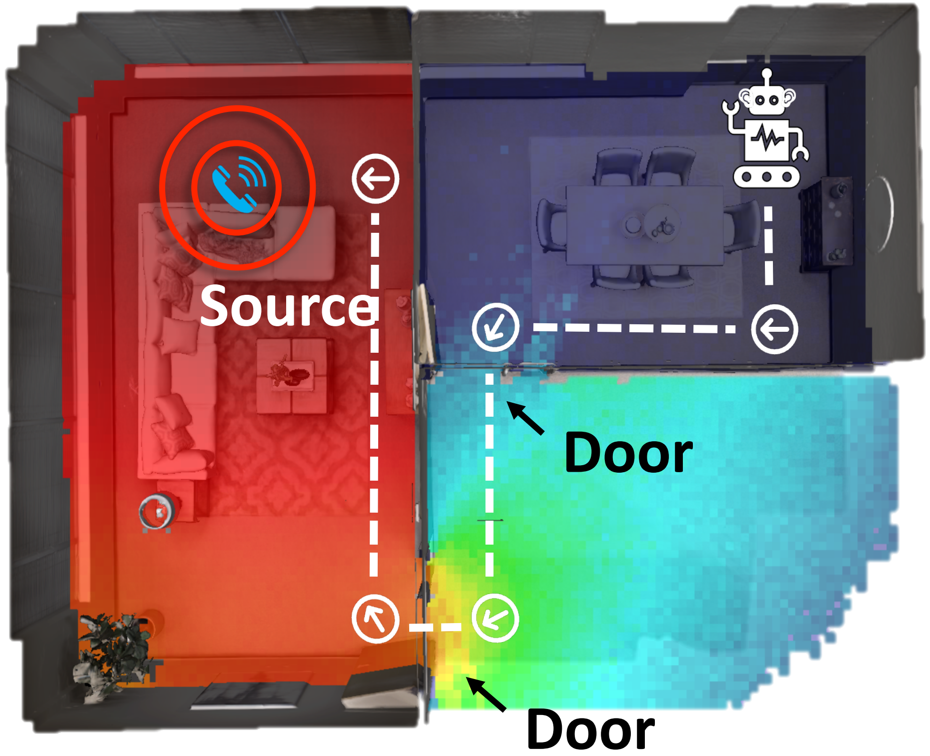
SoundSpaces: Audio-Visual Navigation in 3D Environments

Presenter: Changan Chen

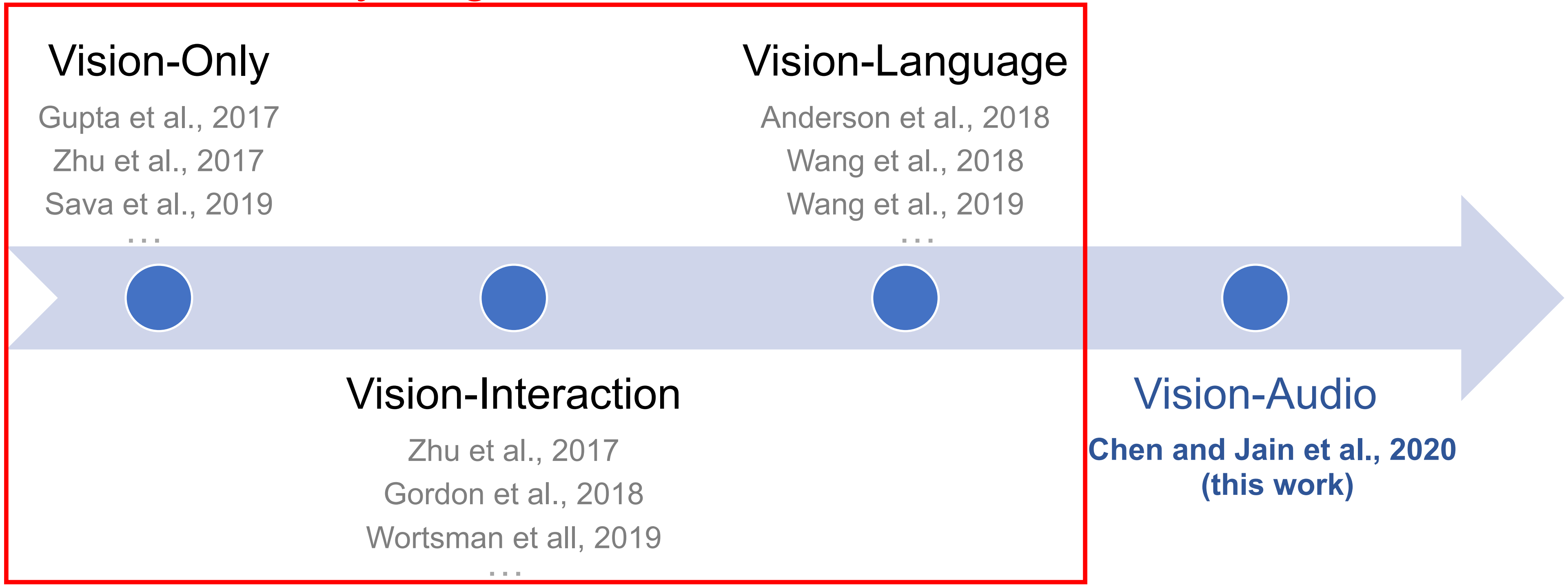
09/23/2021



Embodied Perception Is a Multisensory Experience

We often use *vision*, *audio*, *touch*, *smell* to move around

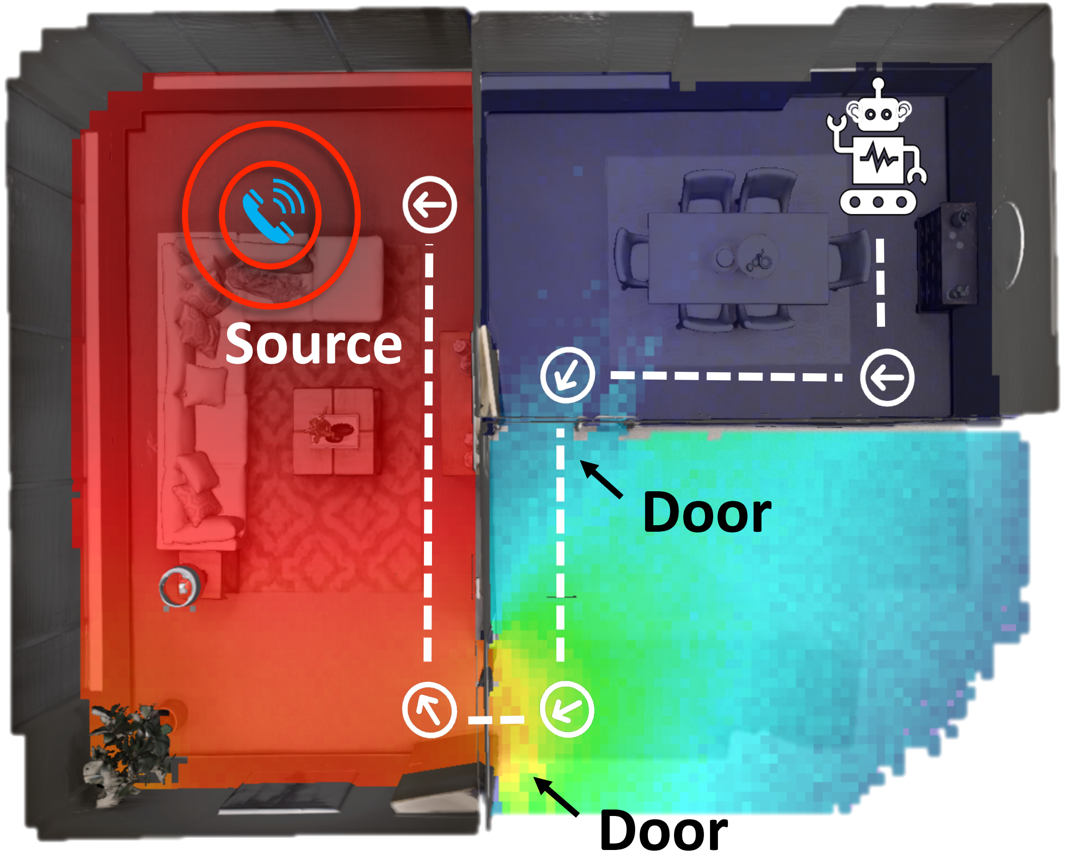
Today's agents are deaf!



Our contribution: audio-visual embodied navigation --- task and simulation

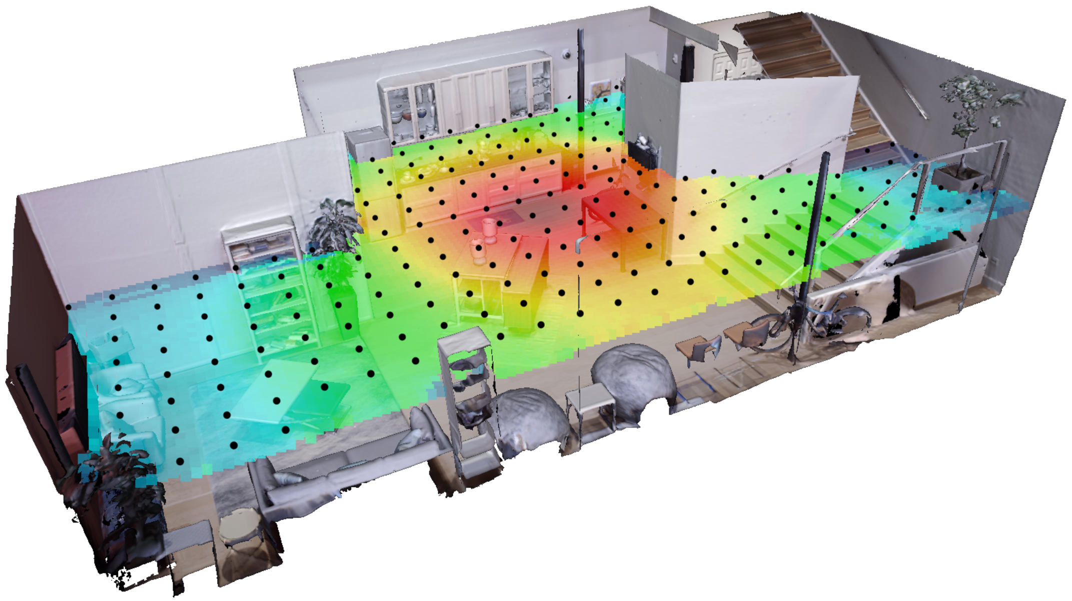
Audio-Visual Navigation in 3D Environments

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



SoundSpaces: Our Audio Simulator

- We introduce SoundSpaces, an audio simulation platform to enable audio-visual navigation for two visually realistic 3D environments: Replica[1] and Matterport3D[2]



	# Scenes	Avg. Area	# Training Eps.
Replica	18	47.24 m^2	0.1M
Matterport3D	85	517.34 m^2	2M

Table: Summary of dataset statistics

[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

- We introduce SoundSpaces, an audio simulation platform to enable audio-visual navigation for two visually realistic 3D environments: Replica[1] and Matterport3D[2]

- Our audio simulator produces realistic audio rendering based on the room geometry, materials, and sound source location

- The platform can play varying sounds of your choice in real time by precomputing a transfer function between locations

[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

Example: Where Is My Phone?

Agent view

Top-down map (unknown to the agent)

Direction: left ear is louder when the agent faces upward on the top-down map
Intensity: overall intensity gets higher as the agent gets closer to the goal

Agent | Goal | Start | Shortest path | Agent path | Seen/Unseen area | Occupied area

Example 2: Where Is The Piano?

Agent view

Top-down map (unknown to the agent)

Agent | Goal | Start | Shortest path | Agent path | Seen/Unseen area | Occupied area

Audio-Visual Navigation Tasks

PointGoal

Gupta et al., 2017
Savva et al., 2019
…

GPS

The agent receives a displacement vector (Δx, Δy) pointing towards the goal at each time step

New tasks

AudioGoal

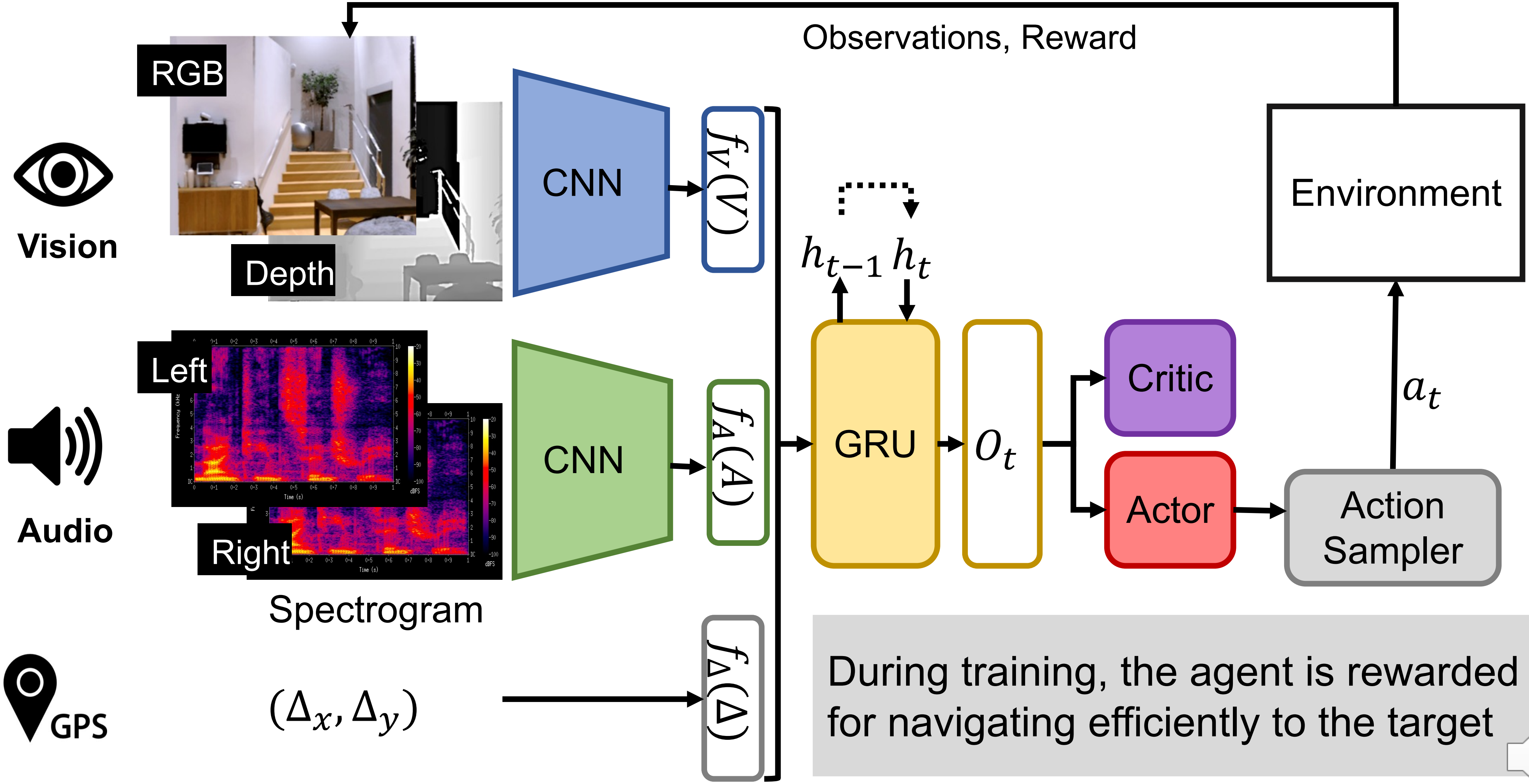
The agent receives an audio signal emitted by the sounding object at each time step

AudioPointGoal

\+ GPS

The agent receives both a displacement vector (Δx, Δy) and an audio signal at each time step

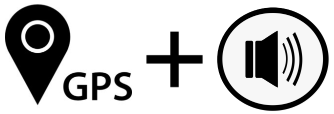
Deep RL for Audio-Visual Navigation



During training, the agent is rewarded for navigating efficiently to the target

Navigation Demo - AudioPointGoal

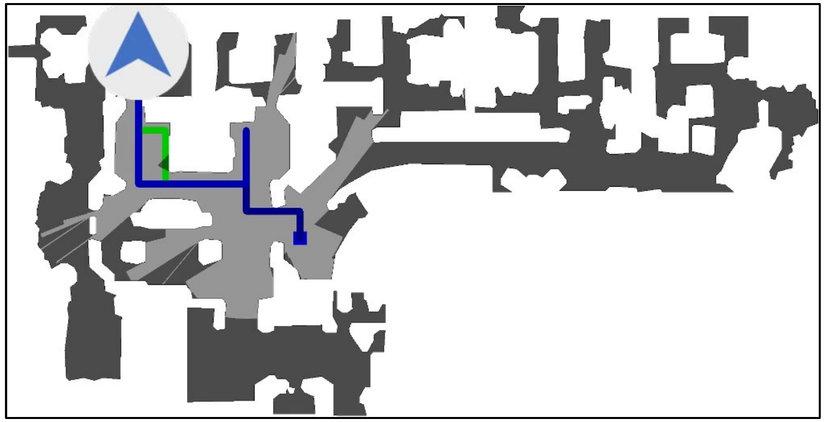
SPL: 1.00



AudioPointGoal agent leverages the complementary information in audio and GPS, and navigates to the goal efficiently

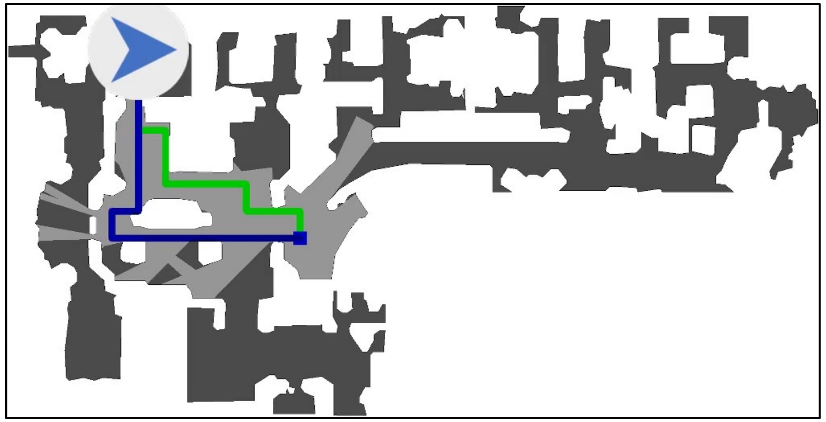
Agent | Goal | Start | Shortest path | Agent path | Seen/Unseen area | Occupied area | Red Frame: ...ision

Navigation Trajectory Comparison



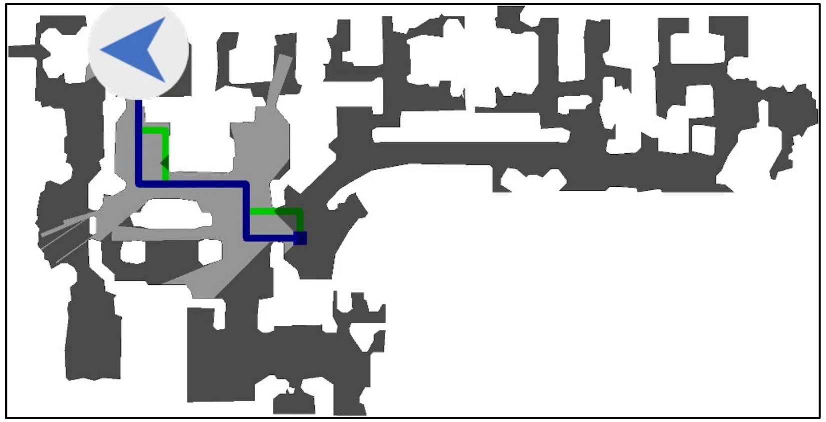
SPL: 0.68

PointGoal agent gets confused about the direction and gets stuck behind the bed.

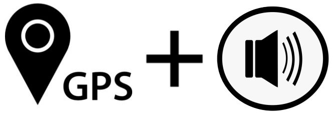


SPL: 0.87

AudioGoal agent figures out the sound comes from the front more quickly than the PointGoal agent



SPL: 1.00



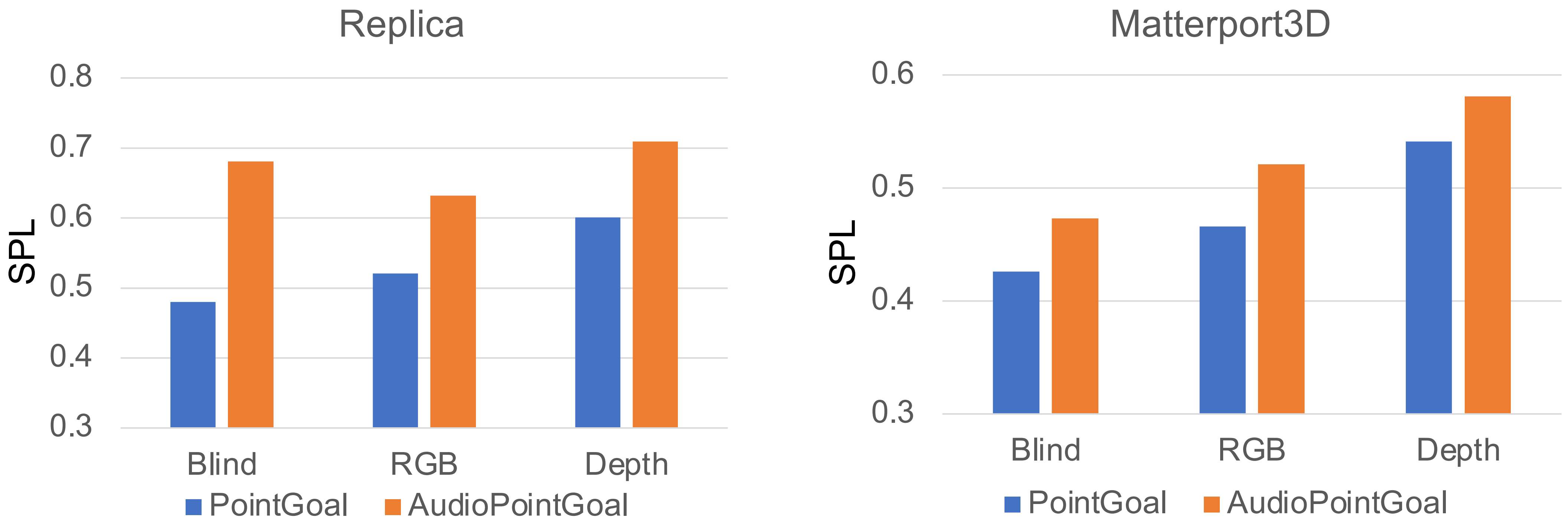
AudioPointGoal agent knows immediately it should go straight and then right and thus follows the shortest path

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

Does Audio Help Navigation?

Comparing PointGoal (PG) and AudioPointGoal (APG):

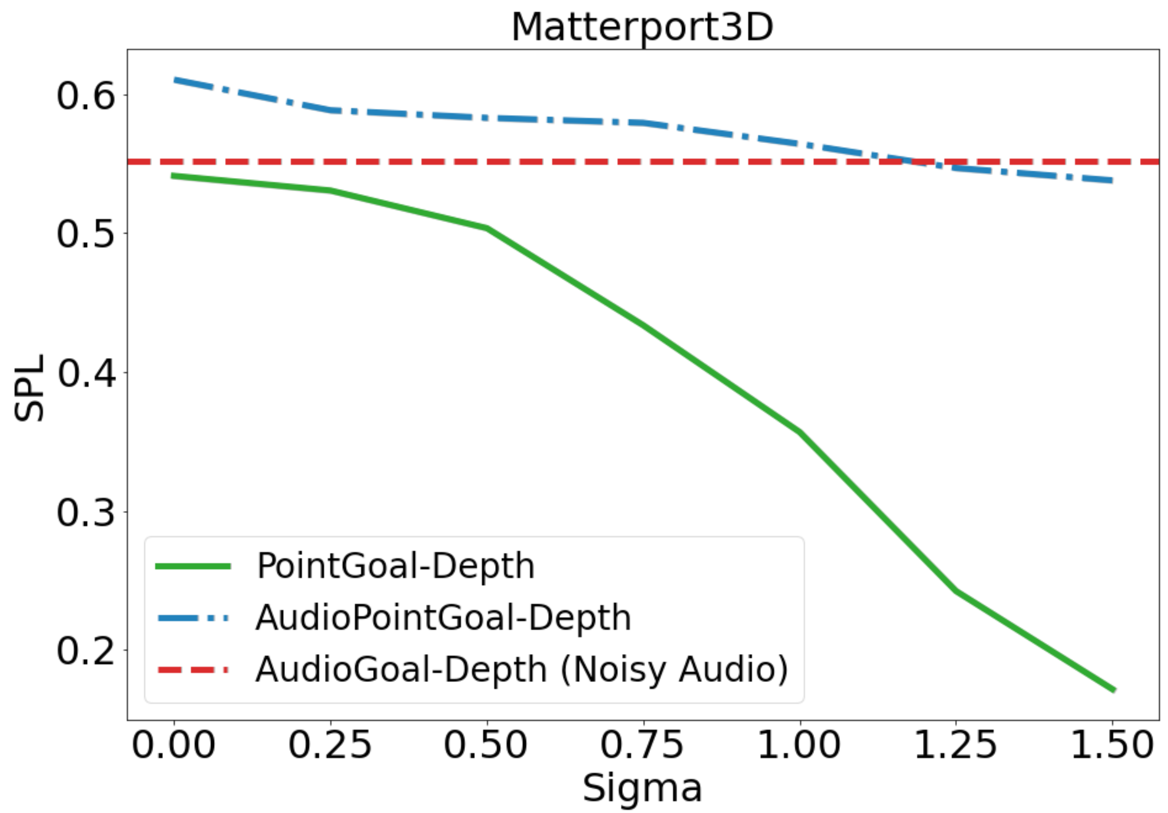
- Audio improves accuracy significantly



Metric: SPL (success weighted by inverse path length)

Can Audio Supplant GPS for AudioGoal?

- AudioGoal is immune to GPS noise (localization error) and robust to microphone noise
- AudioPointGoal degrades less in the presence of GPS noise
- Audio signal gives similar or even better spatial cues than the PointGoal displacements

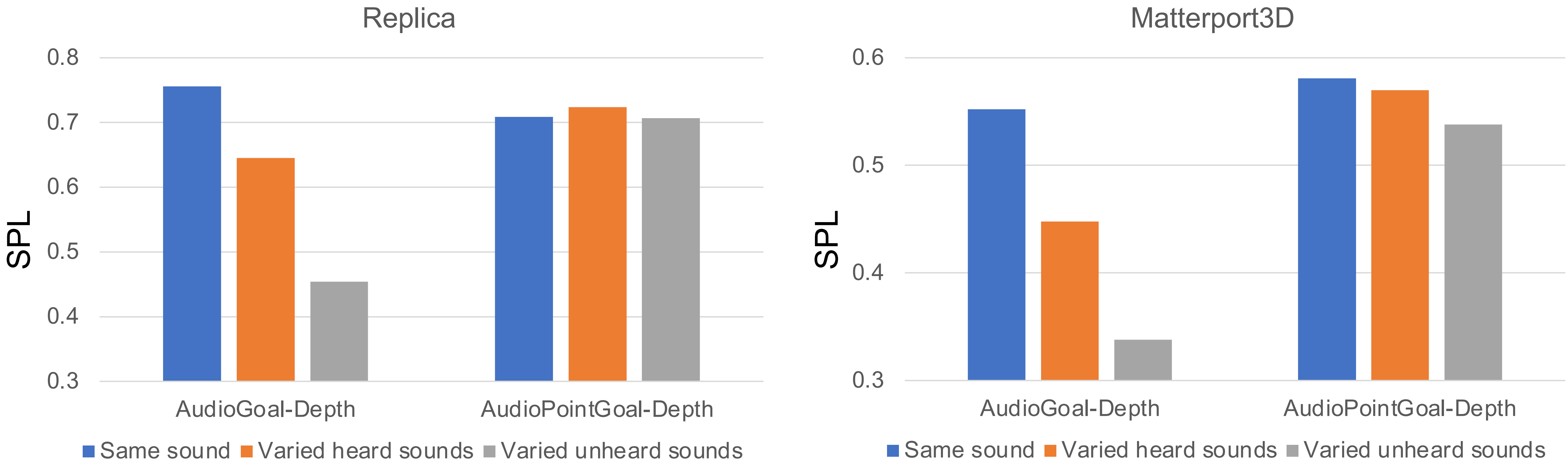


Perfect GPS → Noisy GPS

Effect of Different Sound Sources

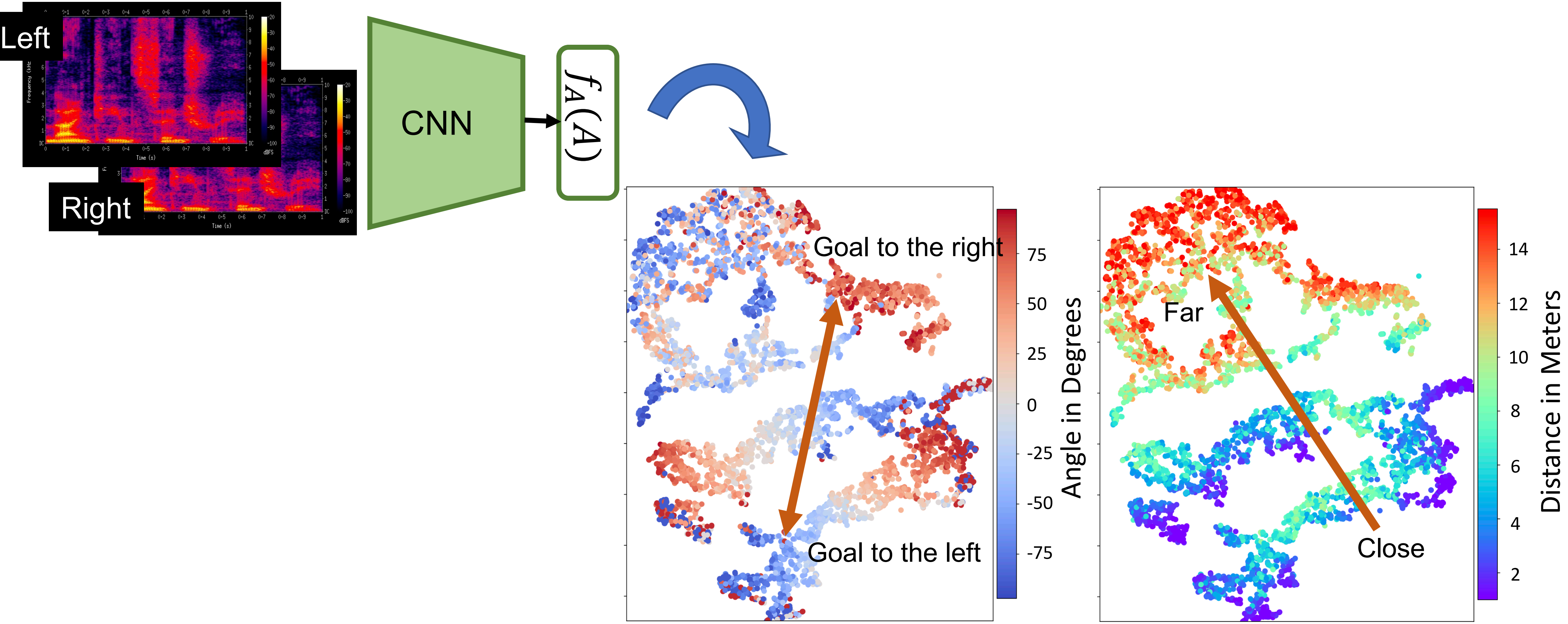
From *same sound* to *varied heard sounds* to *varied unheard sounds*[1]

- AudioGoal accuracy declines with varied heard sounds to unheard sounds
- AudioPointGoal almost always outperforms AudioGoal agent



[1] 102 copyright-free sounds, divided into 73/11/18 for train/val/test

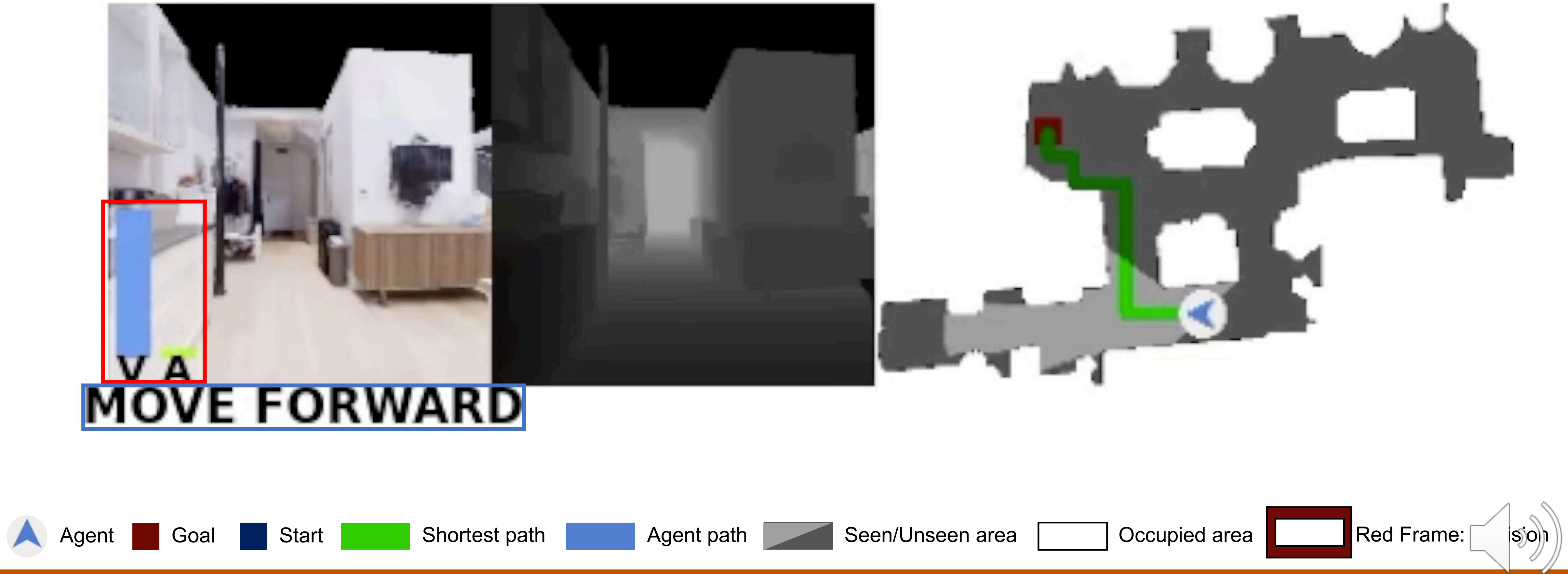
What Do the Learned Audio Features Capture?



T-SNE of audio features from an AudioGoal agent

Relative Importance of Audio and Vision

Each modality plays an important role in action selection, based on the environment context and goal placement



Limitations and Extensions

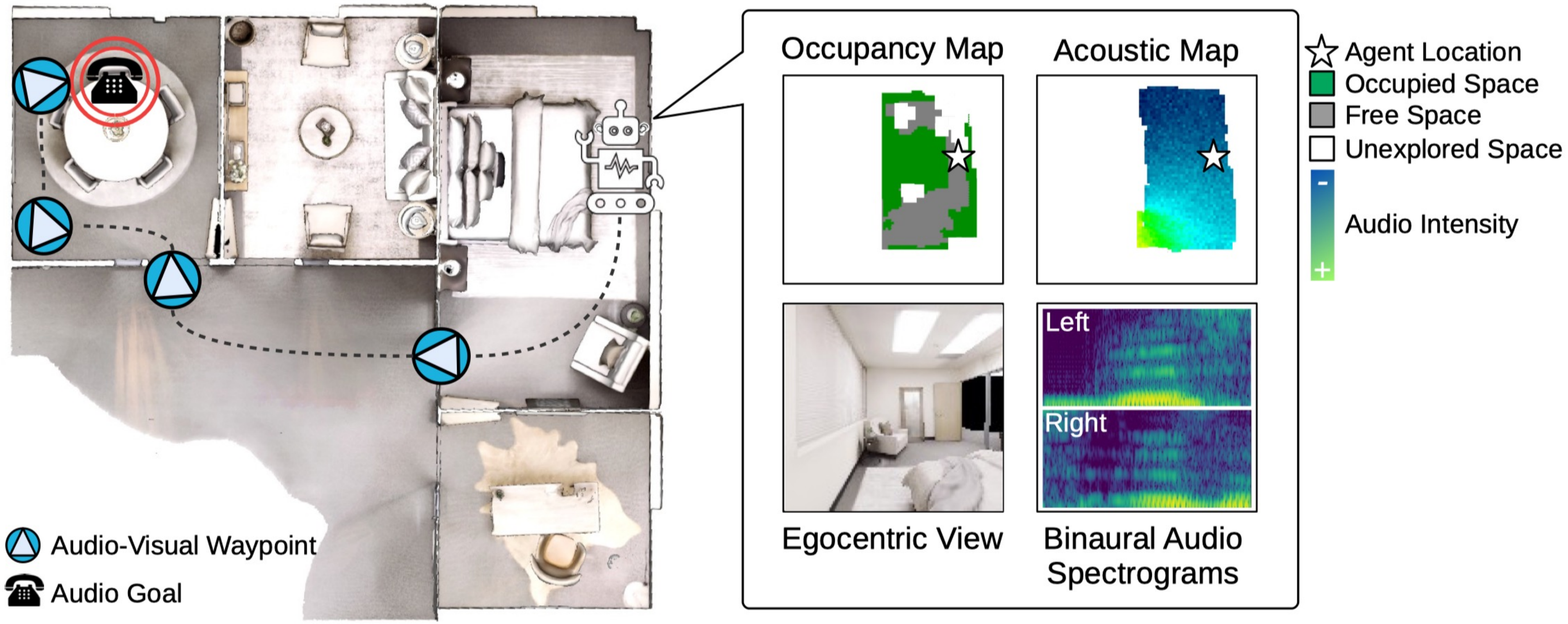
- Step-wise action prediction leads to oscillating behaviors
- Simplified AudioGoal task does not require semantic understanding
- Discuss two extensions:
  - Learning to set waypoints for Audio-Visual Navigation
  - Semantic Audio-Visual Navigation

Learning to Set Waypoints for Audio-Visual Navigation

Changan Chen[1,2], Sagnik Majumder[1], Ziad Al-Halah[1], Ruohan Gao[1,2],
Santhosh Kumar Ramakrishnan[1,2], Kristen Grauman[1,2]

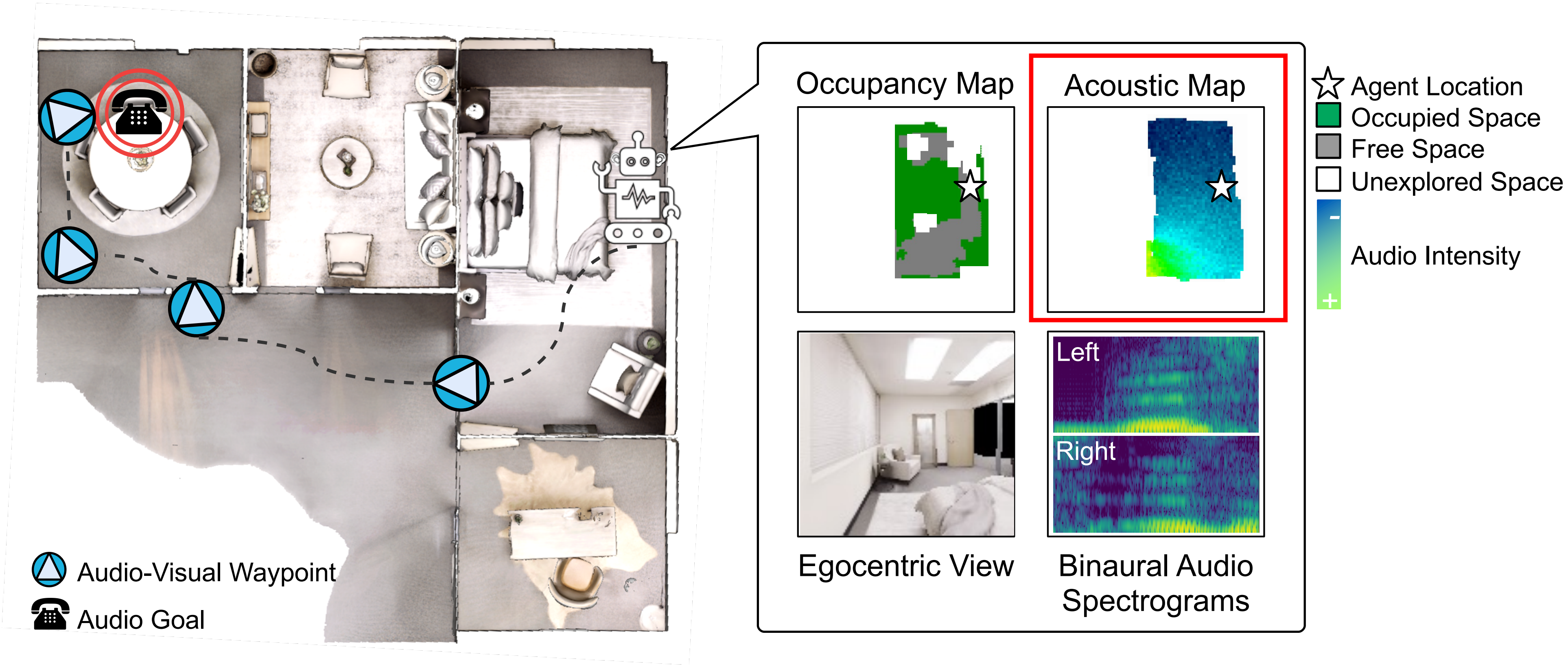
[1]UT Austin, [2]Facebook AI Research

ICLR 2021



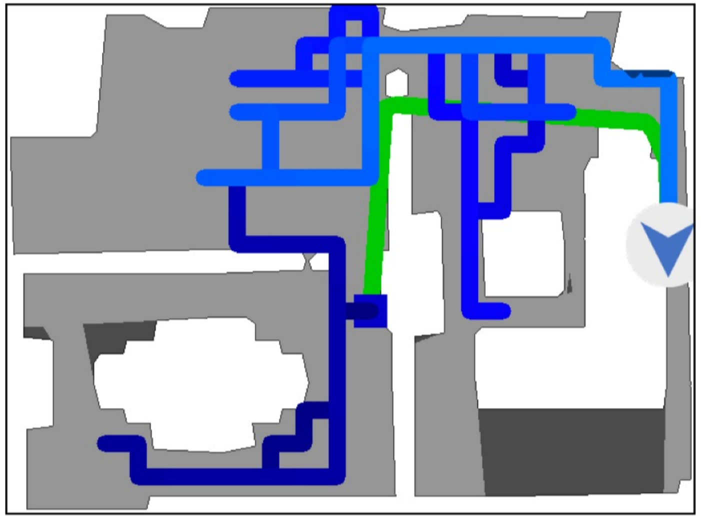
Our Idea

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

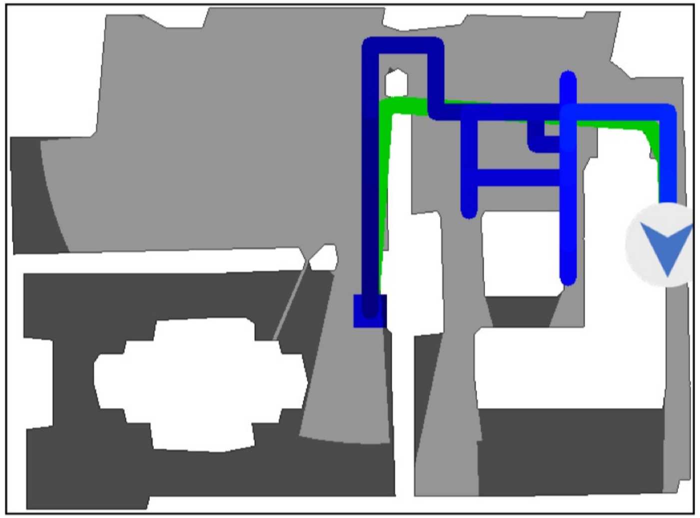


Navigation Trajectories

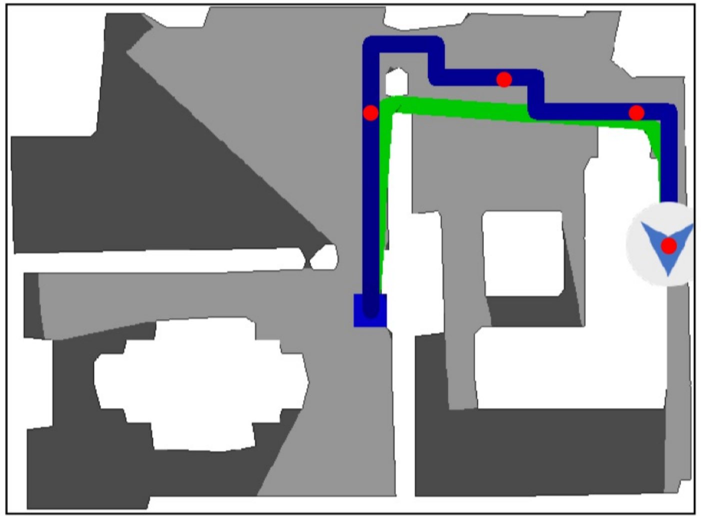
- Gan et al. [ICRA 20]: is prone to errors and often leads the agent to backtrack
- Chen et al. [ECCV20]: oscillates around obstacles
- AV-WaN (Ours): reaches the goal most efficiently



Gan et al. [ICRA20]



Chen et al. [ECCV20]

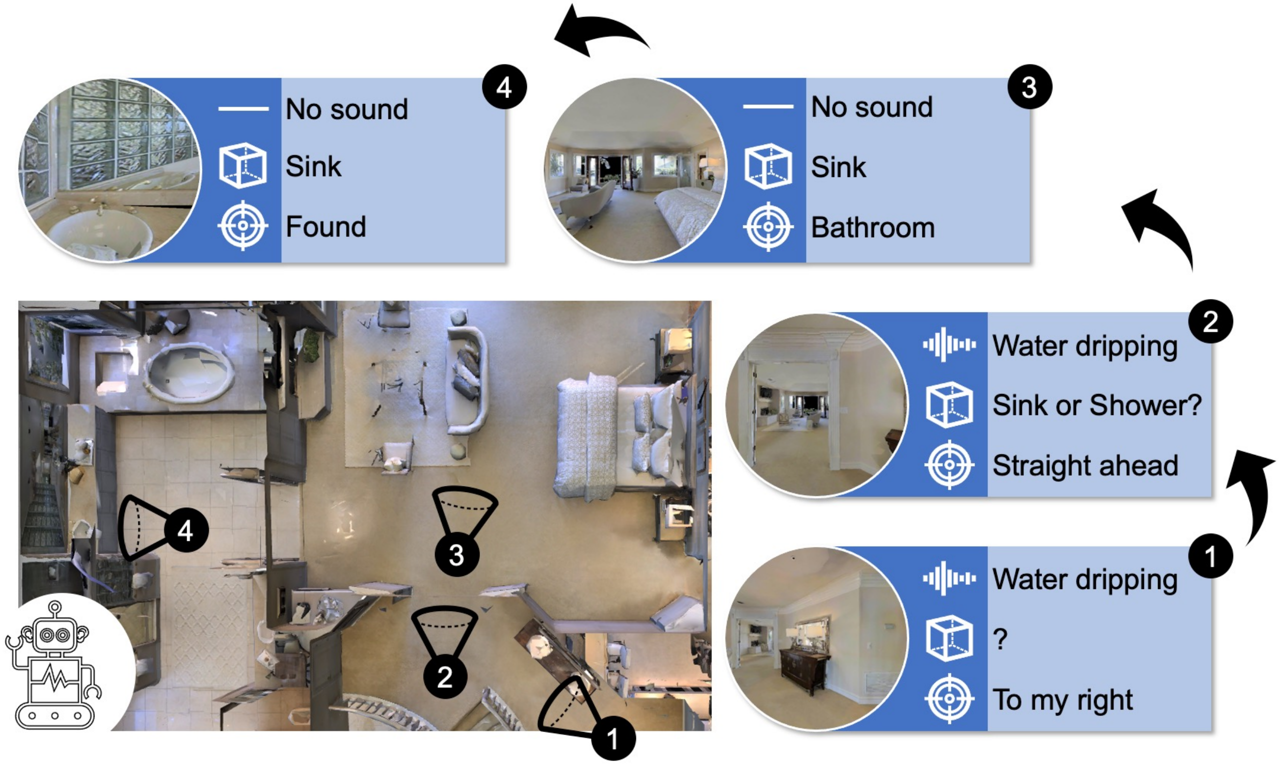


AV-WaN (Ours)

Semantic Audio-Visual Navigation

Changan Chen[1,2], Ziad Al-Halah[1], Kristen Grauman[1,2]
[1]UT Austin,[2]Facebook AI Research

CVPR 2021



AudioGoal Task

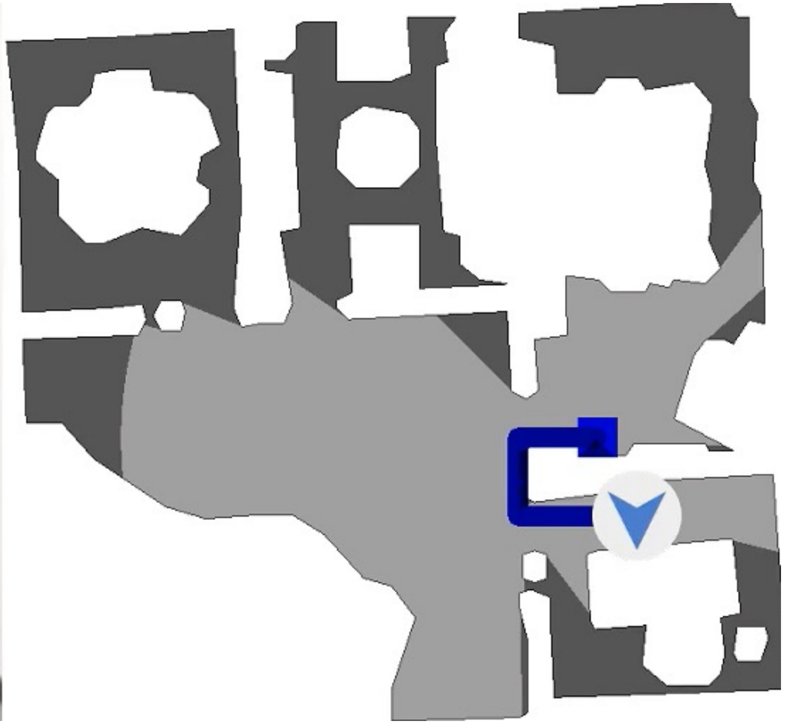
AudioGoal task (Chen et al. ECCV 2020, Gan et al. ICRA 2020):

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

Agent's egocentric view

Telephone not present!

Top-down map

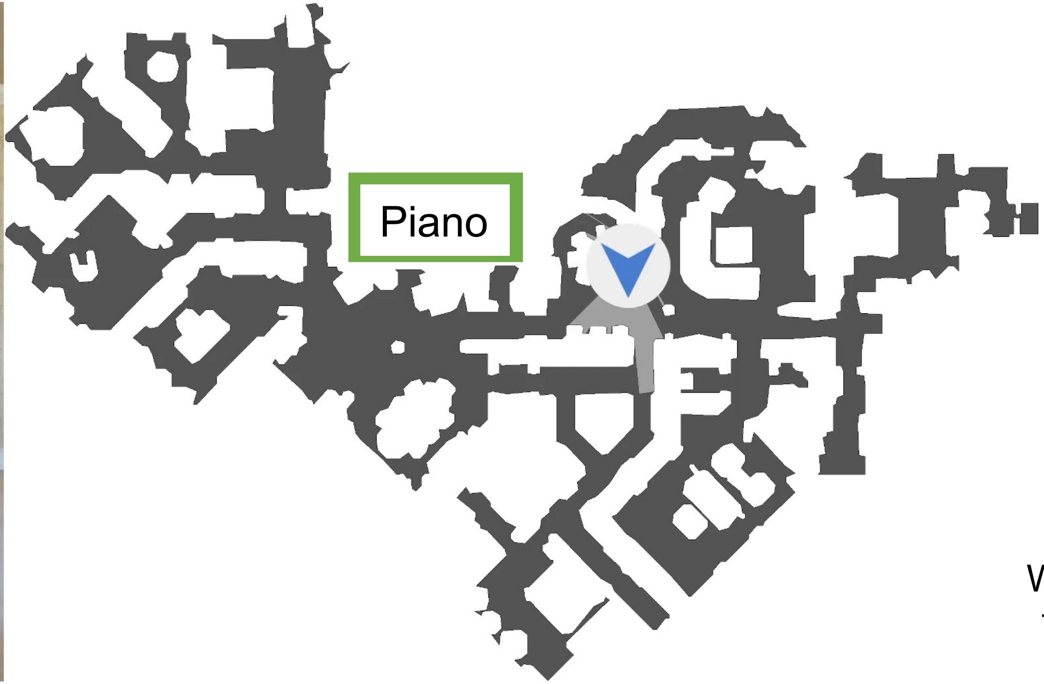


The agent searches for the ringing telephone in an unfamiliar environment

Semantic AudioGoal Task

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