Audio-Visual Embodied AI: From Simulating to Navigating with Sounds in Spaces

Changan Chen
changan.io
UT Austin
10/03/2023

TEXAS
The University of Texas at Austin

Human perception is multisensory

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

Autonomous agents

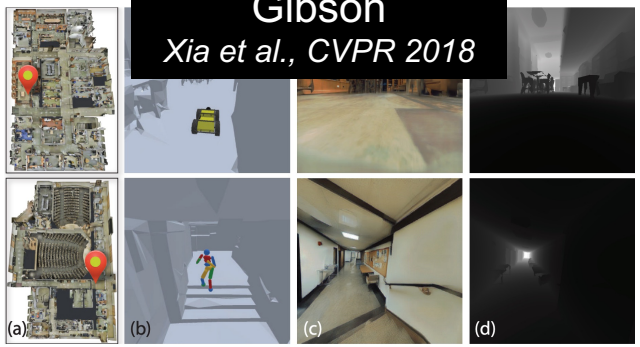
Home assistance robot

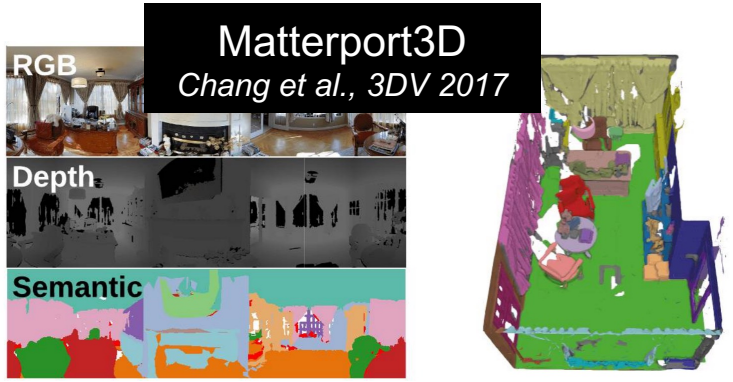
Rescue robot

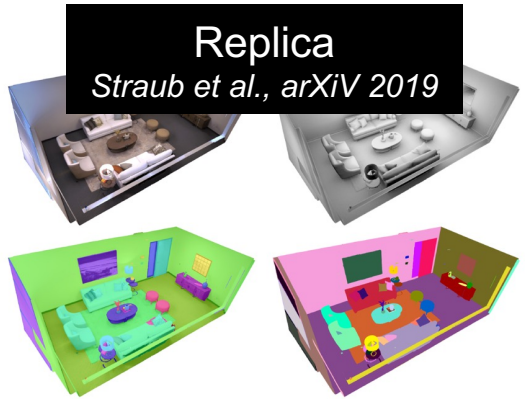
Robots that can **see, hear and act** in the environment, e.g., reacting to humans' commands or locating fire-alarm going off

Simulating embodiment in 3D scenes

Datasets







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 hear and react in the environment

Vision-Only

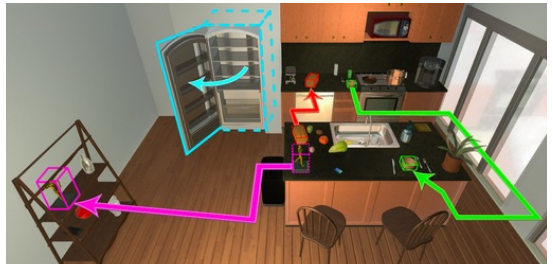
Gupta et al., 2017
Zhu et al., 2017
Sava et al., 2019
…

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 al all, 2019
…



Vision-Audio

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

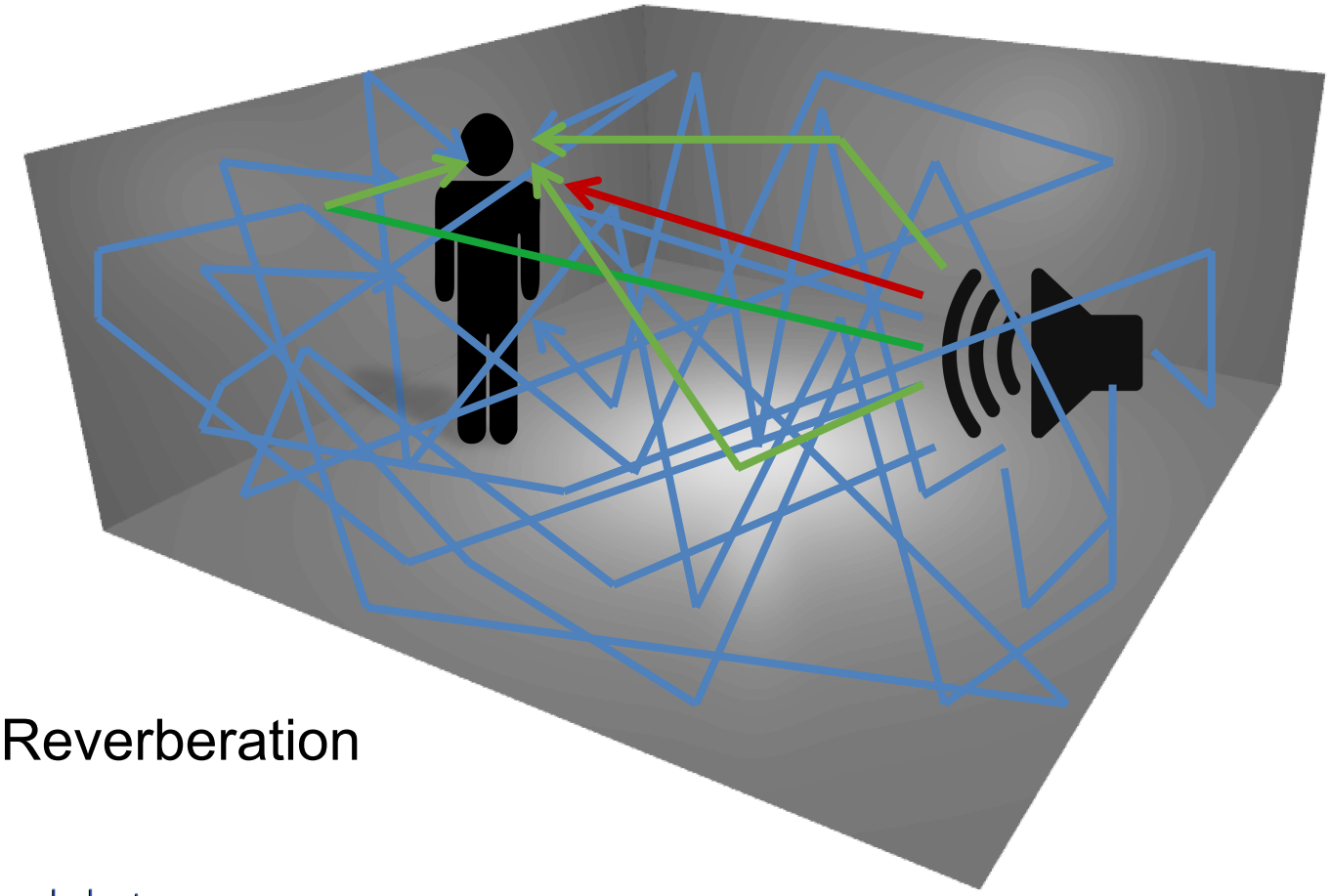
- No existing simulation supports audio-visual rendering
- No existing formulation for audio-visual navigation

SoundSpaces demo

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

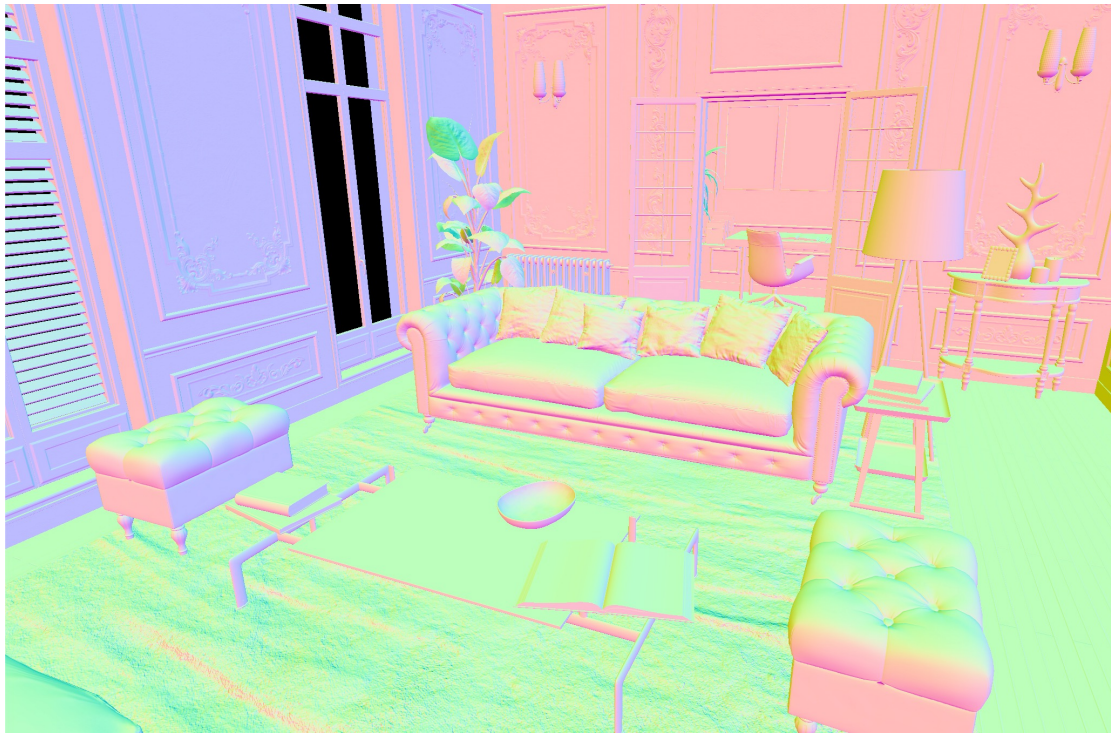
Background: acoustic simulation

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

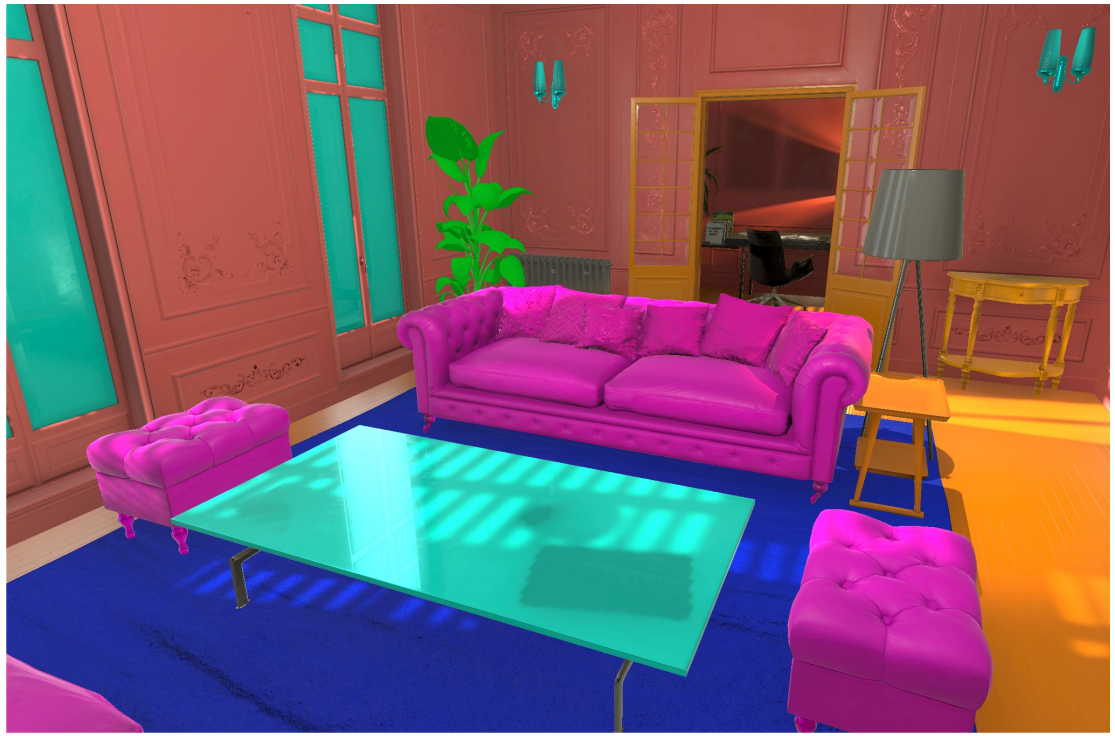


Physics-based audio rendering

3D Geometry



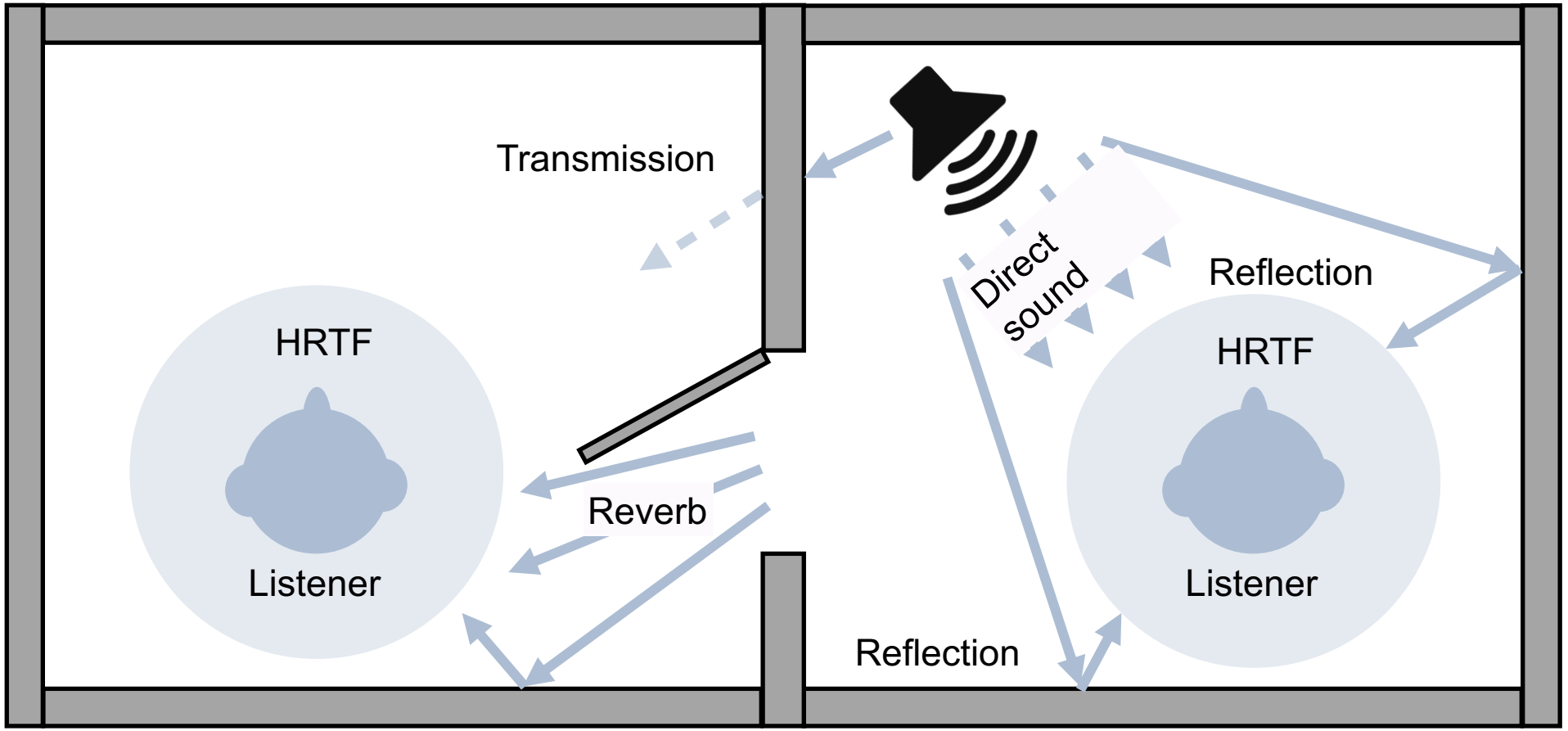
Material Properties



Simulate the sound received by the listener from a source location

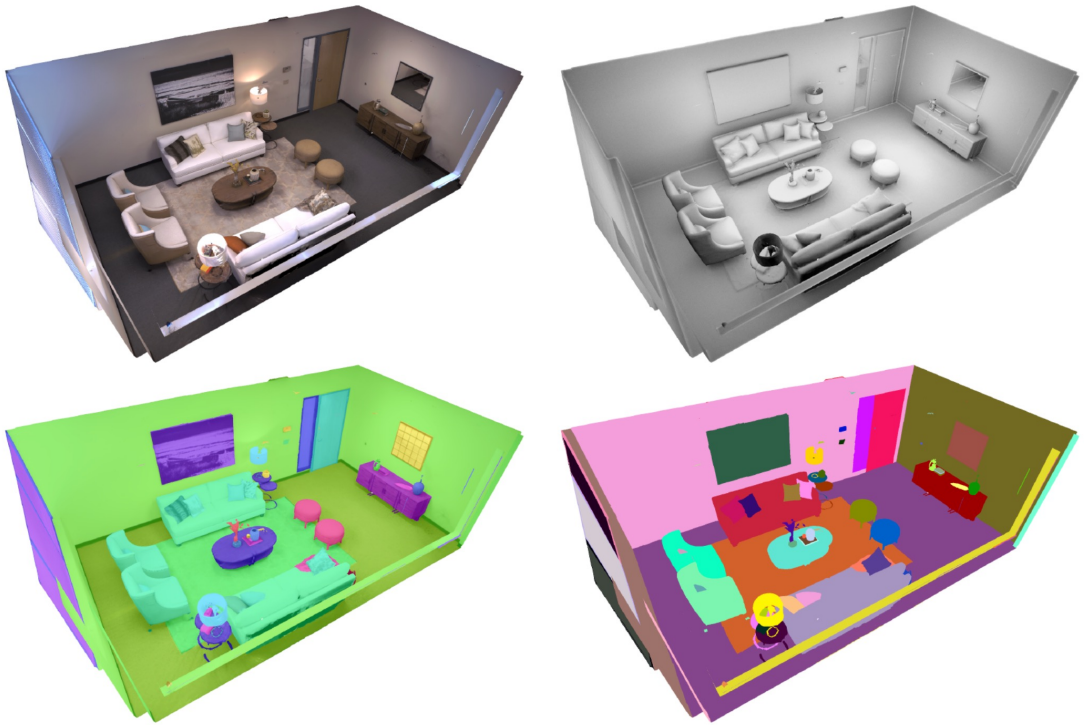
Sound propagation system

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

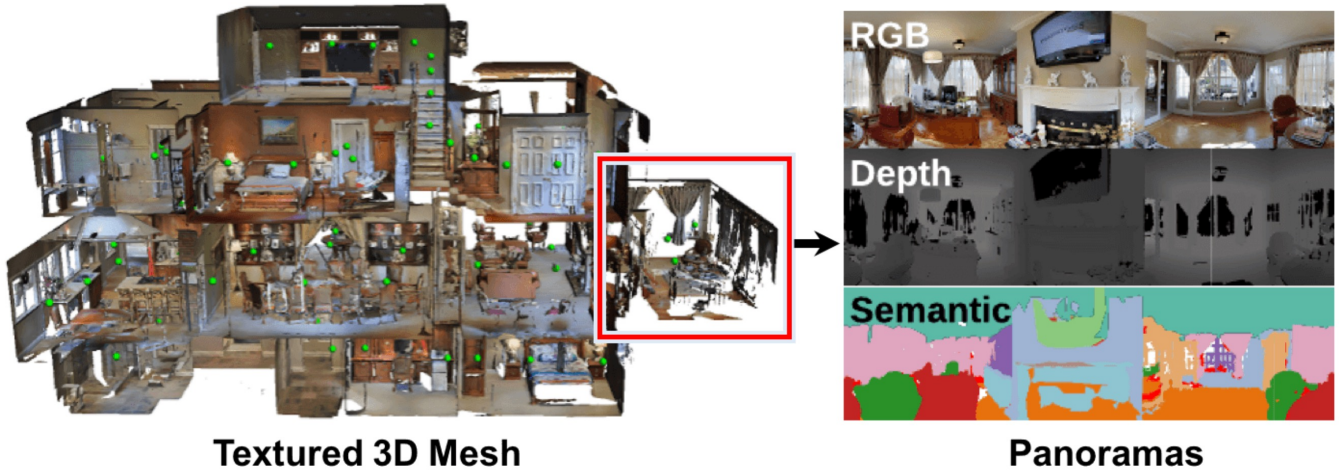


Real-scan environments

Replica[1] dataset



Matterport3D[2] dataset

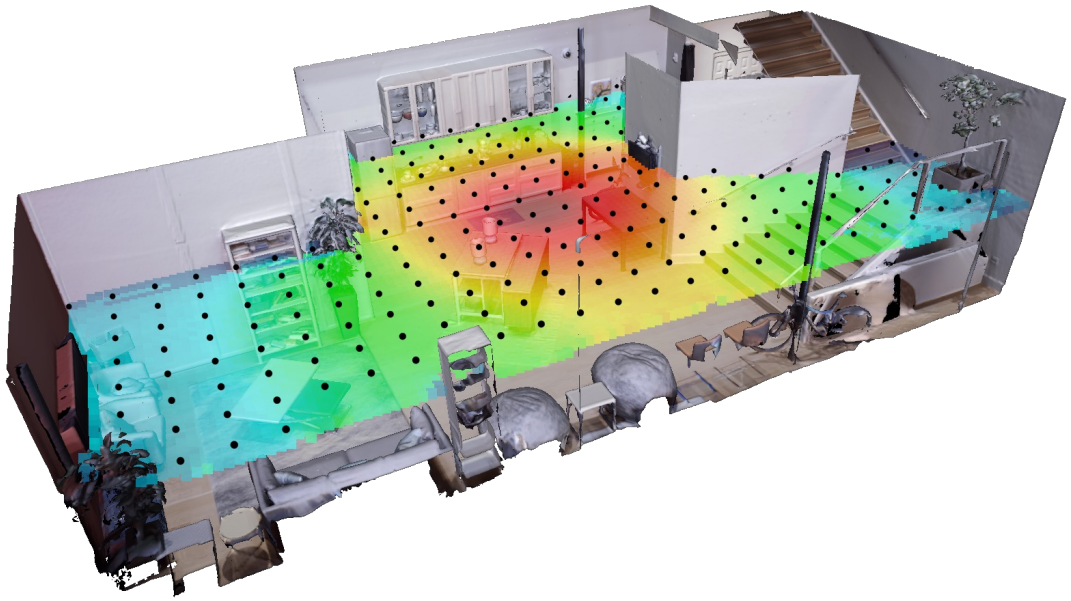


[1]The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019
[2]Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

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

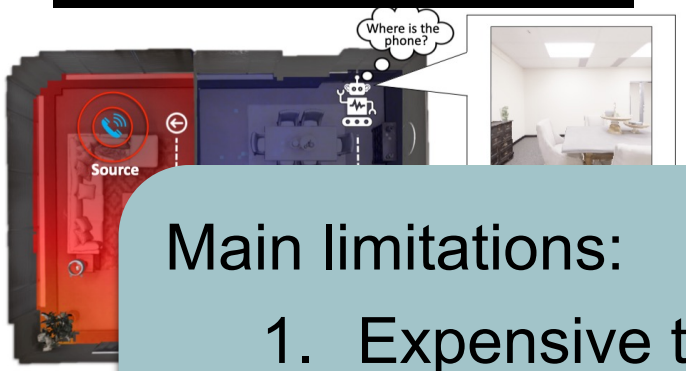
Table: Summary of dataset statistics

Visit soundspaces.org for more information!

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

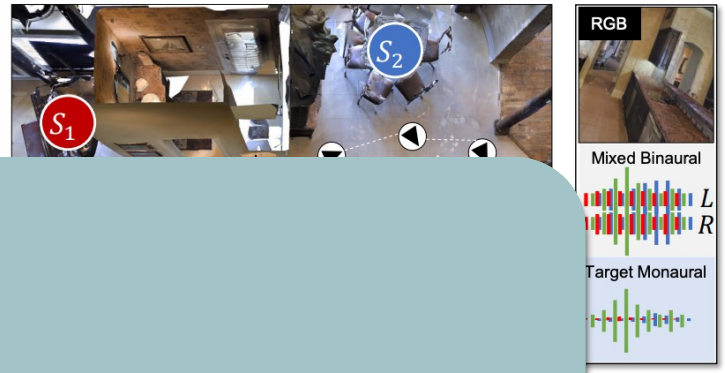
Enabling audio-visual embodied AI and beyond

Audio-Visual Navigation
Chen et al., ECCV 2020



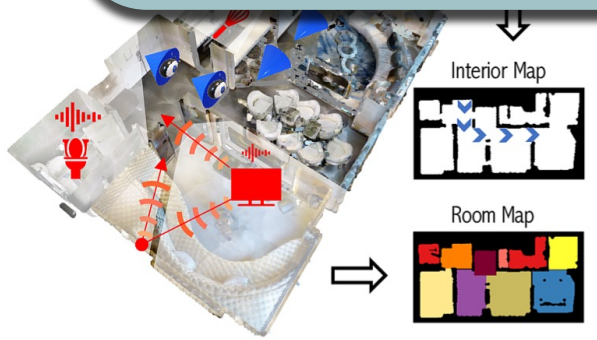
Echolocation Learning
Gao et al., ECCV 2020

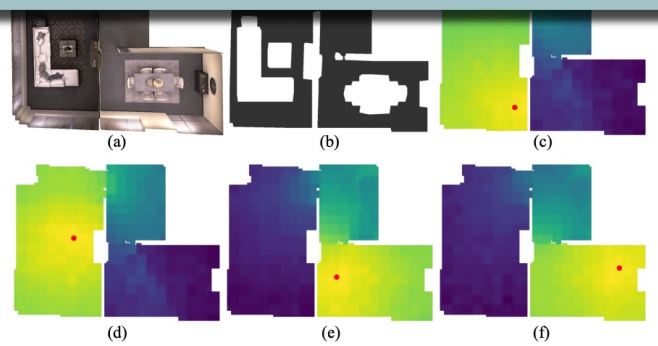
Audio-Visual Separation
Majumder et al., ICCV 2021

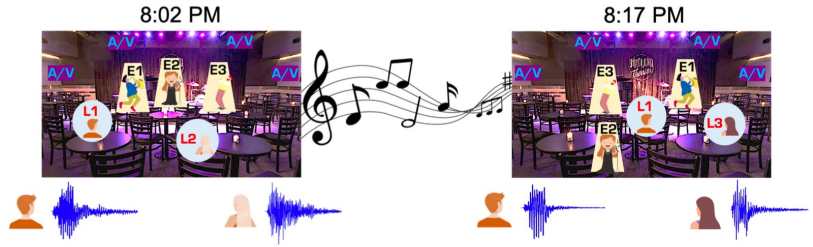


Main limitations:

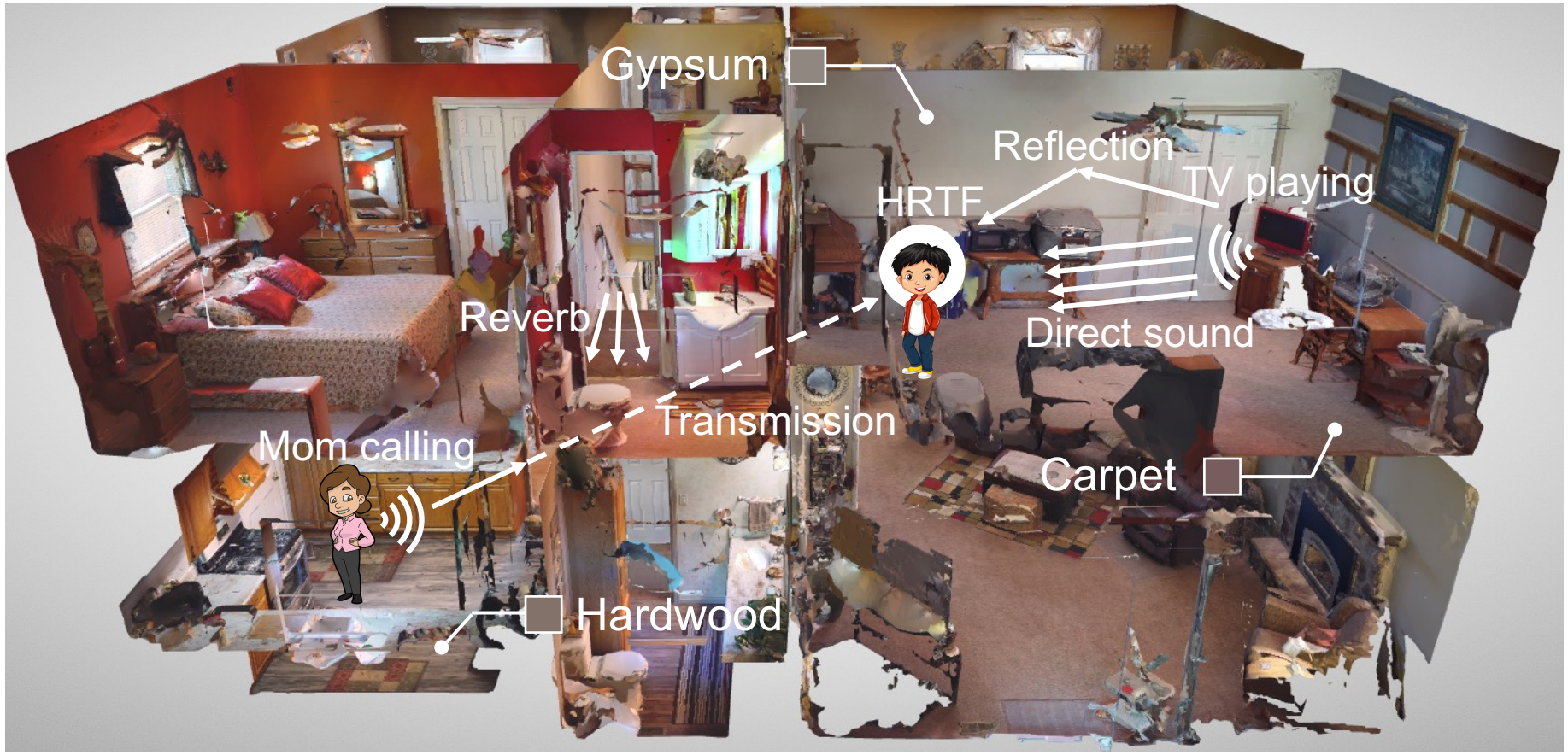
1. Expensive to store IRs
2. Does not generalize to new locations or environments
3. Microphones are not configurable







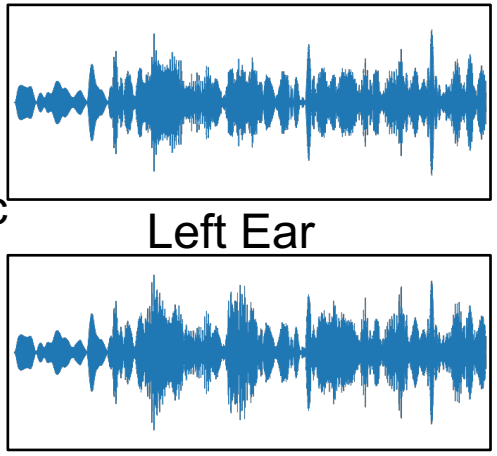
SoundSpaces 2.0: A fast, continuous, configurable and generalizable audio-visual simulation platform



Visual Render

RGB

Acoustic Render



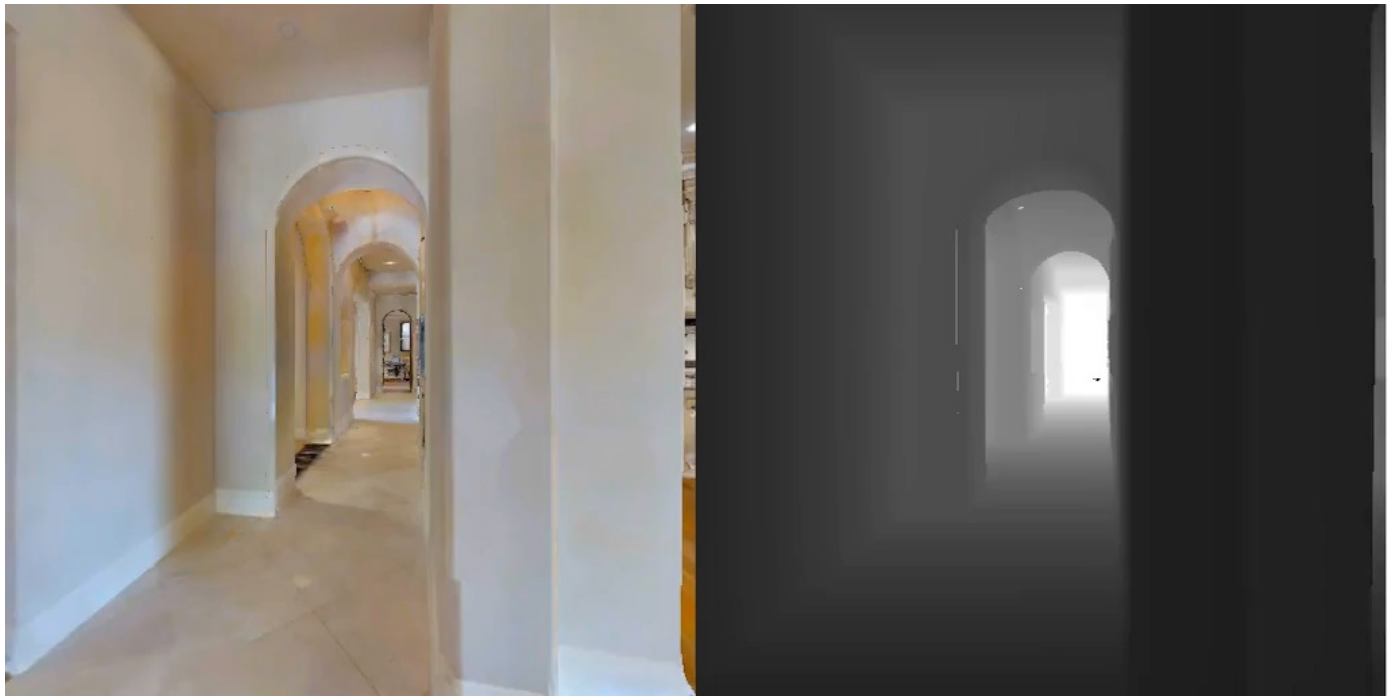
Left Ear

Right Ear

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
- …

Generalizable simulation

We support arbitrary scene datasets.

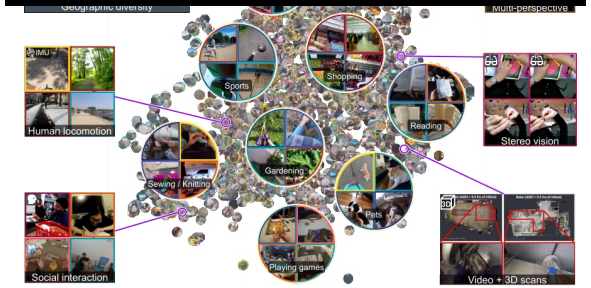
Gibson
Xia et al., CVPR 2018

Matterport3D
Chang et al., 3DV 2017

Replica
Straub et al., arXiV 2019

HM3D
Ramakrishnan et al., NeurIPS 2021

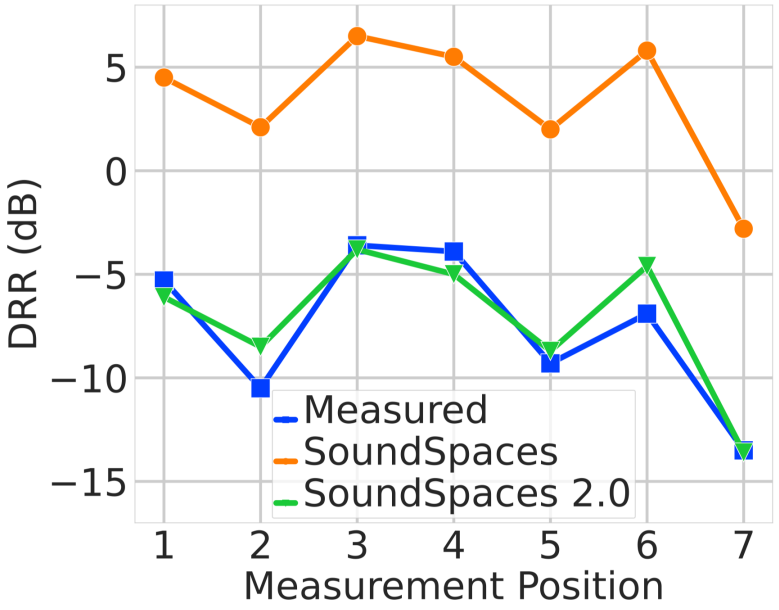
Ego4D
Grauman et al., CVPR 2022



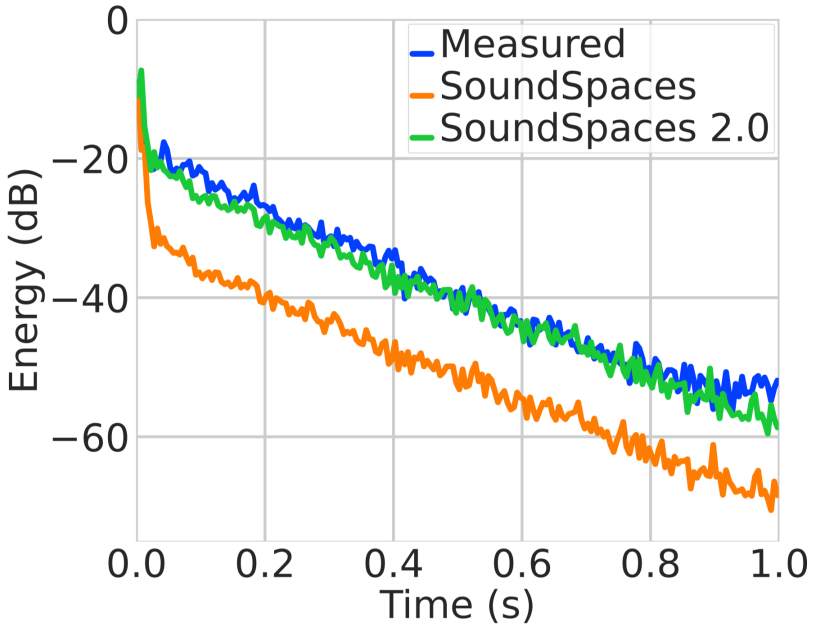
Validating simulation with real IRs

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

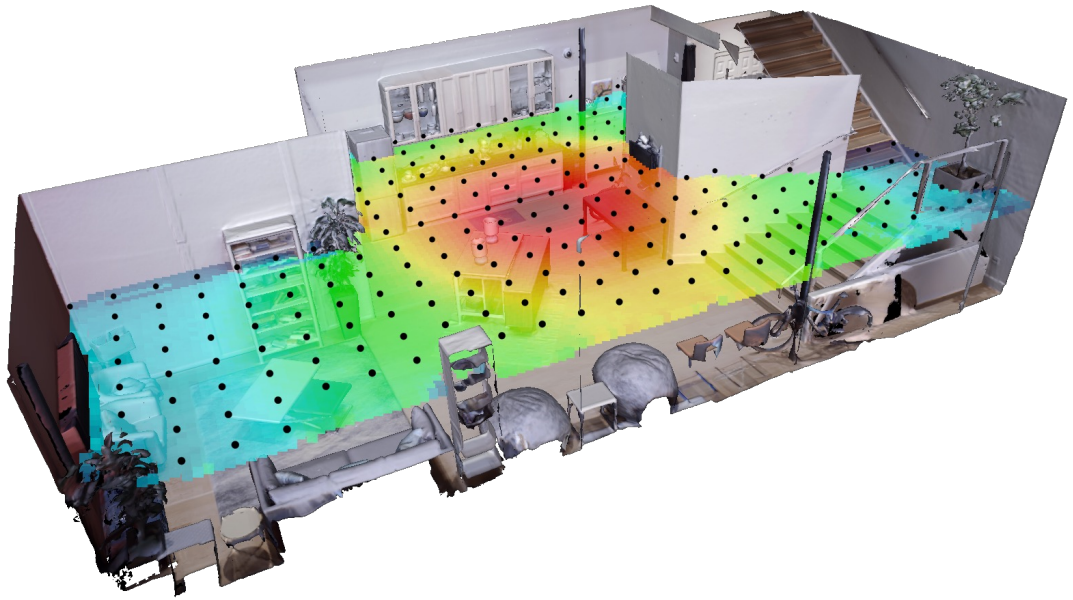


DRR comparison

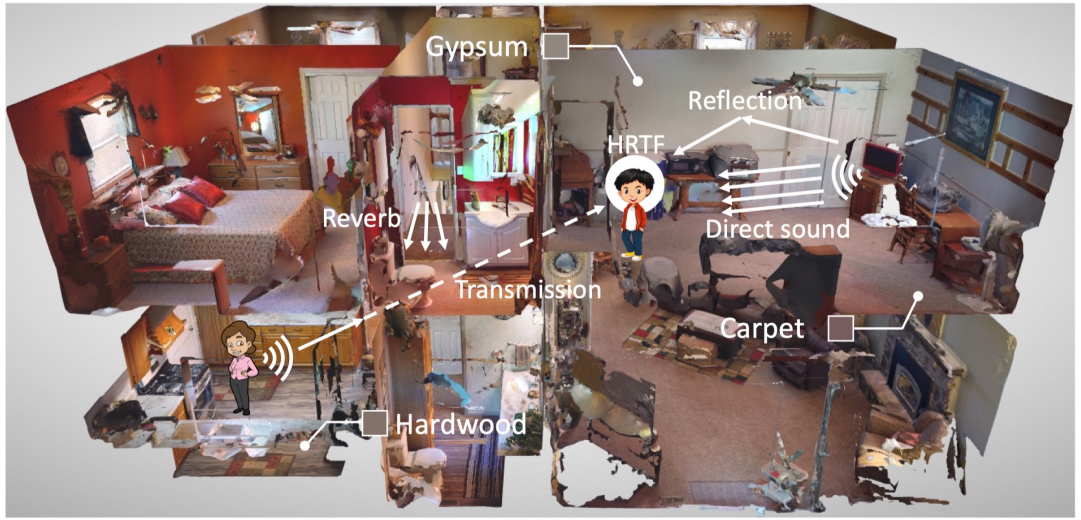


Early decay comparison

Main differences



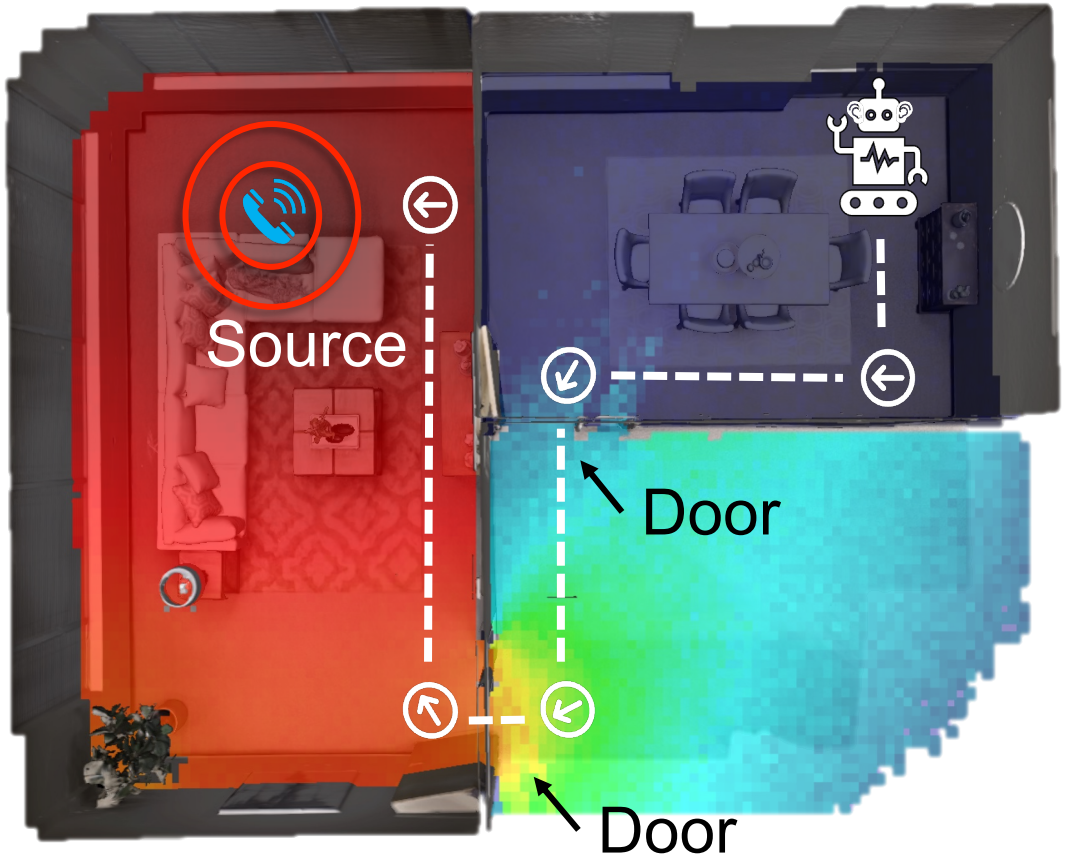
SoundSpaces 1.0
- 500 fps+
- Discrete and unconfigurable



SoundSpaces 2.0
- 30 fps+
- Continuous and configurable

Audio-visual navigation in 3D environments

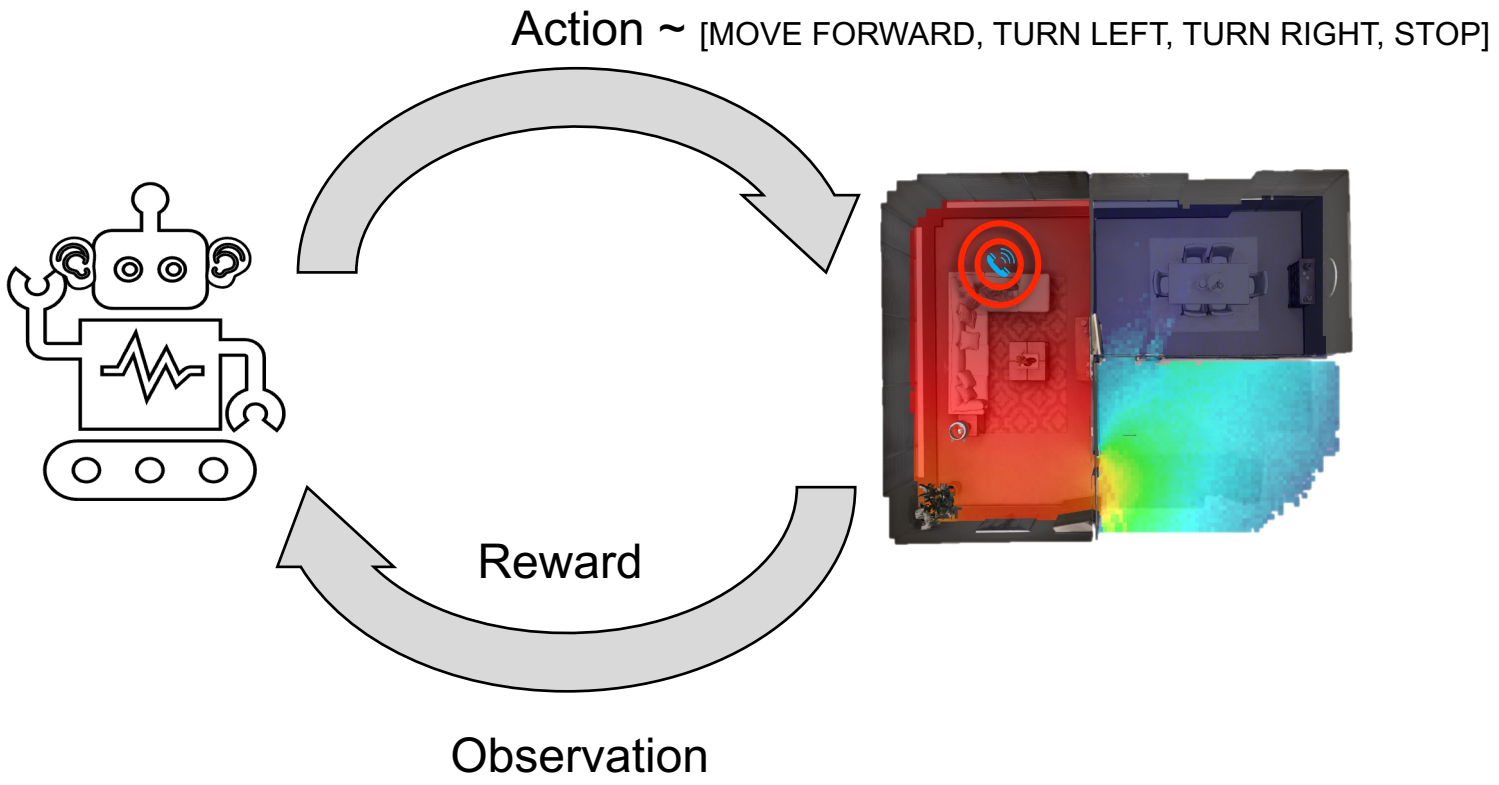
An agent navigates to a sounding object with vision and audio perception



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

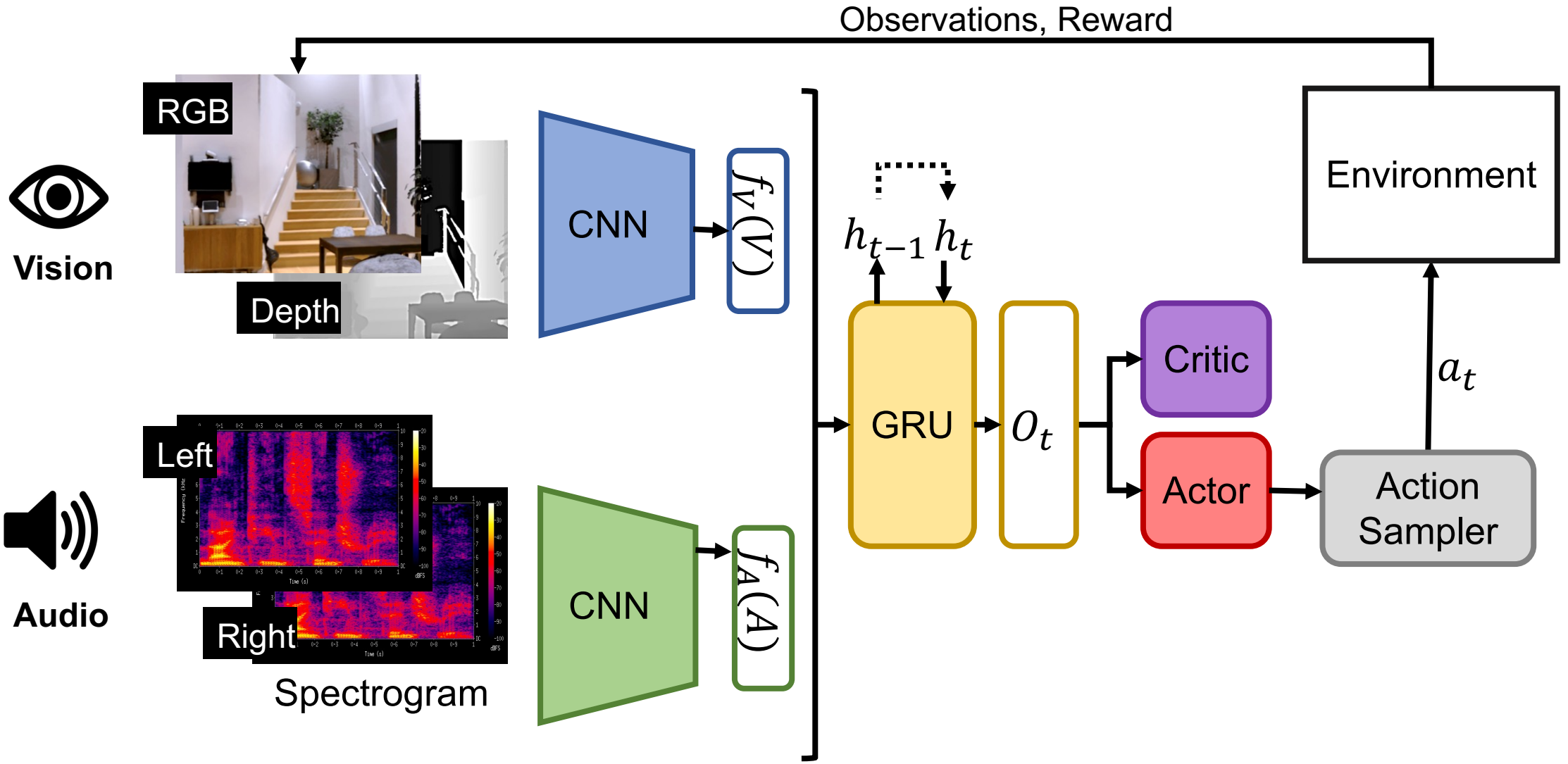
Learning with deep reinforcement learning

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



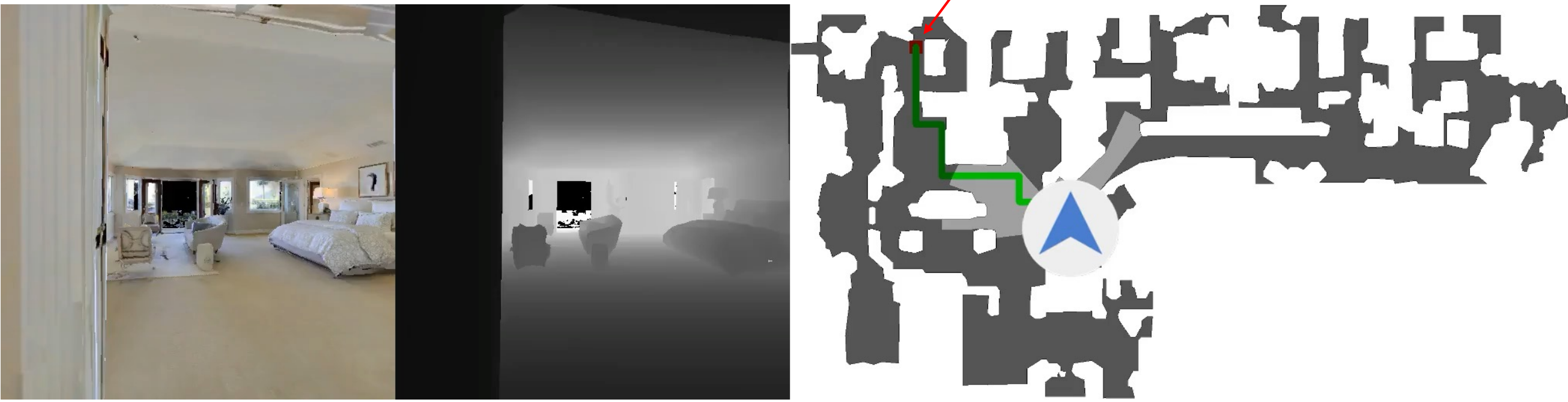
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Navigation policy



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

Navigation example



The agent leverages the complementary spatial information in audio and vision, and navigates to the goal successfully

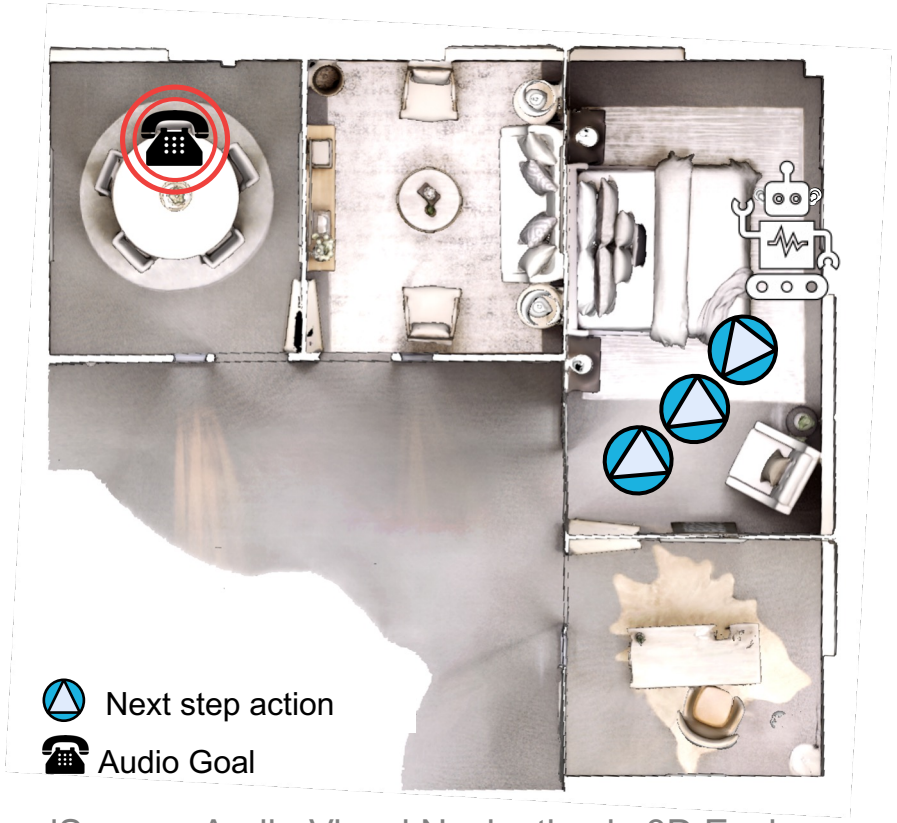
Agent Goal Start Shortest path Agent path Seen/Unseen area Occupied area Red Frame: Collision

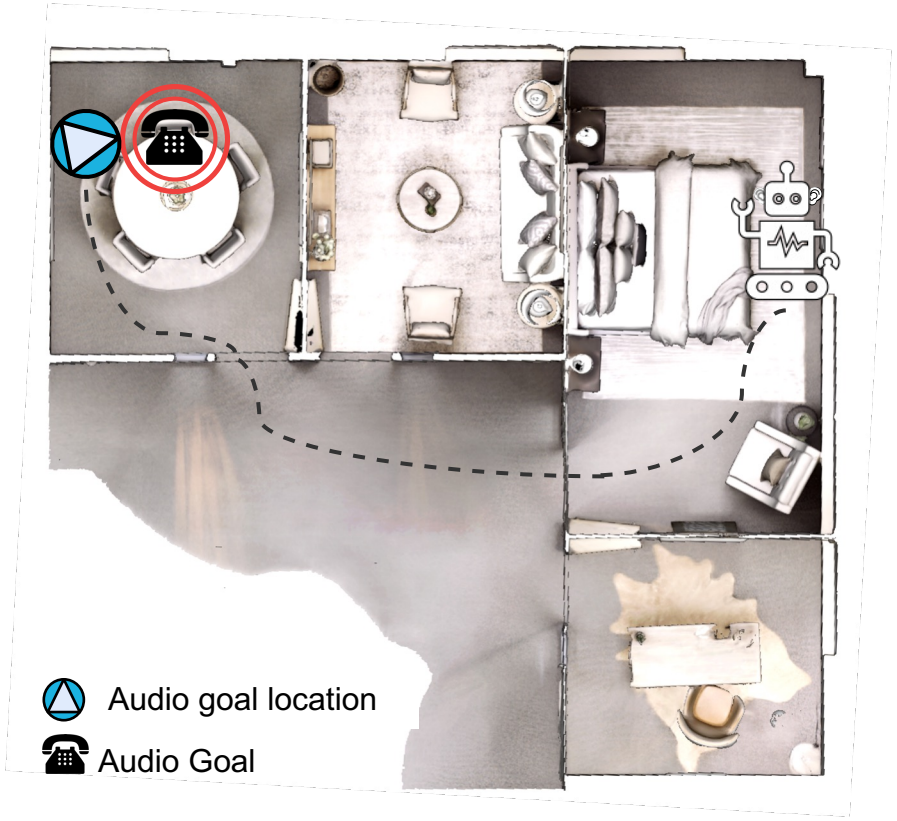
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

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

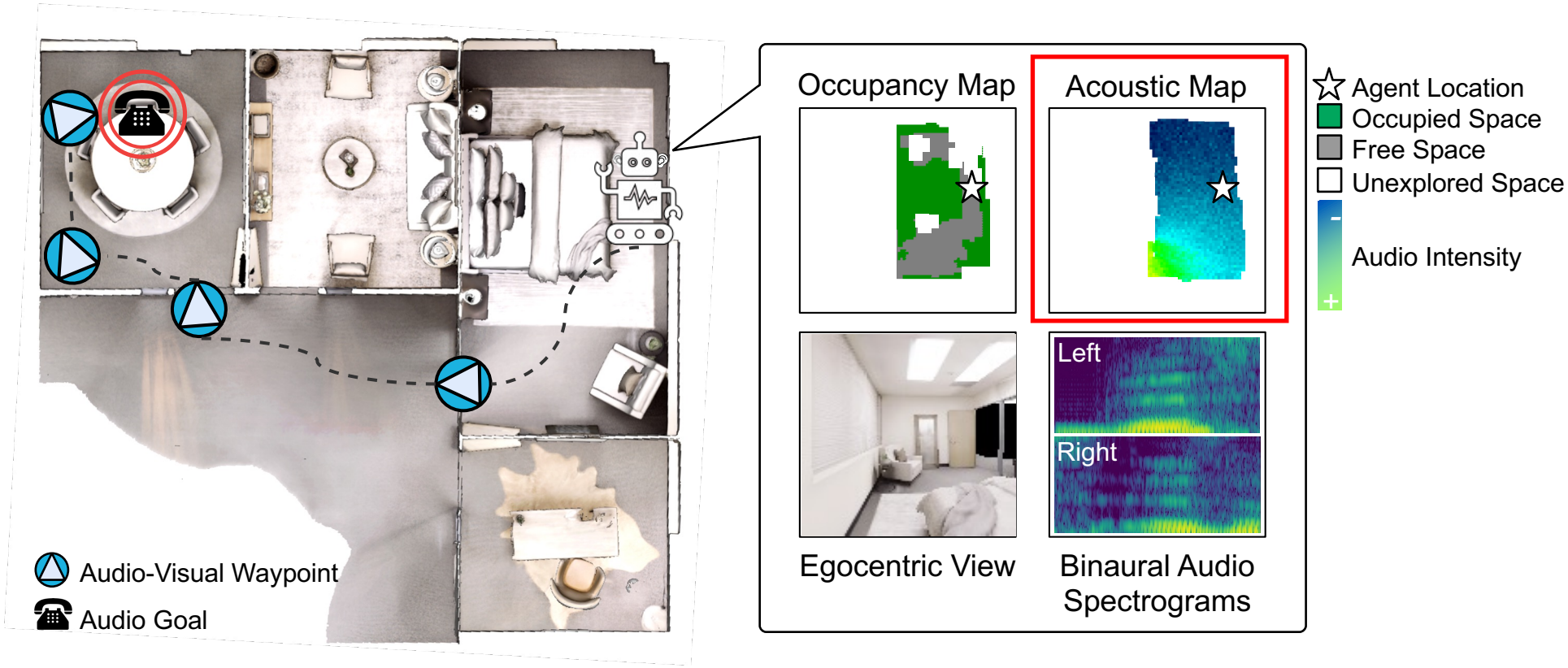




[1]SoundSpaces: Audio-Visual Navigation in 3D Environments, Chen et al., ECCV, 2020
[2]Look, Listen, and Act: Towards Audio-Visual Embodied, Gan et al., ICRA, 2020

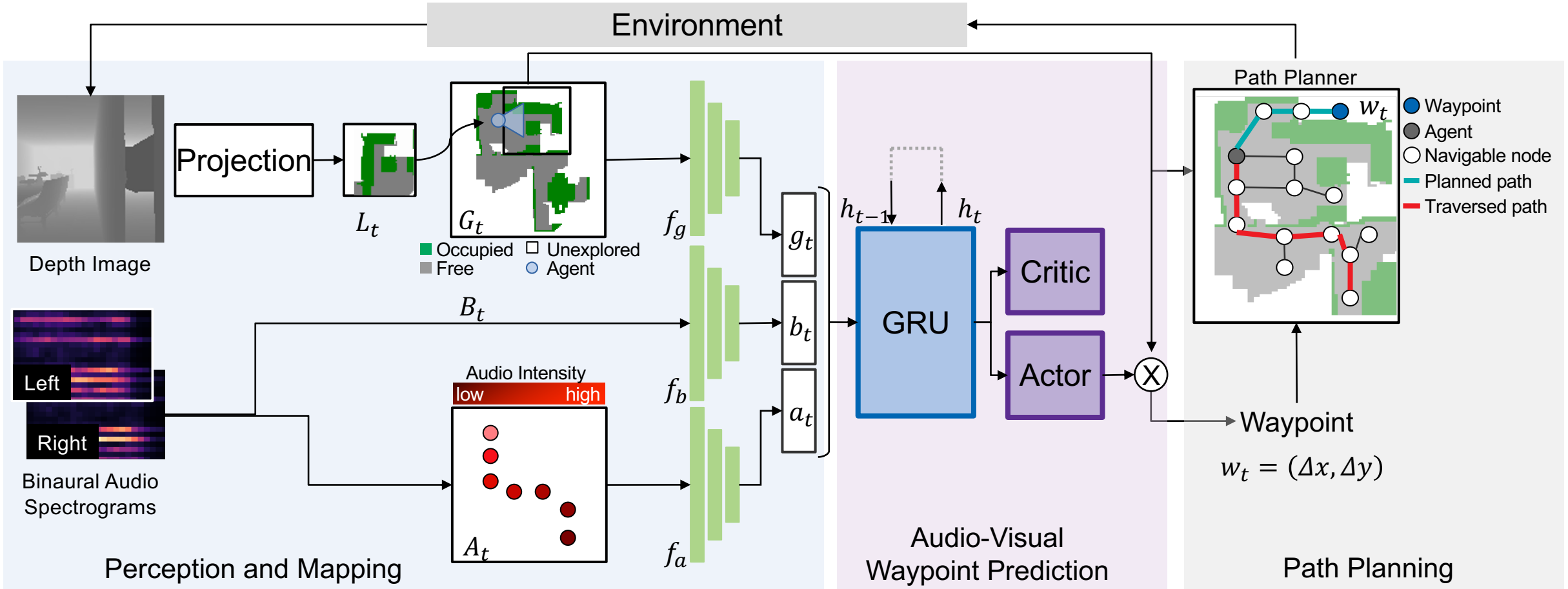
Learning to set waypoints for AV navigation

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



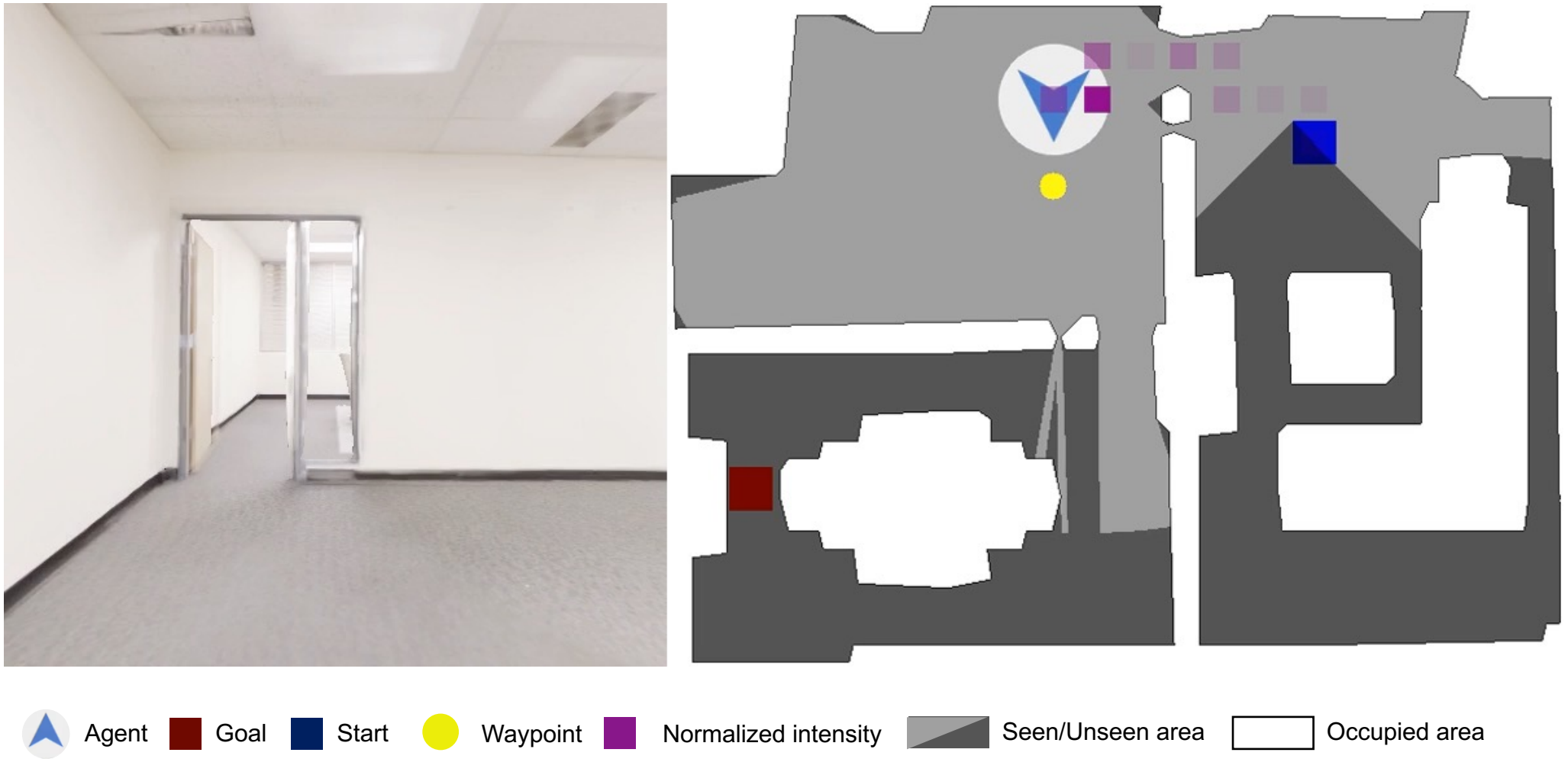
Changan Chen et al., Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021

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

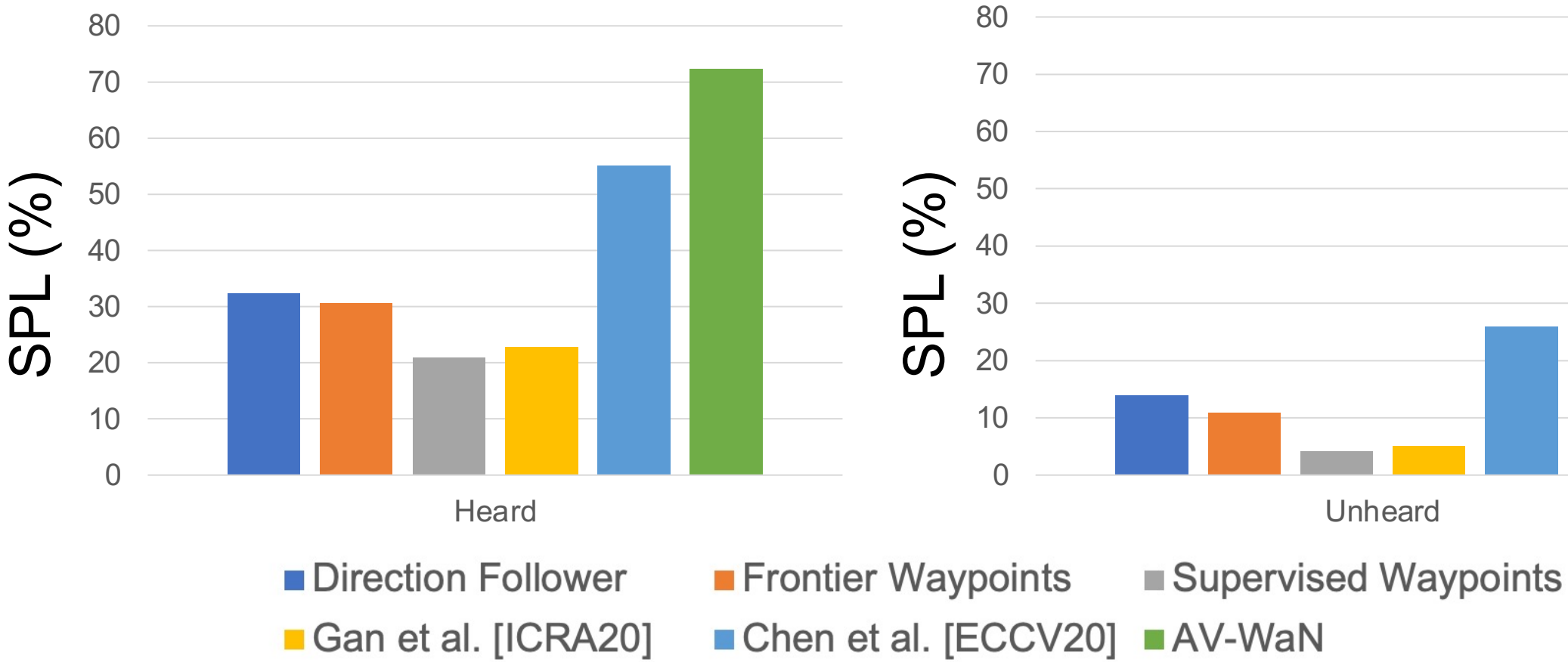


Our model dynamically selects waypoints and builds an acoustic memory as it moves.

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

Navigation results

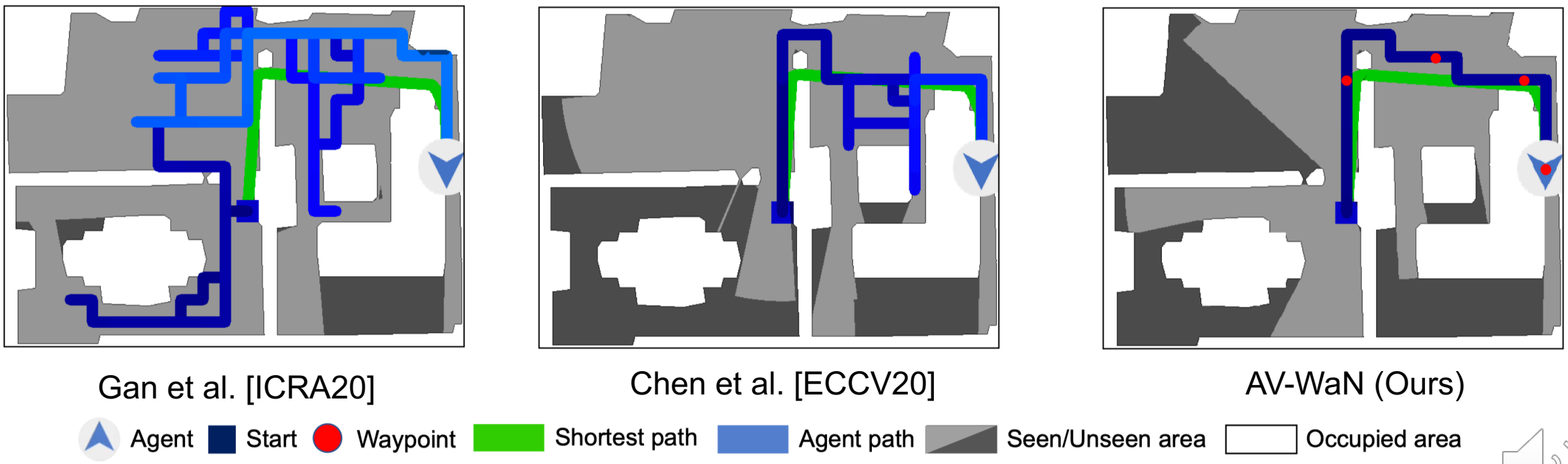
- Our model (AV-WaN) strongly outperforms all baselines and existing methods
- Generalizing to unheard sounds is much more challenging



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

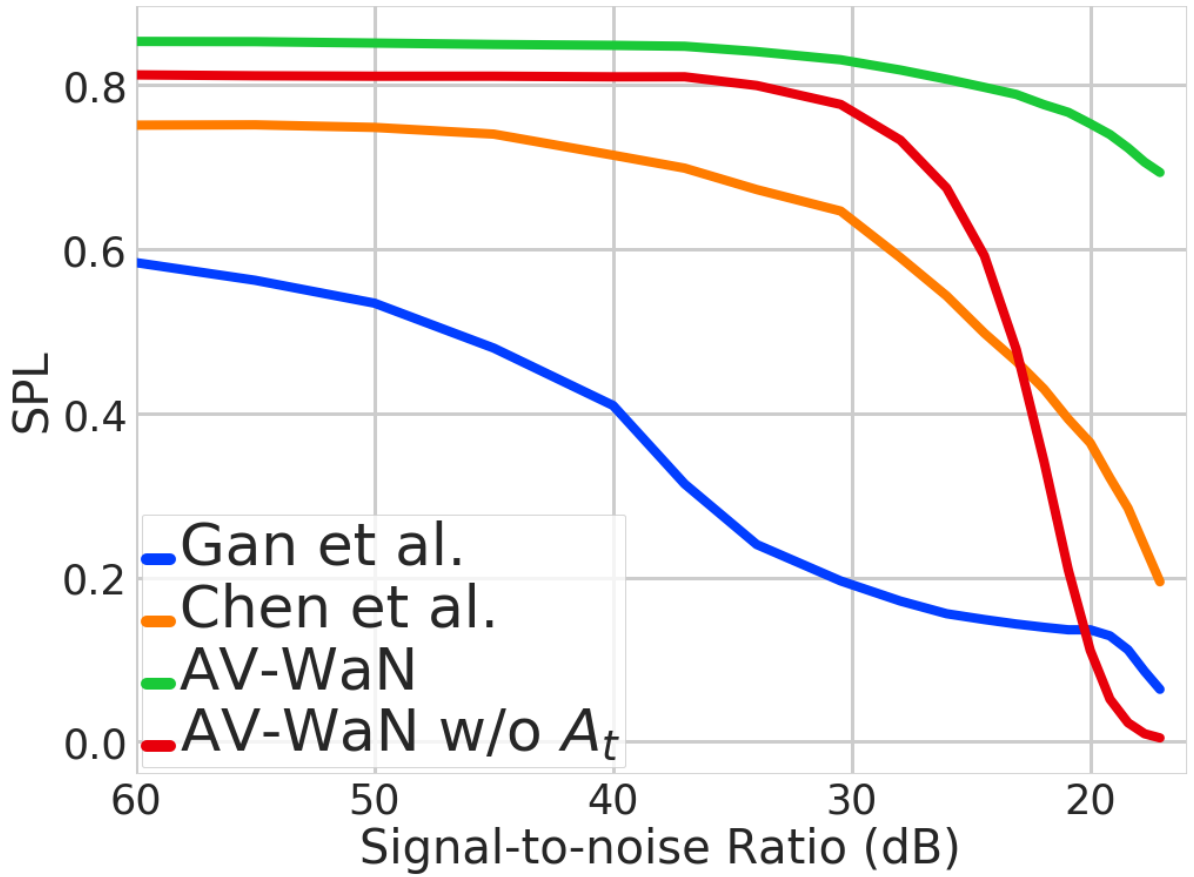
Navigation trajectories

- 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



Robustness to audio noise

AV-WaN with acoustic map is more robust to audio noise



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

Limitations of the AudioGoal task

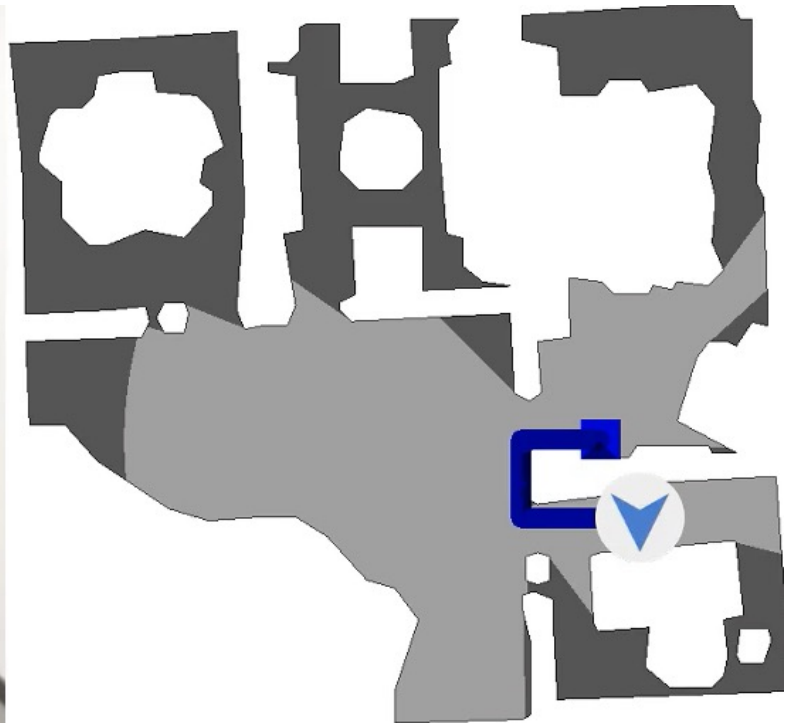
AudioGoal task (Chen et al. ECCV 2020, Gan et al. ICRA 2020):

- 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!

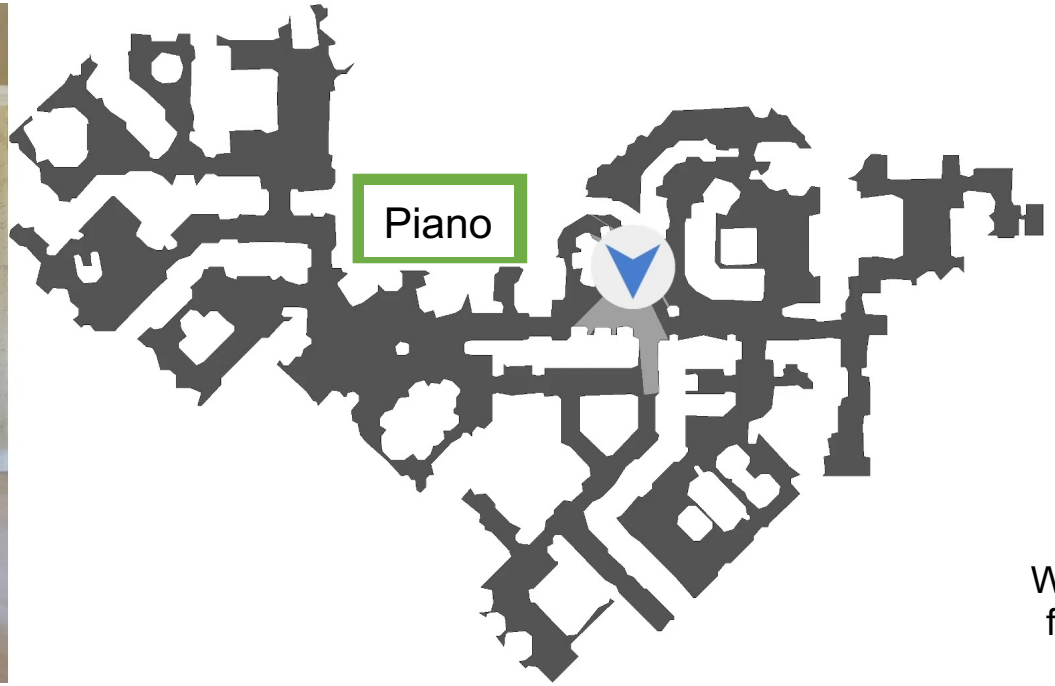


The agent searches for the ringing telephone in an unfamiliar environment

Semantic AudioGoal

Agent's egocentric view

Top-down map



Wear headphones
for spatial sound

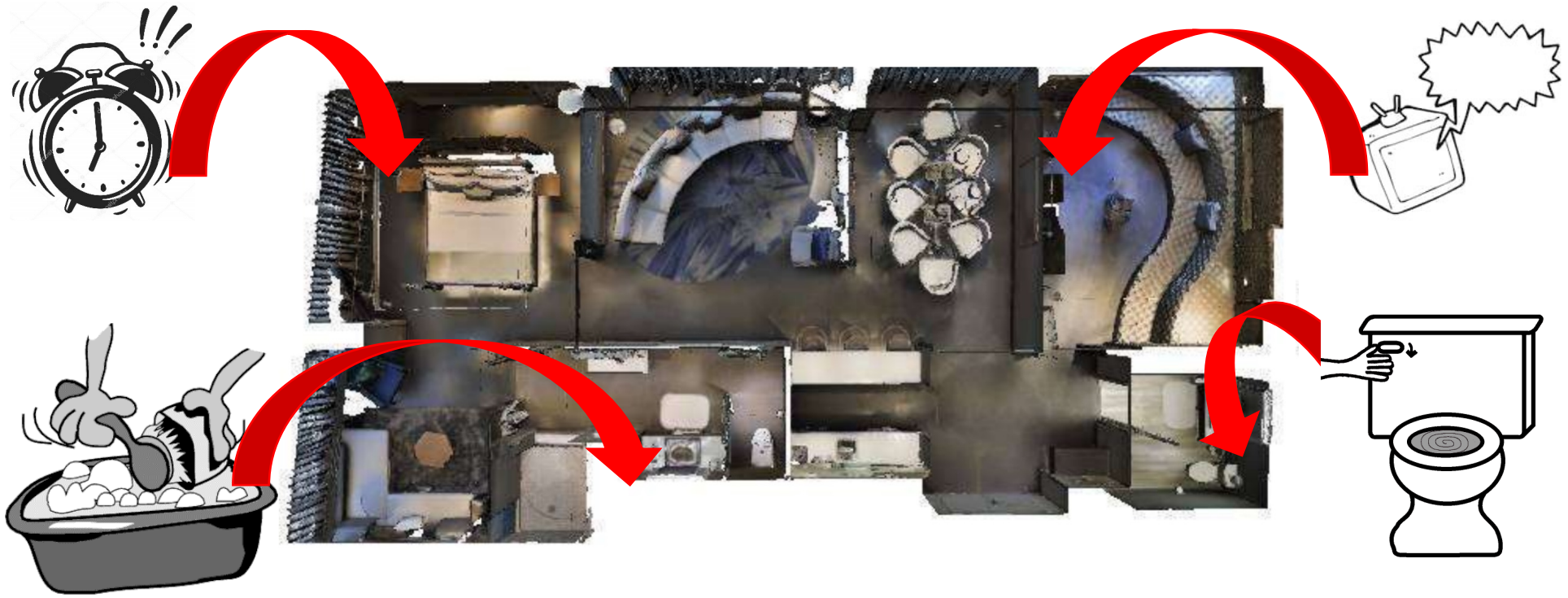
The agent must continue navigating even after the sound stops

Our proposed semantic AudioGoal task:

- The sound is associated with a semantically meaningful object
- The sound is not periodic and has variable length

Semantic AudioGoal dataset

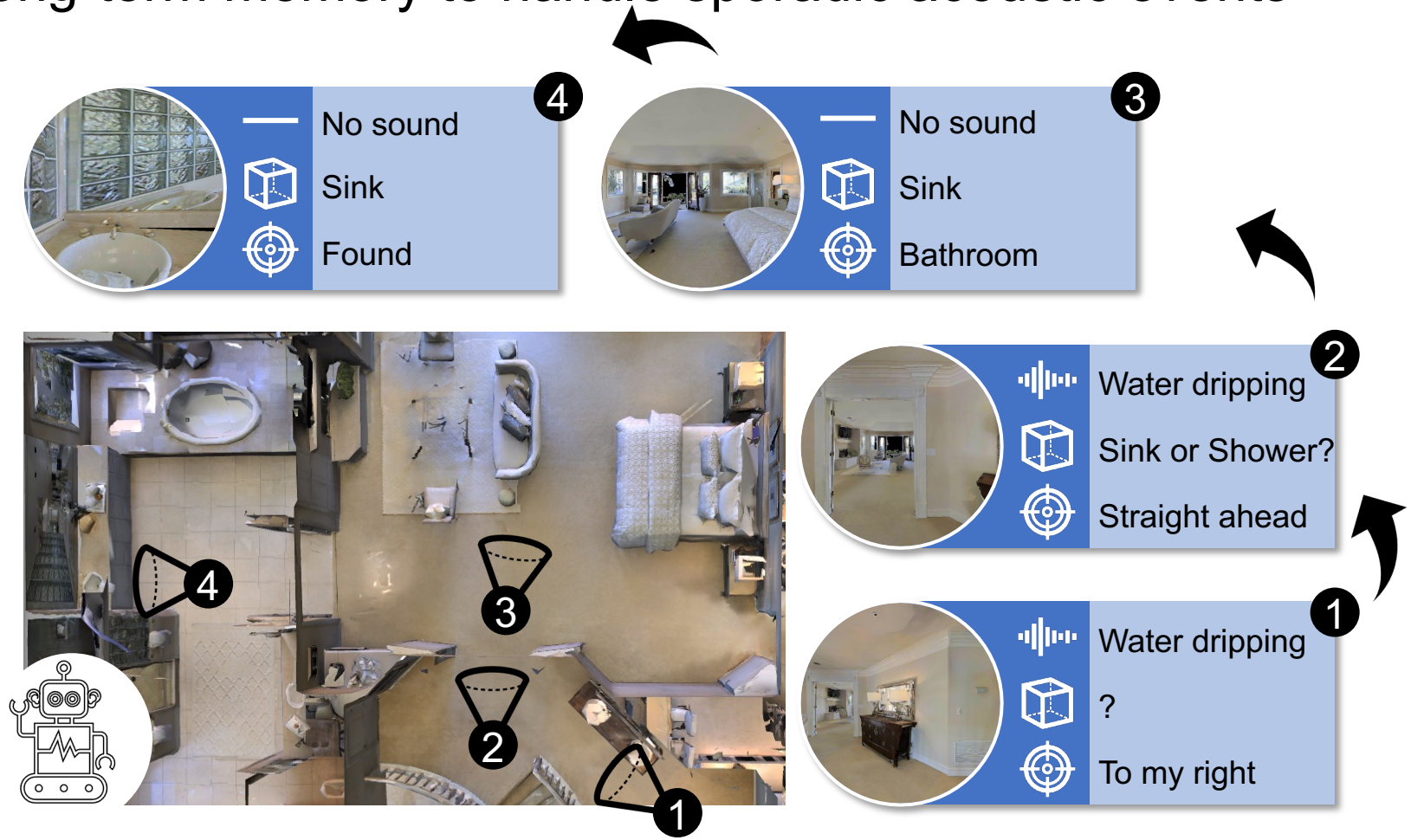
- Augment an existing simulator SoundSpaces[1] with semantic sounds
- 21 object categories in Matterport3D[2]: chair, TV, cabinet, sink etc.
- Object-emitted sounds and object-related sounds



[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

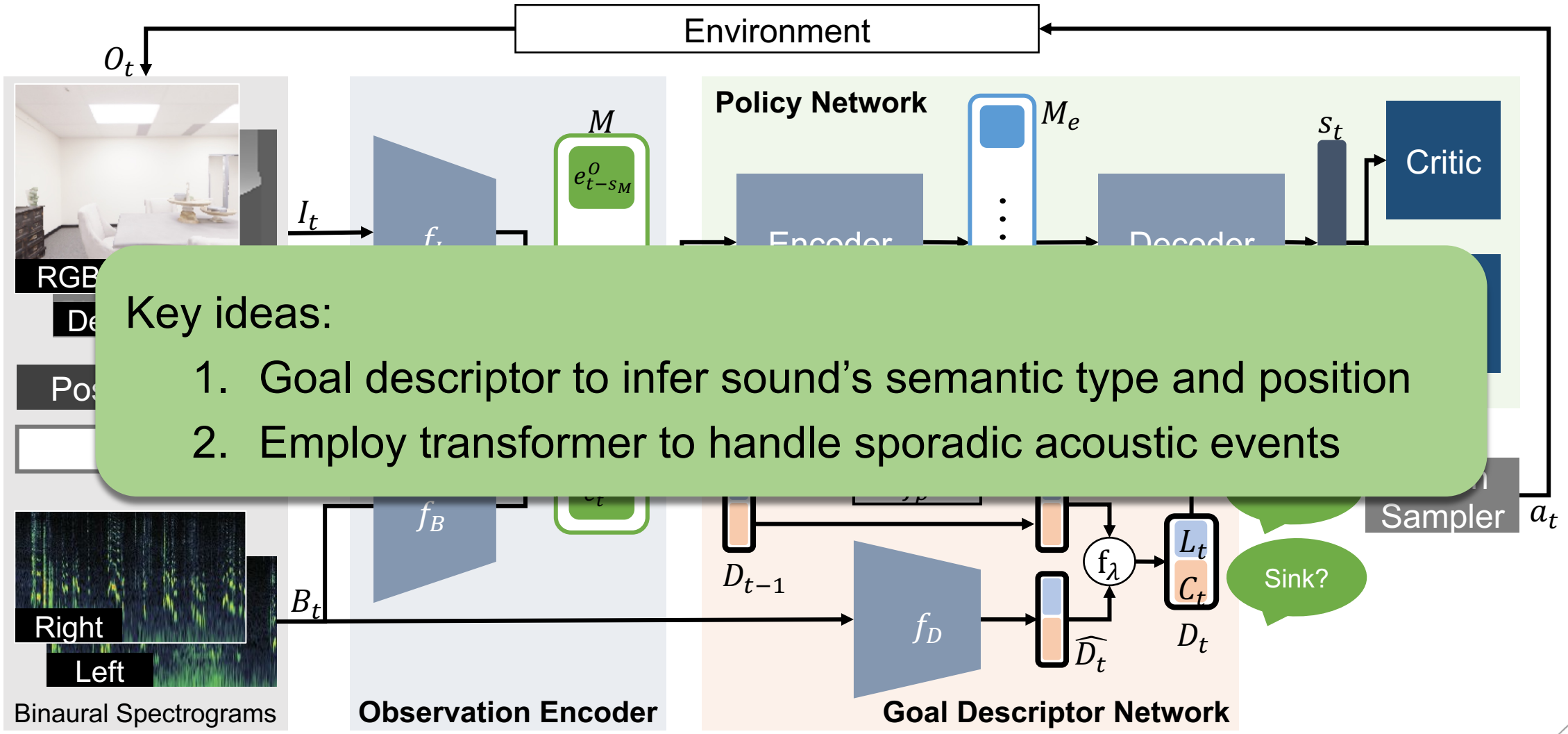
Semantic audio-visual navigation

- Learn the association between how objects look and how they sound
- Leverage long-term memory to handle sporadic acoustic events



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

Semantic audio-visual navigation model (SAVi)



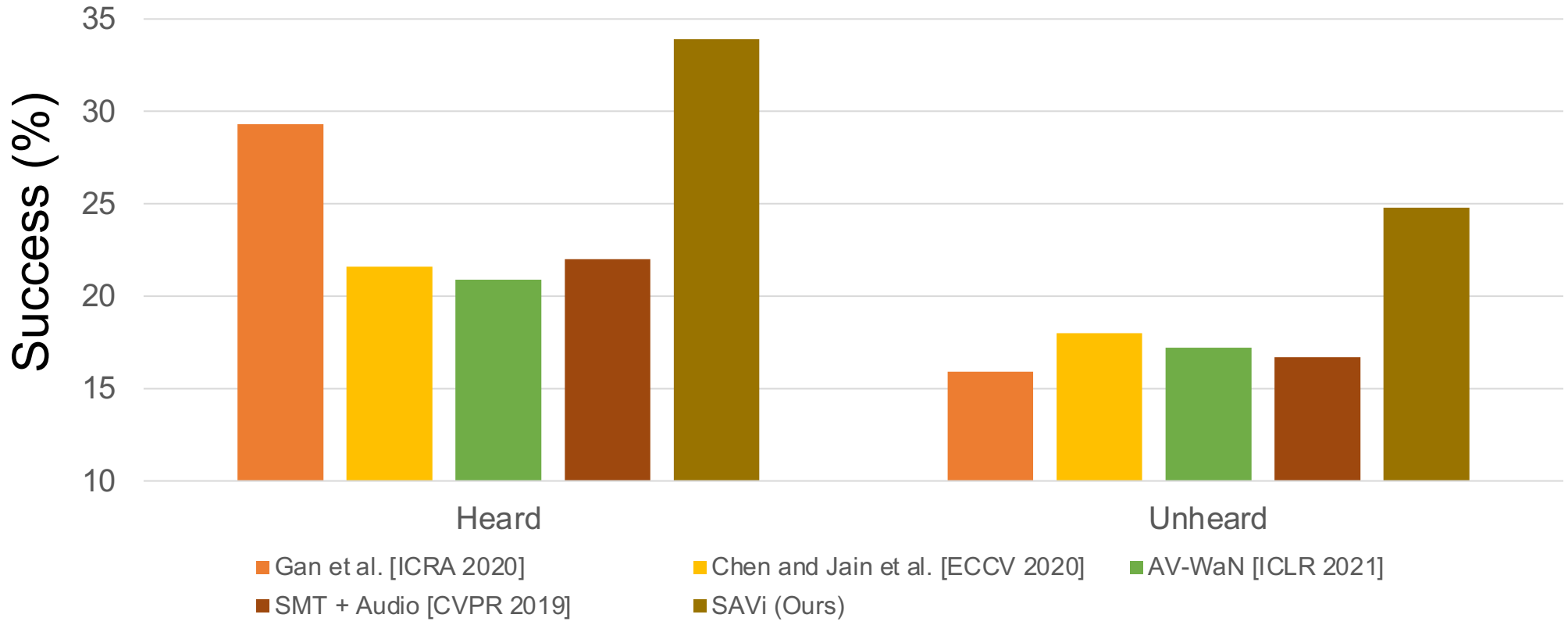
Key ideas:

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

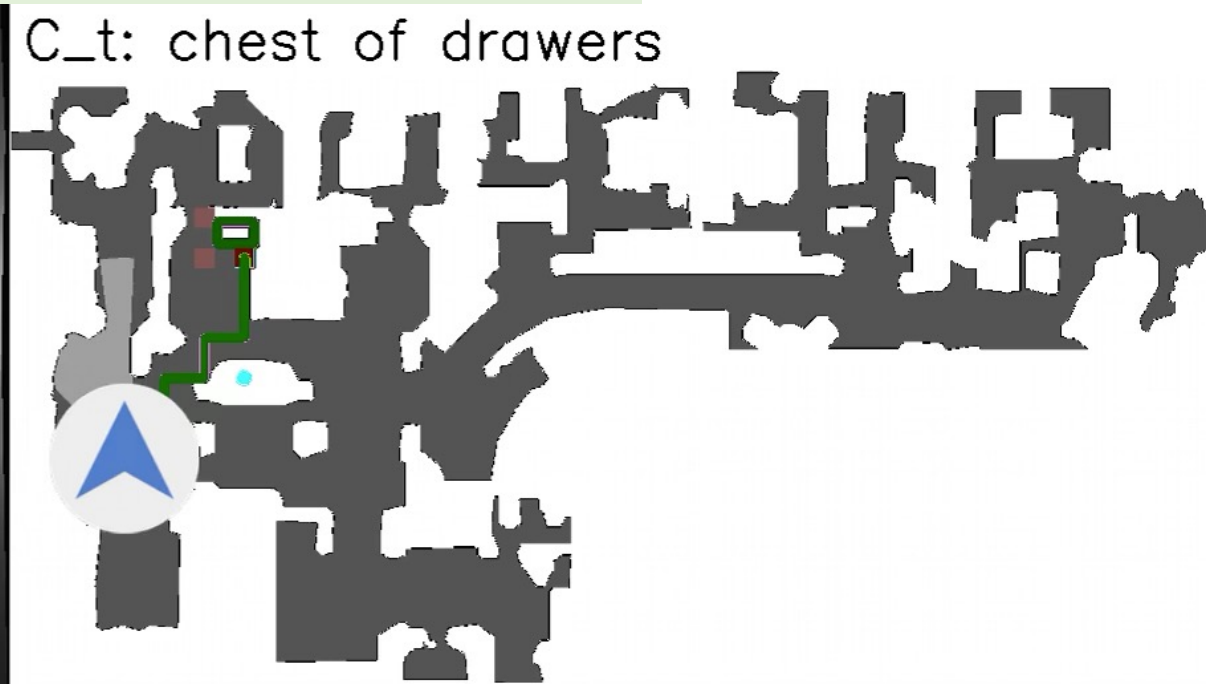
- SAVi strongly outperforms all existing methods
- Generalizing to unheard sounds is much more challenging



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

Navigation example

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



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

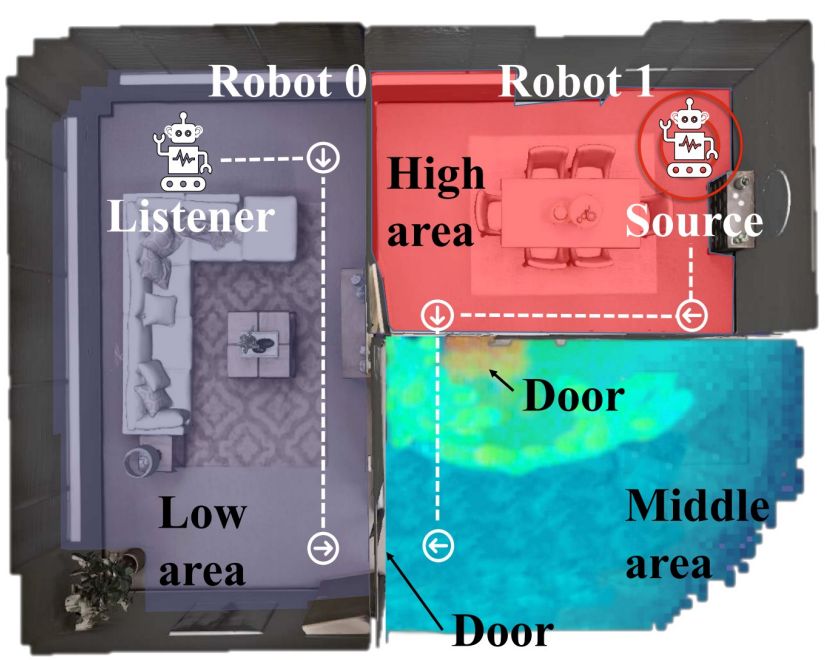
The agent identifies it's drawer sliding sound and locates the target object with vision after the sound stops.

Collaborative acoustic measurements

Measuring acoustics traditionally relies on manual sampling

Multi-agent

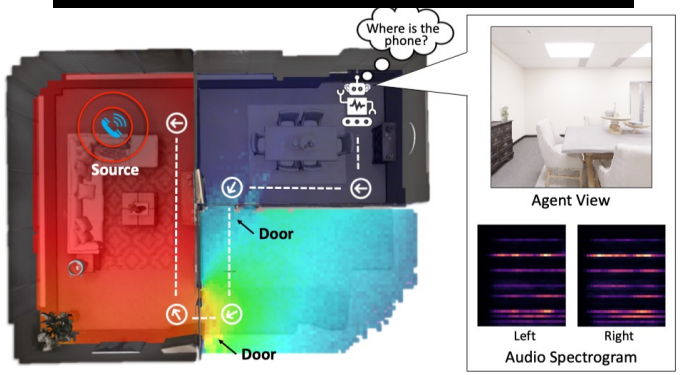
Acoustic exploration



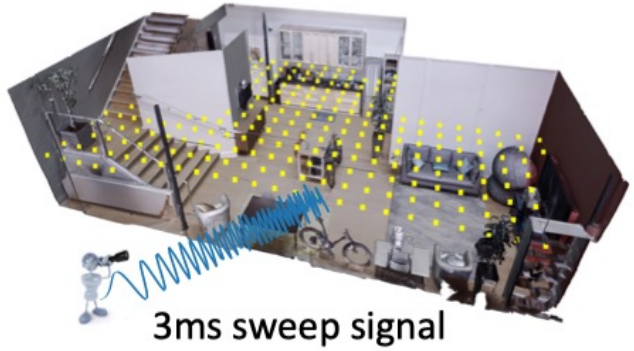
Yinfeng Yu, Changan Chen et al., Measuring Acoustics with Collaborative Multiple Agents, IJCAI 2023

Much more to be explored …

Audio-Visual Navigation
Chen et al., ECCV 2020

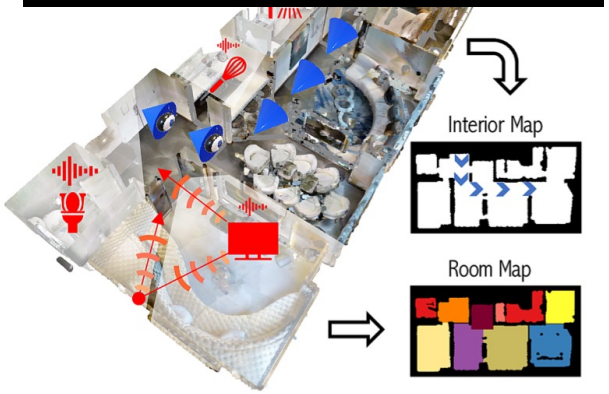


Echolocation Learning
Gao et al., ECCV 2020

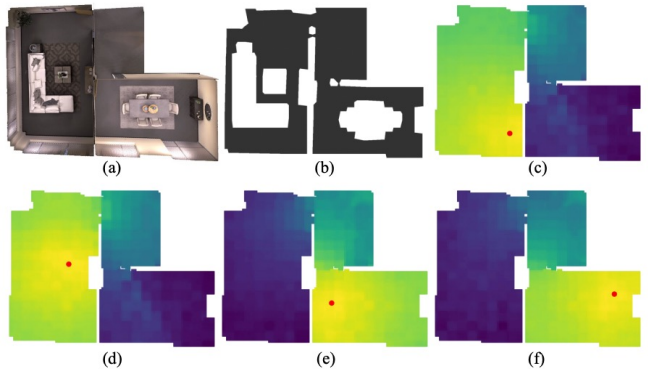


Audio-Visual Separation
Majumder et al., ICCV 2021

Audio-Visual Mapping
Purushwalkam et al., ICCV 2021



Neural Acoustic Rendering
Luo et al., NeurIPS 2022



Audio Scene Reconstruction
Chen et al., ICCV 2023