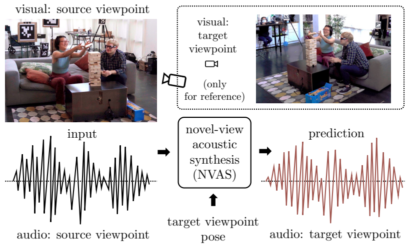


Synthesizing sounds in spaces

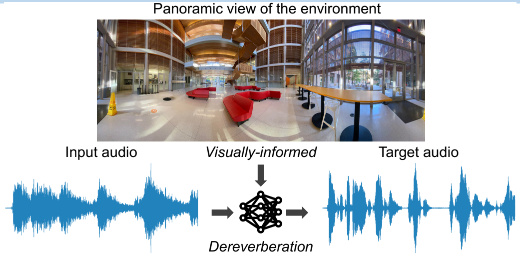
Visual acoustic matching [CVPR22]



Novel-view acoustic synthesis [CVPR23]



Audio-visual dereverberation [ICASSP23]



Matching acoustics

Can we alter the acoustic signature of the sound if we understand the acoustics of the space based on visuals?

Augmented reality

Film dubbing

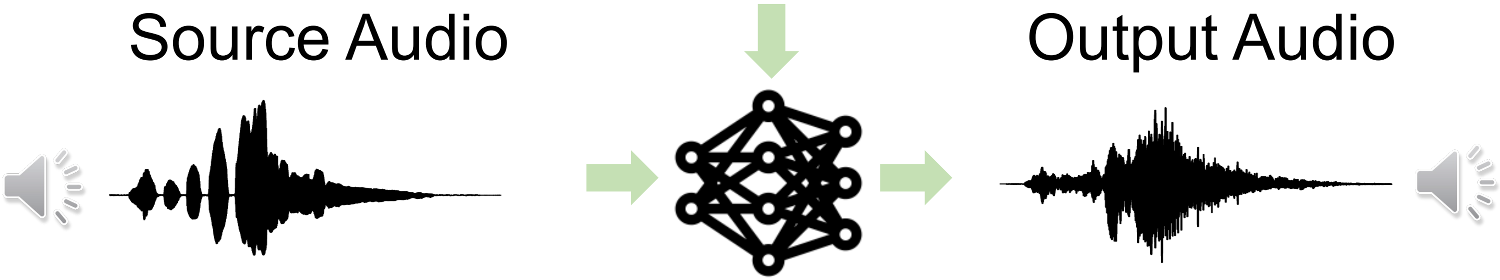
Video conferencing

Changan Chen et al., Visual Acoustic Matching, CVPR 2022

The visual acoustic matching task

We propose to transform the sound recorded in one space to another depicted in the target visual scene.

Target Space

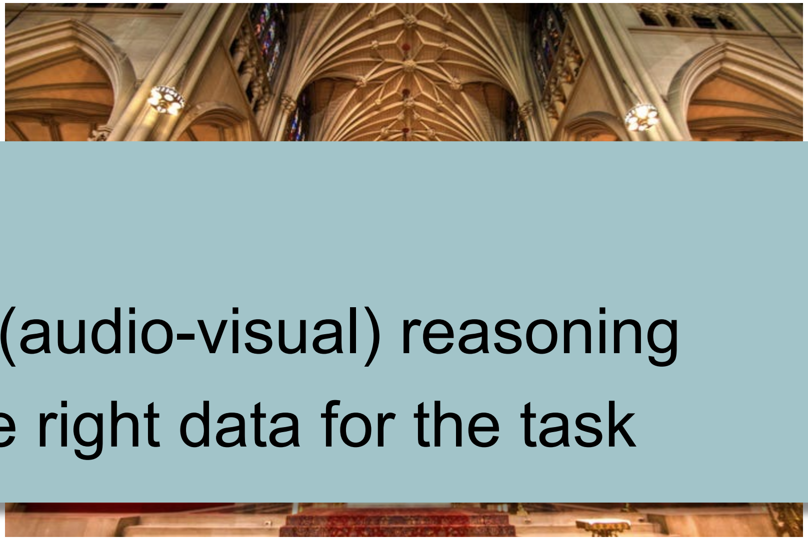


Changan Chen et al., Visual Acoustic Matching, CVPR 2022

The visual acoustic matching task

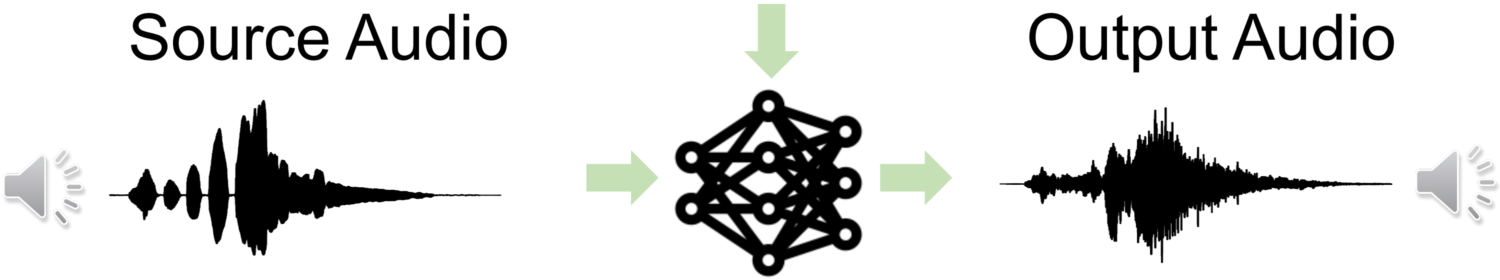
We propose to transform the sound recorded in one space to another depicted in the target visual scene.

Target Space



Main challenges:

1. Crossmodal (audio-visual) reasoning
2. Obtaining the right data for the task



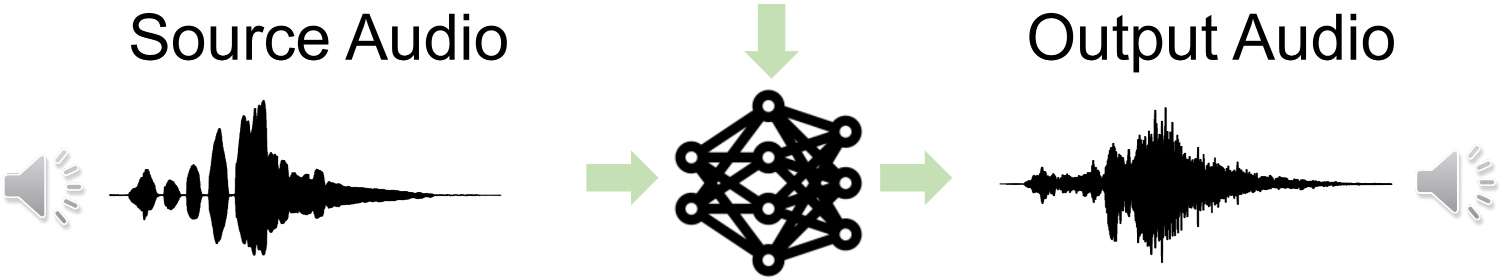
The visual acoustic matching task

We propose to transform the sound recorded in one space to another depicted in the target visual scene.

Target Space

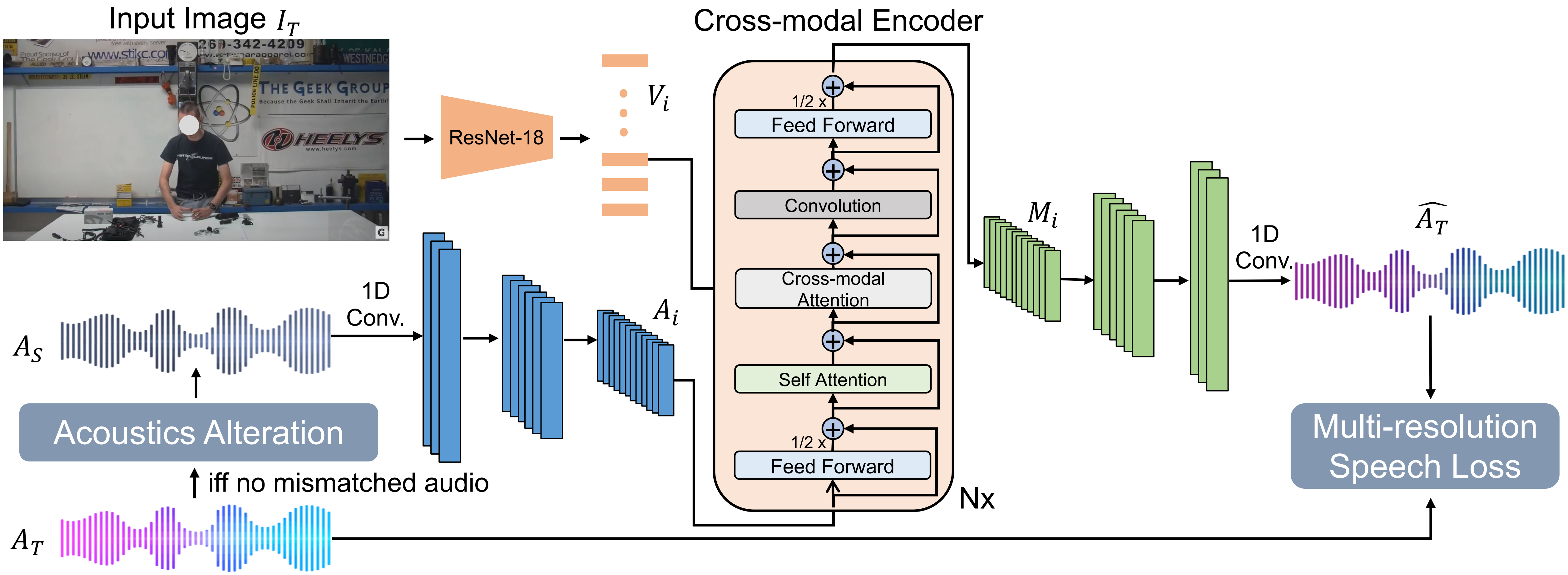
Key ideas:

1. Reasoning how image regions affect acoustics with attention
2. Leveraging Web videos with self-supervision for learning



Changan Chen et al., Visual Acoustic Matching, CVPR 2022

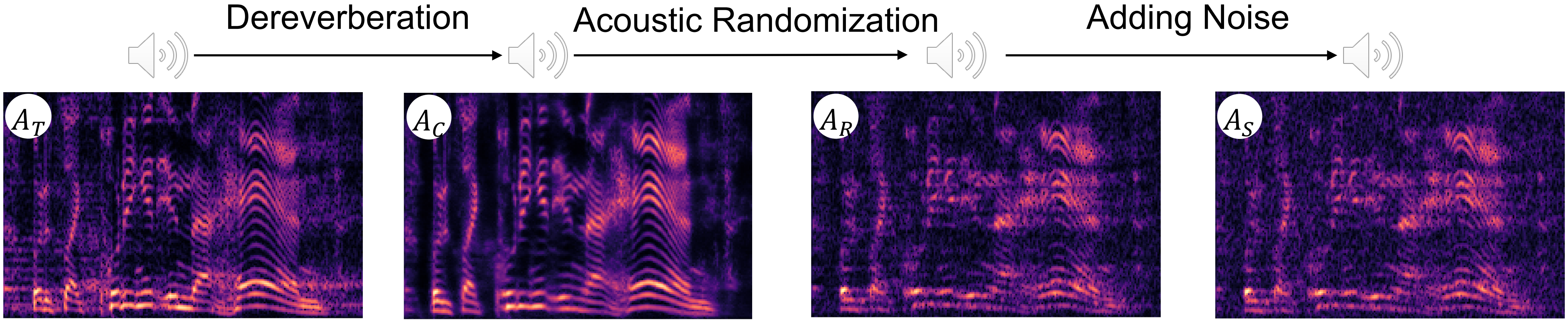
Audio-Visual Transformer for Audio Generation



Changan Chen et al., Visual Acoustic Matching, CVPR 2022

Acoustics alteration strategy

Goal: create audio with the same content but different acoustics as self-supervision.



Changan Chen et al., Visual Acoustic Matching, CVPR 2022

Experiment results

- Experiment on both synthetic and web video datasets
- Strongly outperforms traditional and heavily supervised approaches

	SoundSpaces-Speech						*Acoustic AVSpeech* [4]			
	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 [1]	1.338	0.044	0.312	-	-	-	-	-	-	-
Image2Reverb [2]	2.538	0.293	0.508	2.318	0.317	0.518	-	-	-	-
AV U-Net [3]	**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**

STFT: distance between mag spectrogram
RTE: errors of RT60 (time of reverb decaying by 60dB)
MOSE: errors of MOS (measures speech quality)

[1] More than 50 years of artificial reverberation, Vesa Valimaki, et al., DREAMS 2016
[2] Image2reverb: Cross-modal reverb impulse response synthesis, Nikhil Singh et al., ICCV 2021
[3] 2.5d visual sound, Ruohan Gao and Kristen Grauman, CVPR 2019
[4] Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation, Ariel Ephrat et al., SIGGRAPH 2018

Changan Chen et al., Visual Acoustic Matching, CVPR 2022

Examples on SoundSpaces-Speech

In this example, we show comparison of our model with baselines on SoundSpaces-Speech (unseen).

Anechoic GT Target AViTAR Image2Reverb[1] AV U-Net [2]

[1] Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis, Singh et al., ICCV 2021
[2] 2.5D Visual Sound, Gao et al., CVPR 2019

Matching different environments on AVSpeech

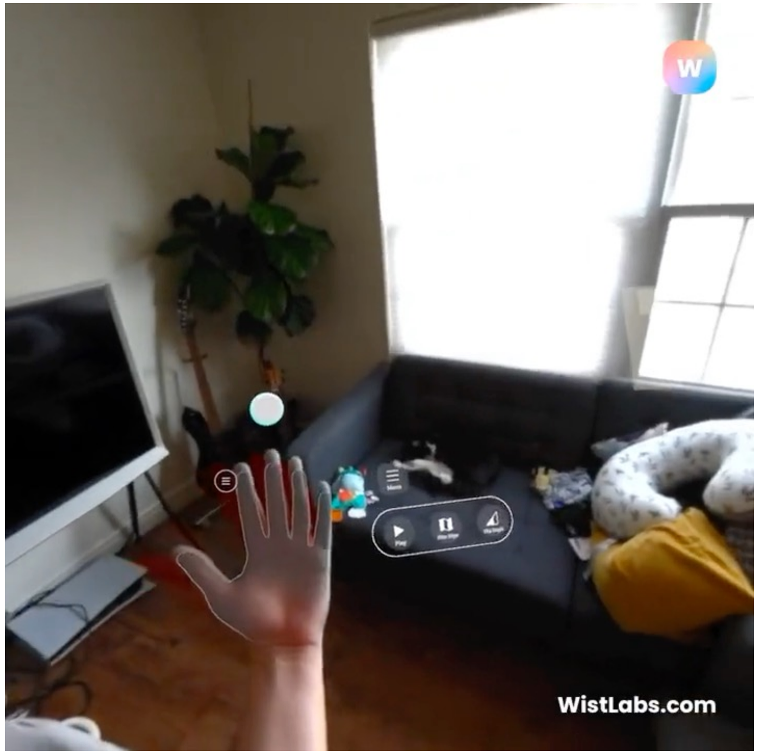
	Office	Garage	Auditorium
Input			
AViTAR			
RT60	0.34s	0.40s	0.58s

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

Changan Chen et al., Visual Acoustic Matching, CVPR 2022

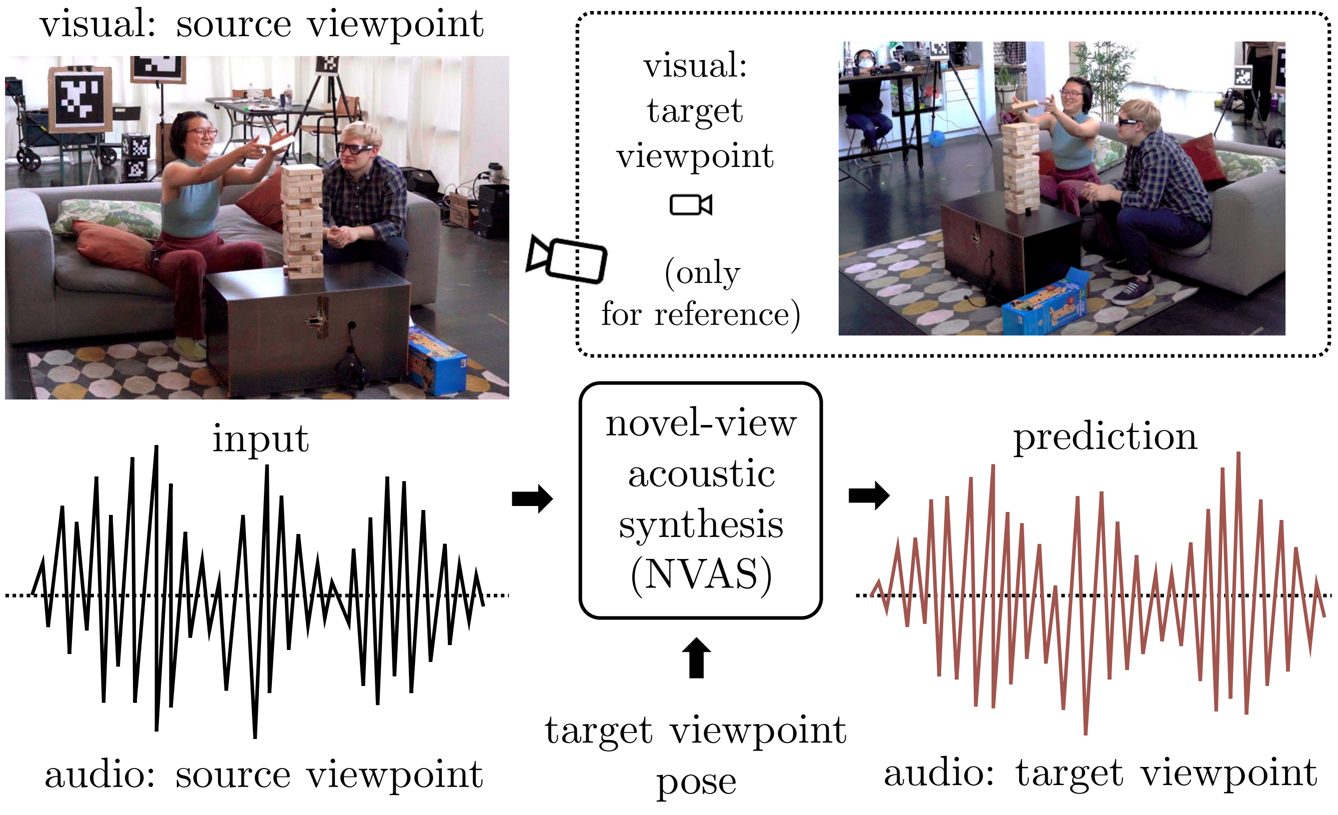
Can we synthesize fine-grained acoustics?

- Many of our important life moments are recorded in videos
- Videos are however passively collected from one viewpoint
- Recreating the moment in 3D is important for immersive AR/VR applications
- Novel-view synthesis (NVS) is vision-only and does not handle sound



Novel-view acoustic synthesis

We propose the novel-view acoustic synthesis task:



Changan Chen et al., Novel-View Acoustic Synthesis, CVPR 2023

Difference between NVS & NVAS

Novel-view synthesis (NVS):

- 3D scenes change limitedly during the recording
- Camera captu... directional ma...
- Frequency of ... providing spa... triangulation and segmentation

Novel-view acoustic synthesis (NVAS):

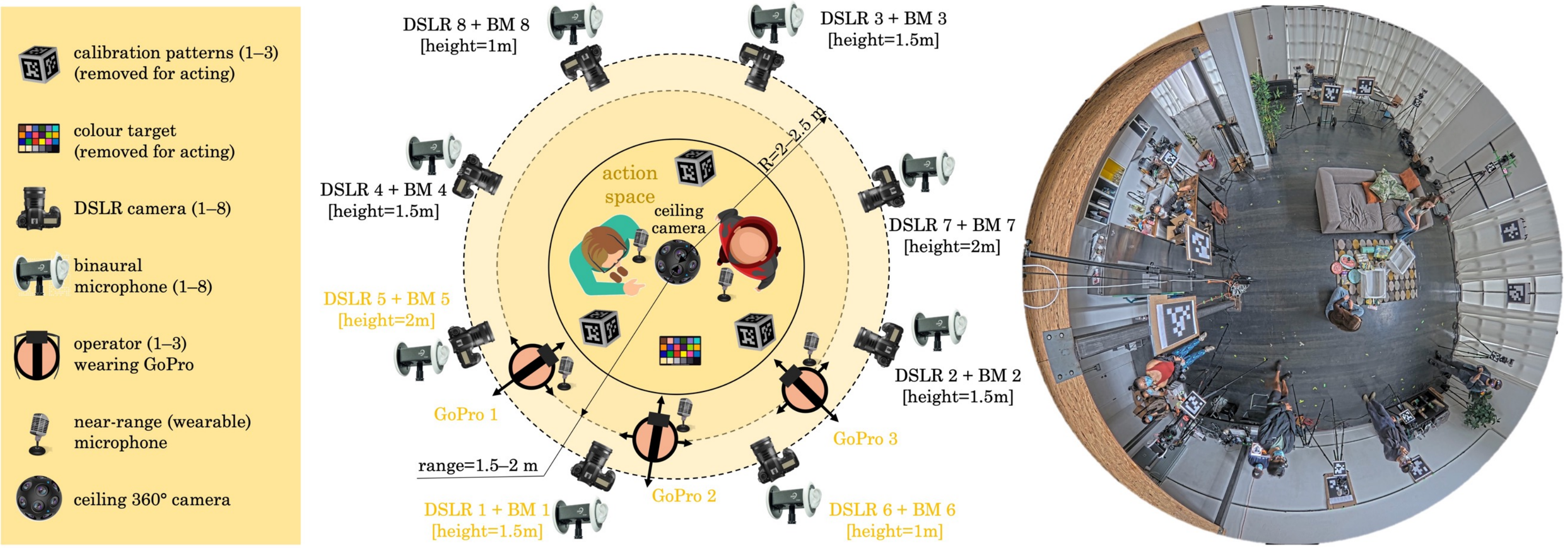
- Sound changes substantially over time
- ...t best weakly
- ... wide range of
- Sounds are often mixed together

1. Lack of supporting dataset and benchmark
2. Lack of existing model that is capable of NVAS

Changan Chen et al., Novel-View Acoustic Synthesis, CVPR 2023

Replay-NVAS dataset

- 68 scenes of social interactions, 2-4 actors per scene
- 8 surrounding viewpoints, equipped with DSLR cameras and binaural mics
- Each actor has a near-range mic to record their voice
- Over 50 hours of video data



Changan Chen et al., Novel-View Acoustic Synthesis, CVPR 2023

Replay-NVAS example

Changan Chen et al., Novel-View Acoustic Synthesis, CVPR 2023

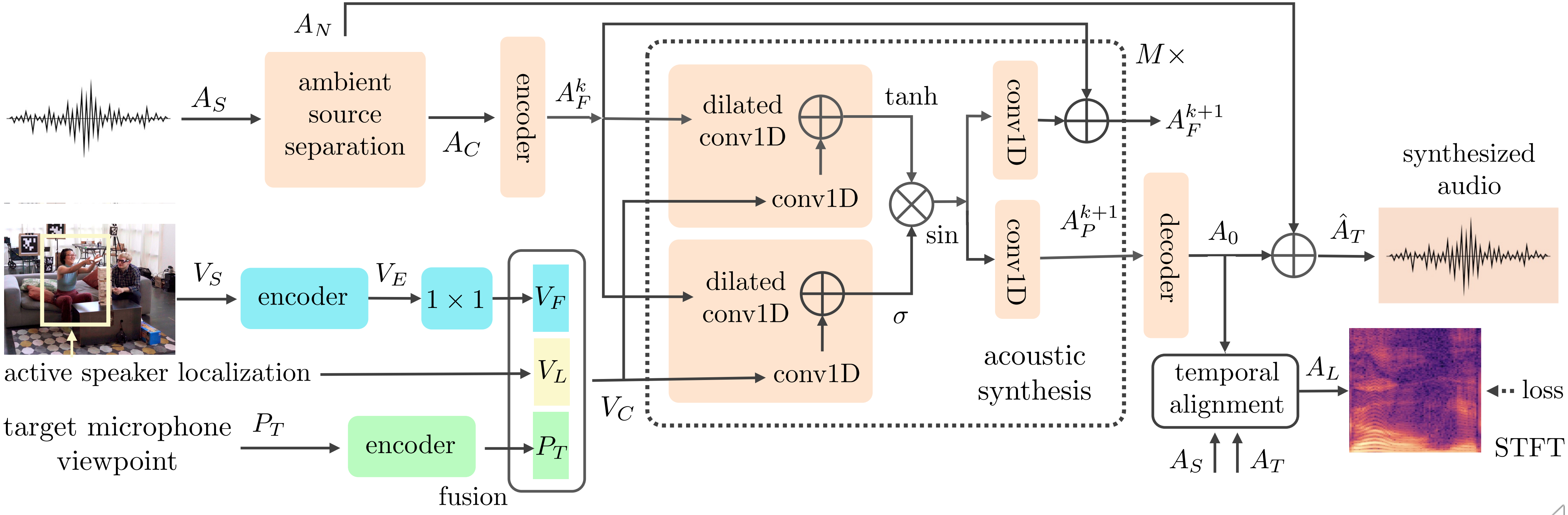
SoundSpaces-NVAS dataset

- Constructed based on SoundSpaces 2.0[1] audio-visual simulator
- Renders acoustic effects such as direct sound, reverberation, transmission, and diffraction
- Use LibriSpeech[2] (audio book) as the source audio
- 1,000 speakers, 120 3D scenes, 200K viewpoints and 1.3K hours of audio-visual data

[1]SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Chen et al., NeurIPS 2022
[2]Librispeech: An ASR corpus based on public domain audio books, Chen et al., ICASSP 2015

Visually Guided Acoustic Synthesis (ViGAS)

Learn an **implicit neural transfer function** that reasons the sound source location, acoustics of the space and the target pose in 3D to synthesize the target sound.



Changan Chen et al., Novel-View Acoustic Synthesis, CVPR 2023

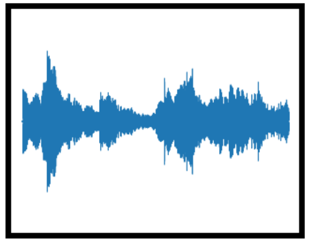
Results

- Experiment on both single environment and novel environment
- Outperforms traditional approaches and audio-only ablation
- Generalizing to novel environment with single view is non-trivial

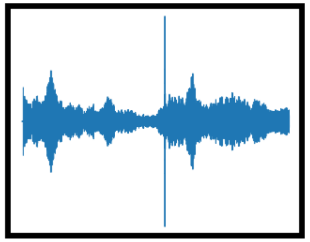
	SoundSpaces-NVAS						**Replay-NVAS**		
	Single Environment			*Novel Environment*			*Single Environment*		
	Mag	LRE	RTE	Mag	LRE	RTE	Mag	LRE	RTE
Input audio	0.225	1.473	0.032	0.216	1.408	0.039	0.159	1.477	0.046
TF Estimator [1]	0.359	2.596	0.059	0.440	3.261	0.092	0.327	2.861	0.147
DSP [2]	0.302	3.644	0.044	0.300	3.689	0.047	0.463	1.300	0.067
VAM [3]	0.220	1.198	0.041	0.235	1.131	0.051	0.161	0.924	0.070
ViGAS w/o visual	0.173	0.973	0.031	0.181	1.007	0.036	0.146	0.877	**0.046**
ViGAS	**0.159**	**0.782**	**0.029**	**0.175**	**0.971**	**0.034**	**0.142**	**0.716**	0.048

[1] Extrapolation, interpolation, and smoothing of stationary time series. Norbert Wiener. Report of the Services 19, 1942
[2] Introduction to head-related transfer functions (hrtfs): representations of hrtfs in time, frequency, and space. Cheng et al., AES 200
[3] Visual Acoustic Matching, Chen et al., CVPR 2022

Replay-NVAS example 1

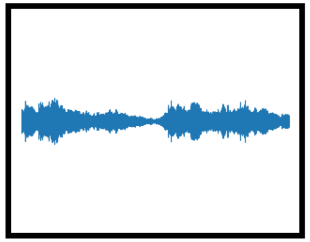


Left channel

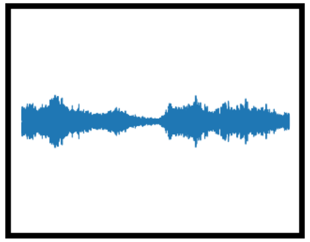


Right channel

Source

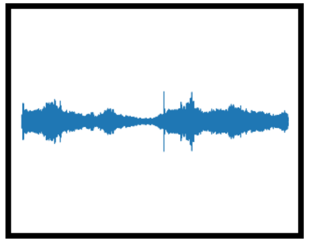


Left channel

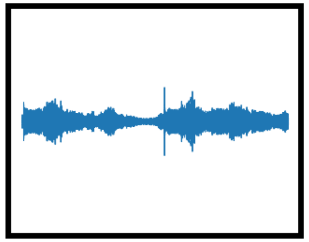


Right channel

Target



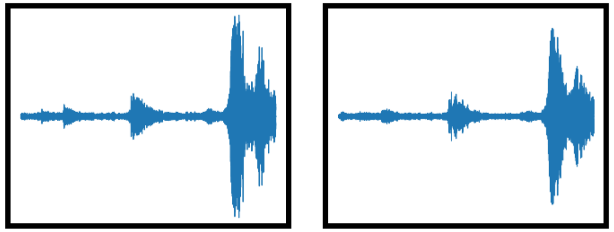
Left channel



Right channel

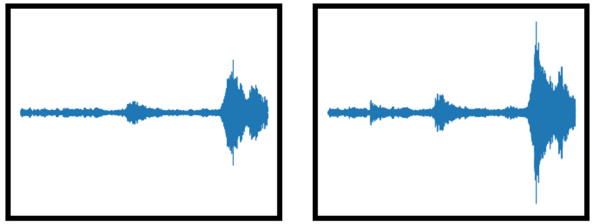
Ours

Replay-NVAS example 2



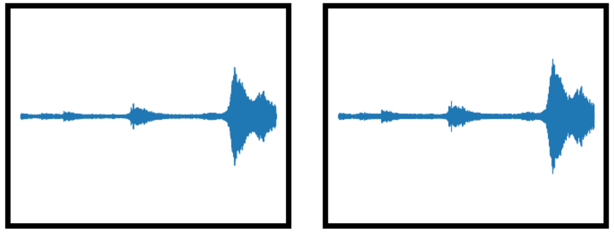
Left channel Right channel

Source



Left channel Right channel

Target

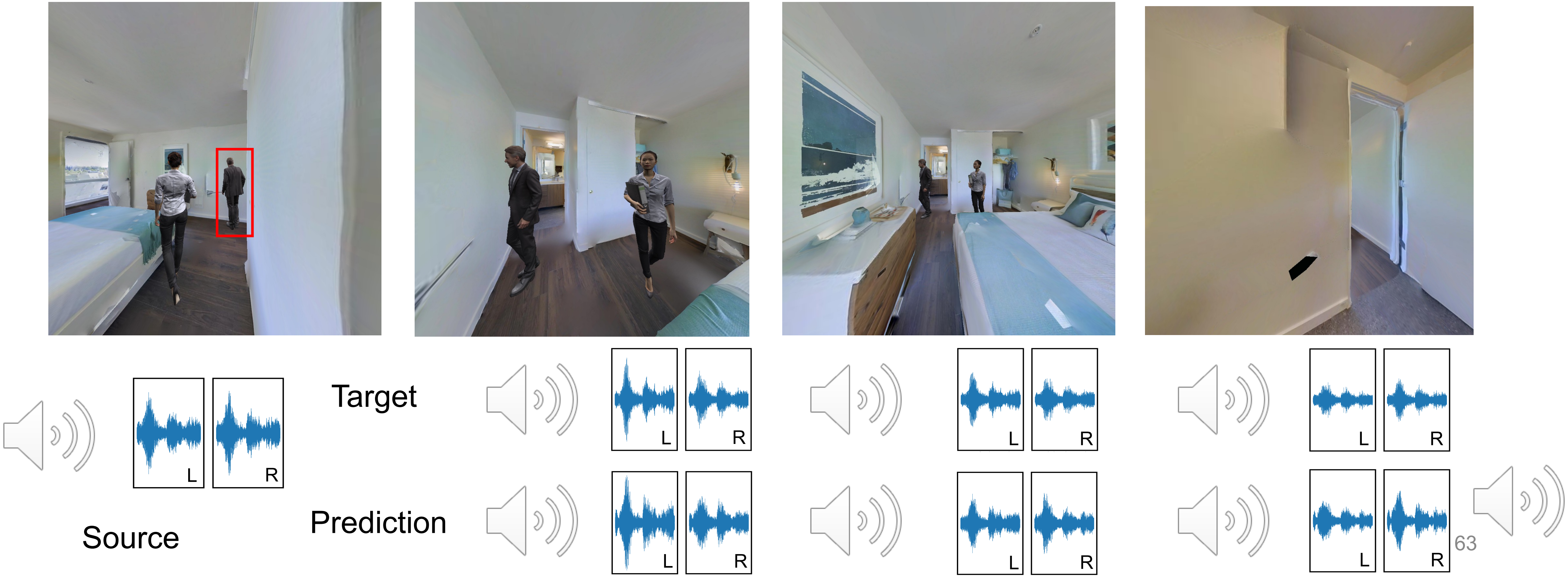


Left channel Right channel

Ours

Qualitative examples on SoundSpaces-NVAS

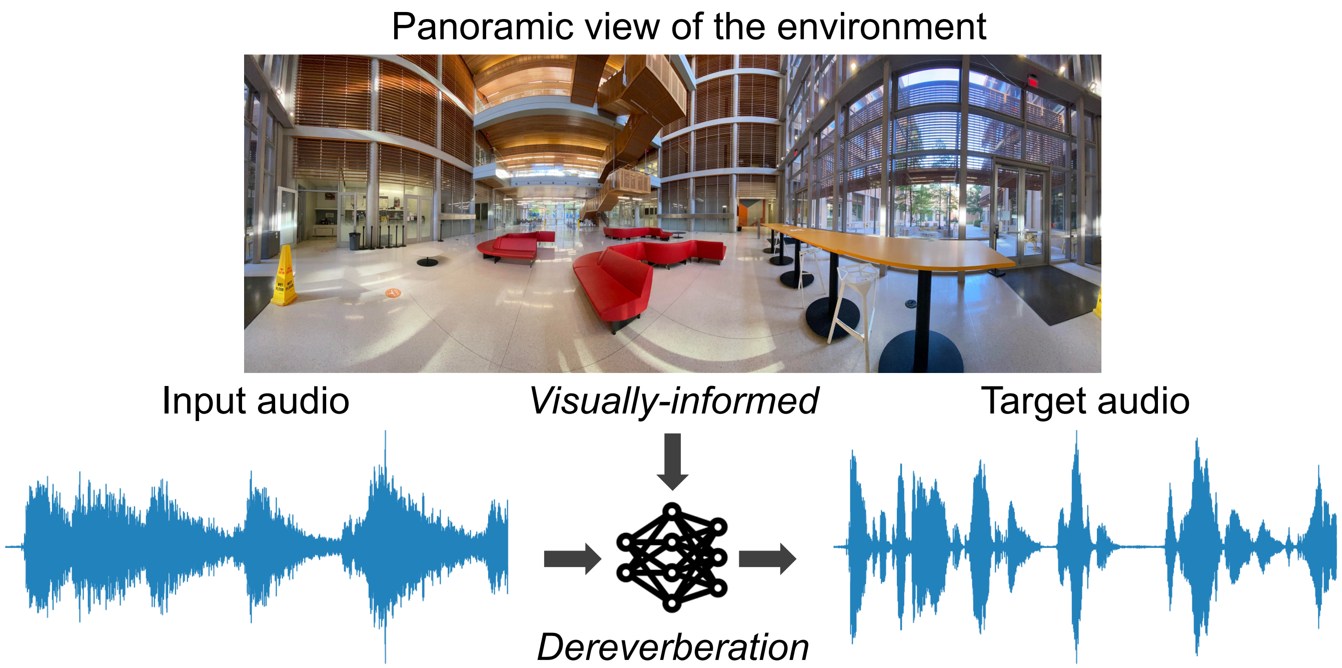
Here we show that for one source viewpoint, our model predicts the audio for four different viewpoints.



Audio-visual dereverberation

Can we strip away reverberation with visual cues?

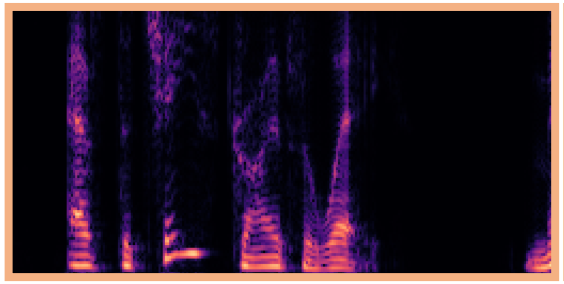
- We propose the audio-visual dereverberation task
- Model dereverberates better with visual information
- Demonstrates on several downstream tasks

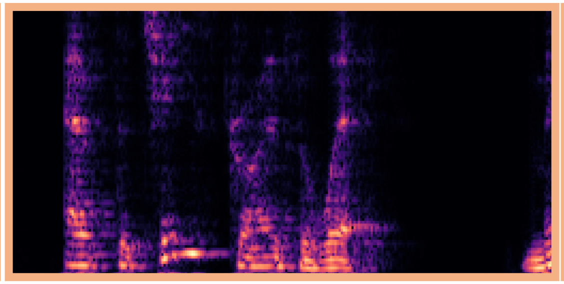


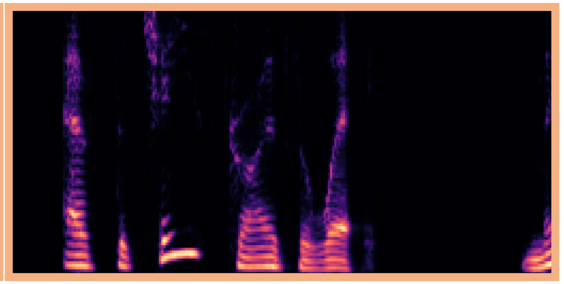
Changan Chen et al., Learning Audi-Visual Dereverberation, ICASSP 2023

Qualitative examples

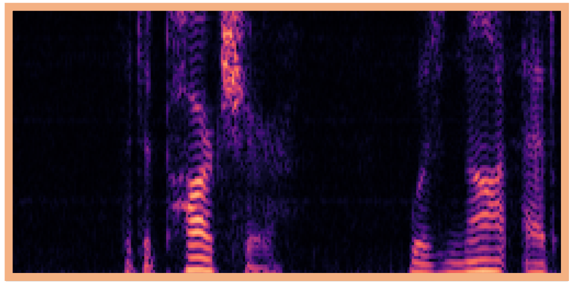
Panorama RGB | Clean (GT) | Reverberant | De-reverberated

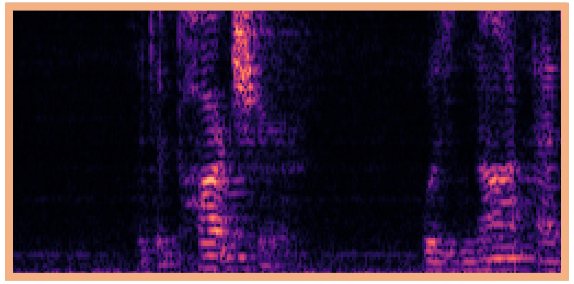


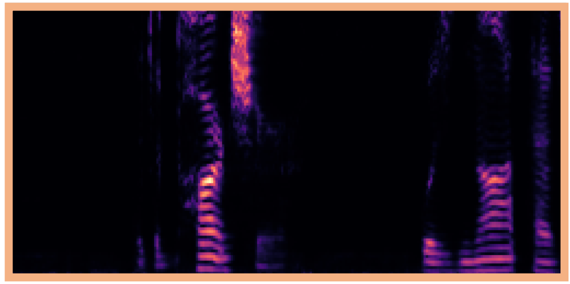




Long corridor, distance speaker



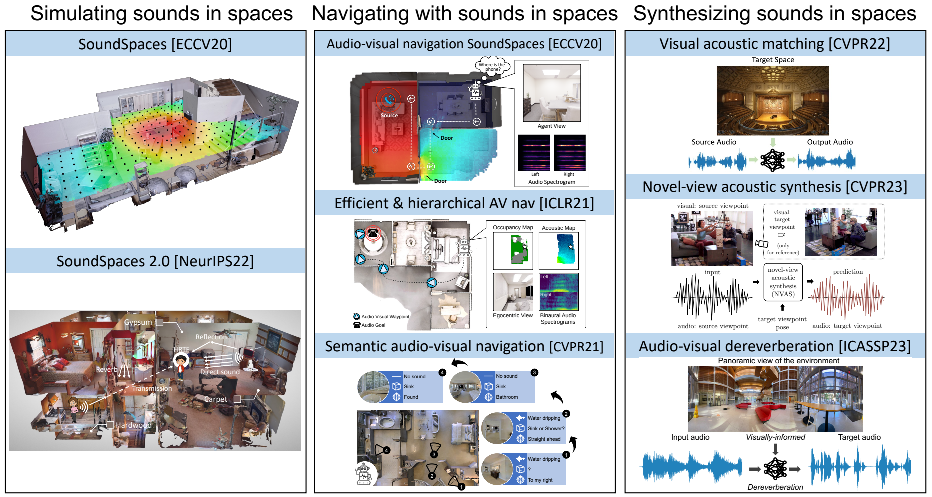




Classroom, close speaker

Changan Chen et al., Learning Audi-Visual Dereverberation, ICASSP 2023

Summary



Simulator & Datasets

- SoundSpaces 1.0 & 2.0
- SoundSpaces derived
- Multi-view AV datasets

Tasks

- Audio-visual embodied AI
- Visual-acoustic learning
- Multimodal NVS

Algorithms

- Multimodal navigation policies
- Self-supervision for VAM
- Multimodal fusion & generation

Thank you!