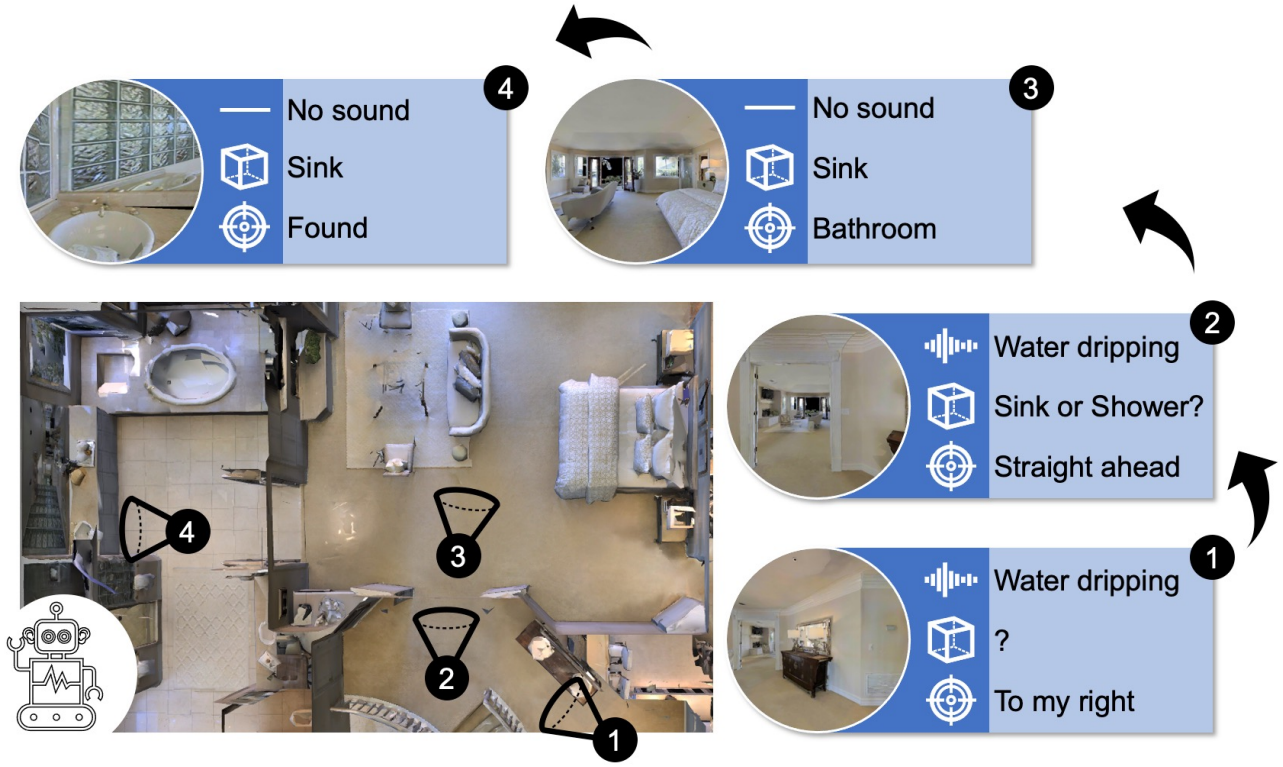
Semantic Audio-Visual Navigation

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

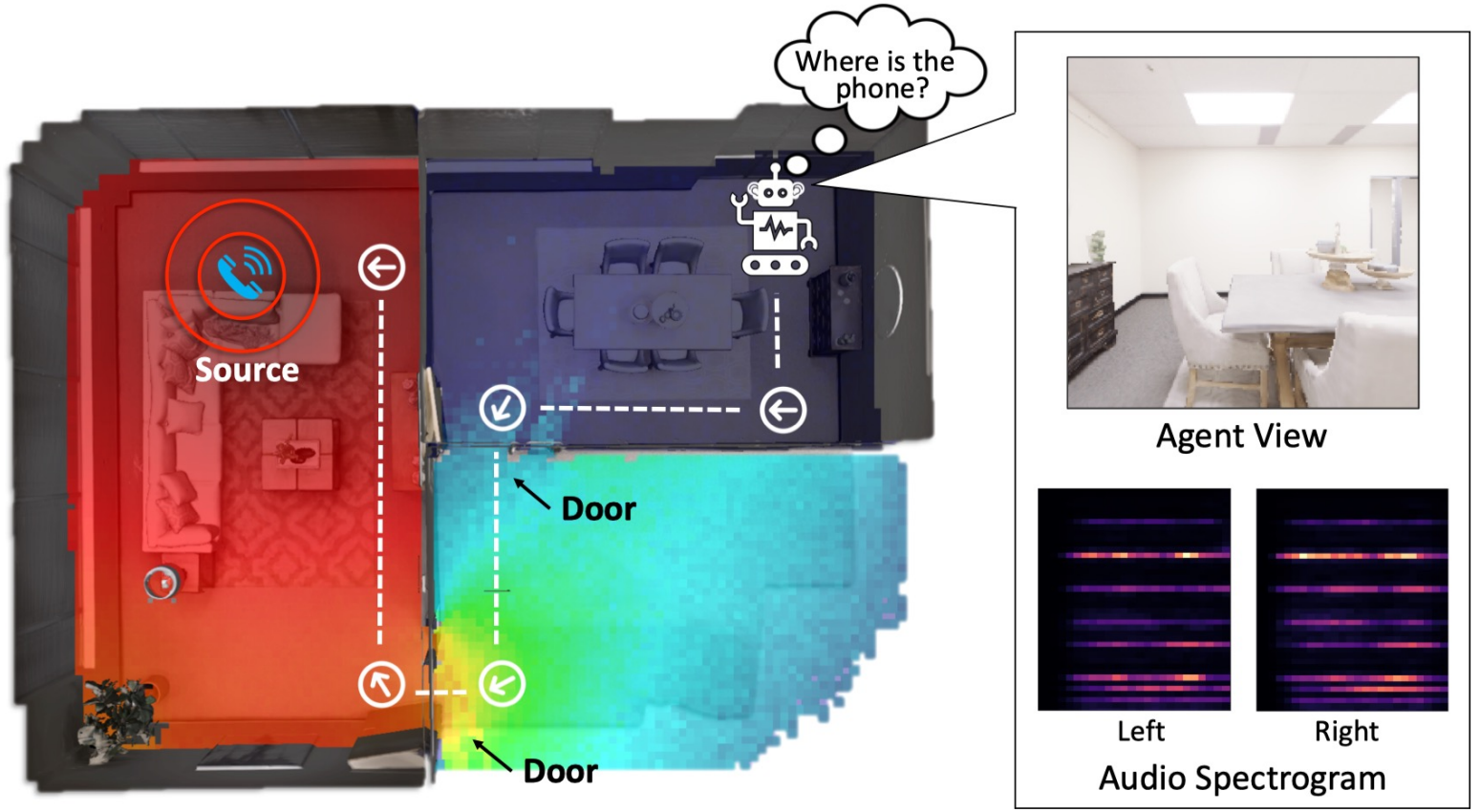
CVPR 2021



Audio-Visual Navigation

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

- Travel to an unknown sounding object in an unmapped environment
- Receive both visual and auditory observations



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

AudioGoal Task

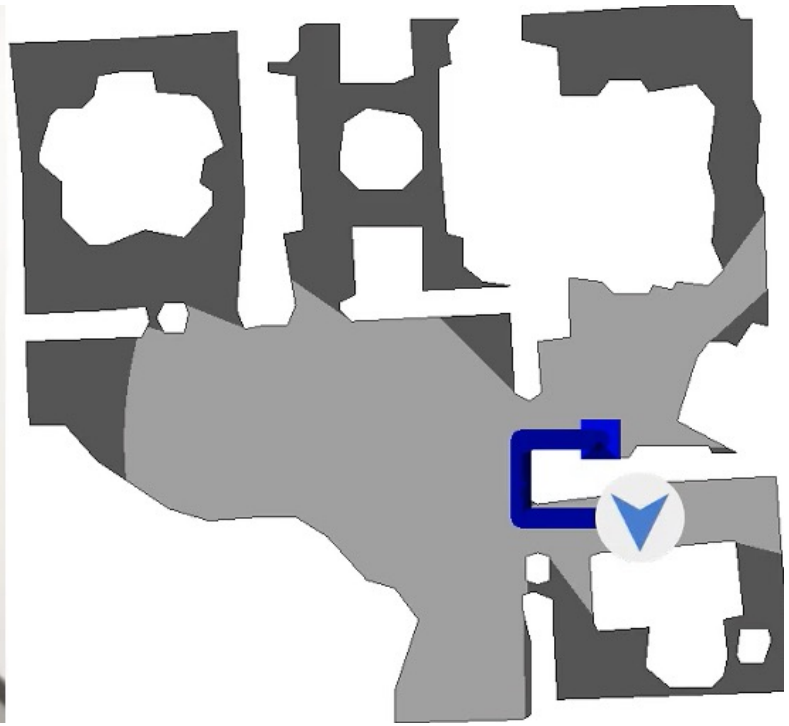
AudioGoal task (Chen et al. ECCV 2020, Gan et al. ICRA 2020):

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

Agent's egocentric view

Top-down map

Telephone not present!

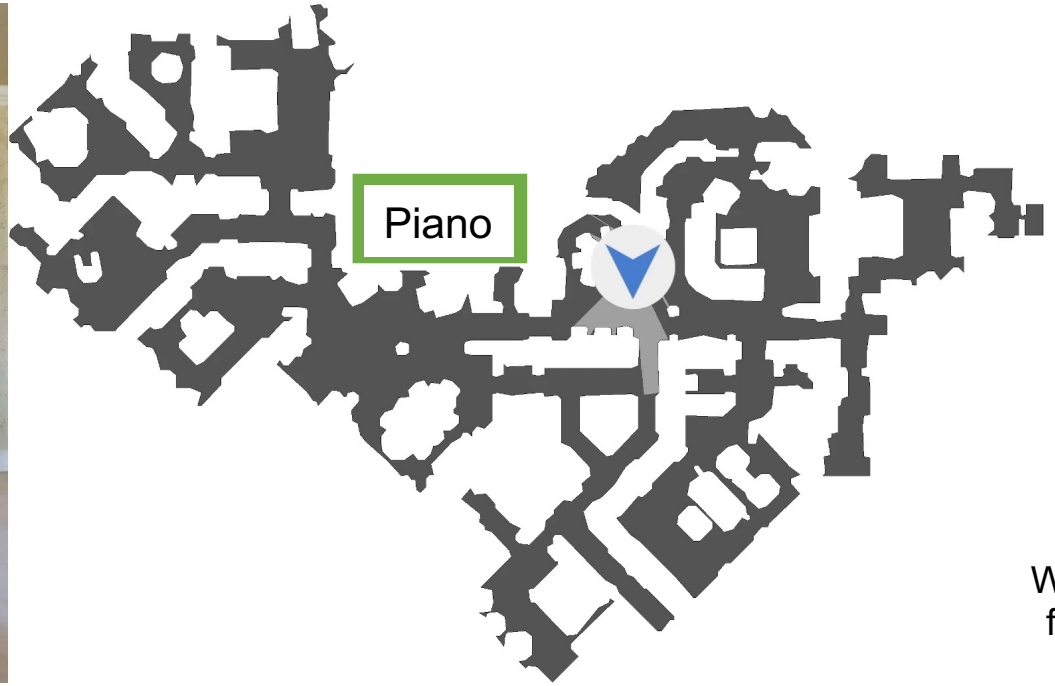


The agent searches for the ringing telephone in an unfamiliar environment

Semantic AudioGoal Task

Agent's egocentric view

Top-down map



Wear headphones
for spatial sound

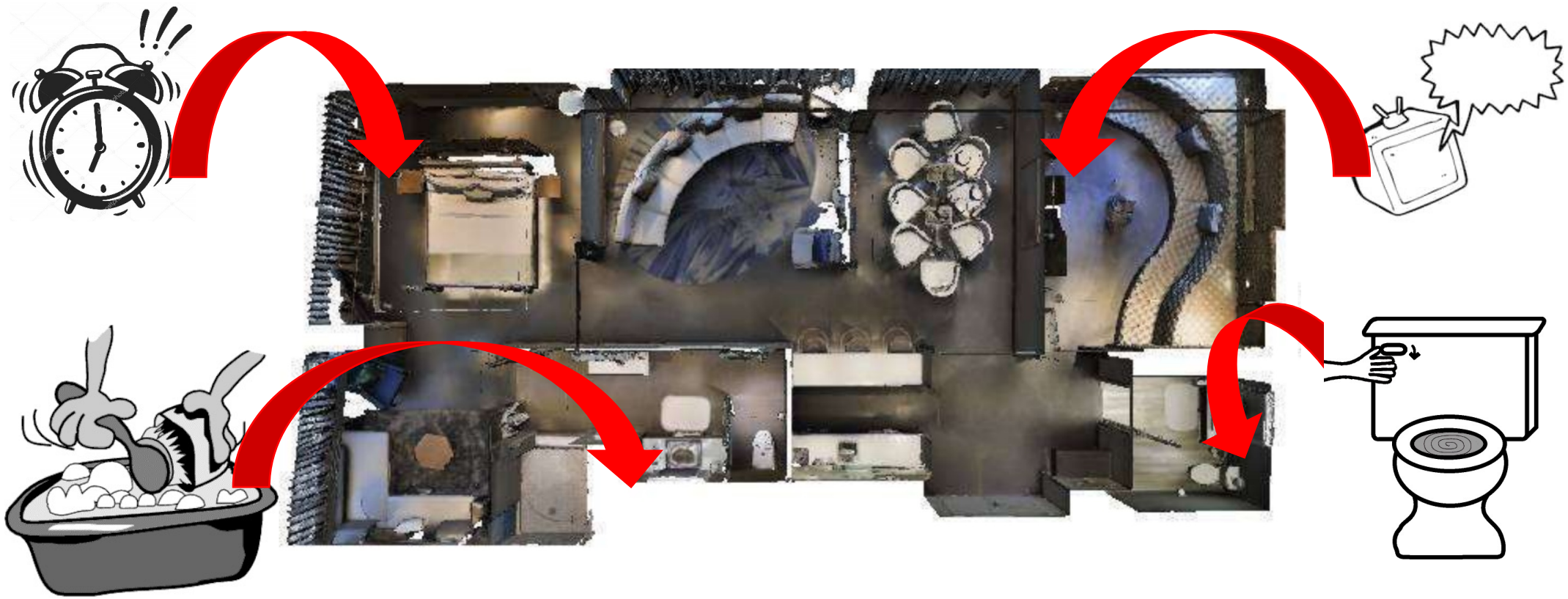
The agent must continue navigating even after the sound stops

Our proposed semantic AudioGoal task:

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

Semantic AudioGoal Dataset

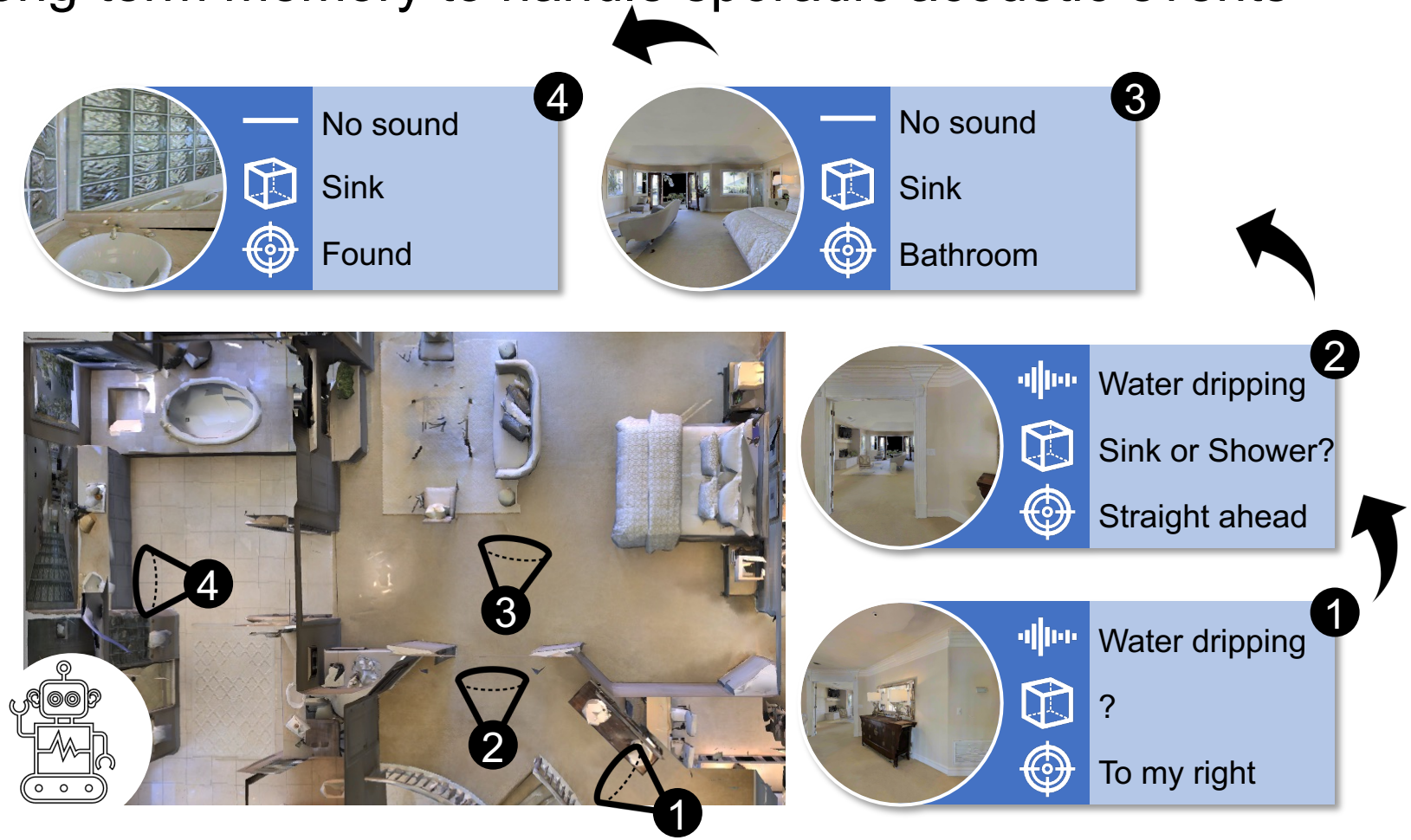
- Augment an existing simulator SoundSpaces[1] with semantic sounds
- 21 object categories in Matterport3D[2]: chair, TV, cabinet, sink etc.
- Object-emitted sounds and object-related sounds



[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

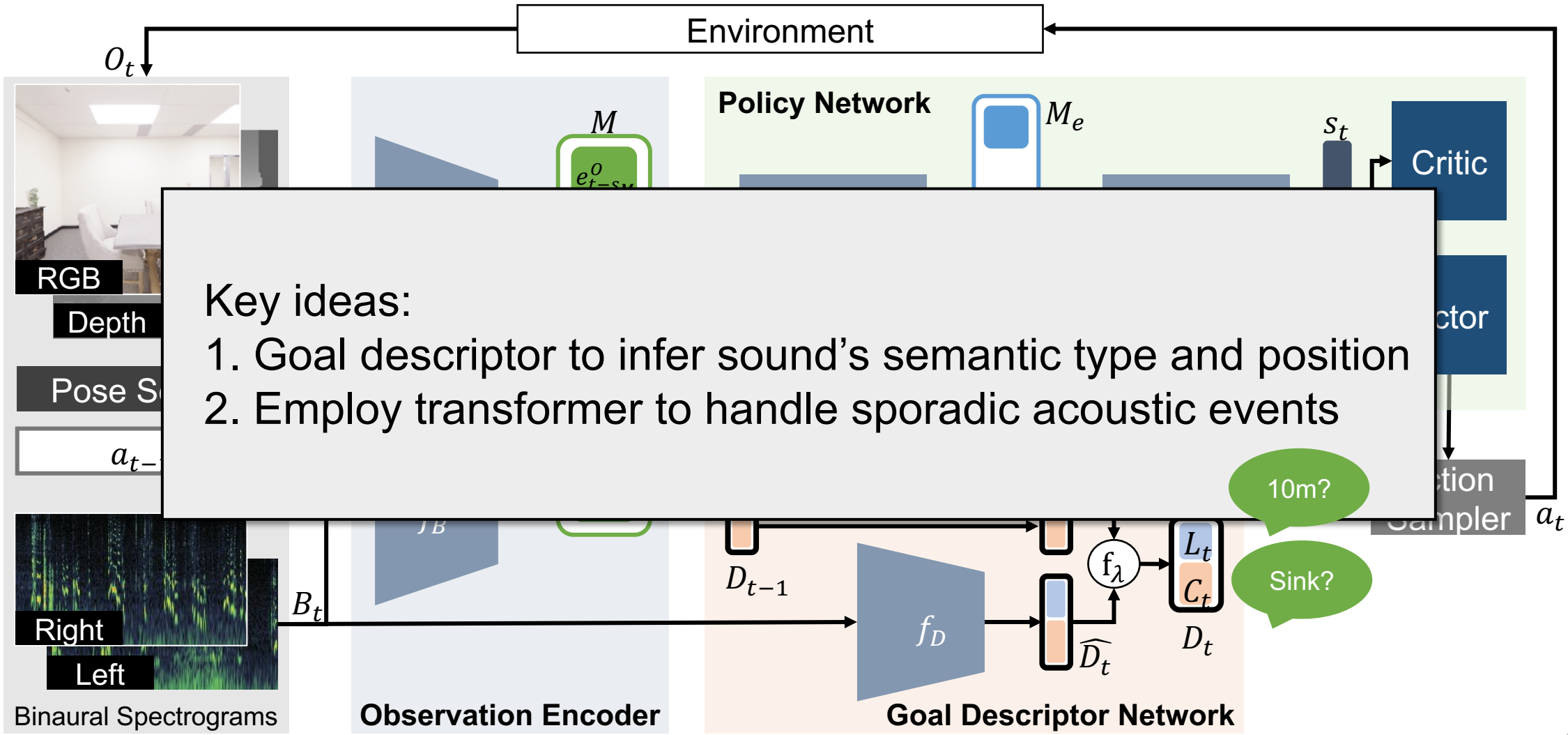
Our Idea

- 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 (SAVi)

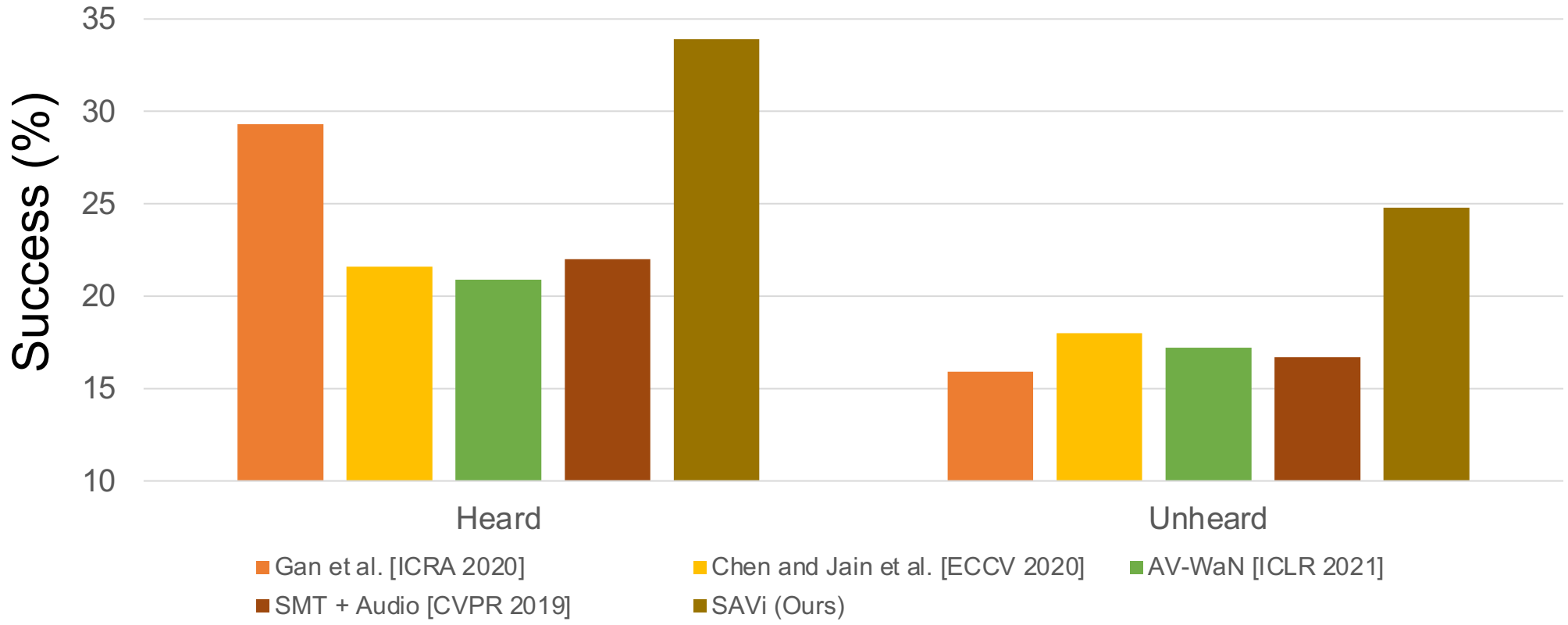


Key ideas:
1. Goal descriptor to infer sound's semantic type and position
2. Employ transformer to handle sporadic acoustic events

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

Navigation Results

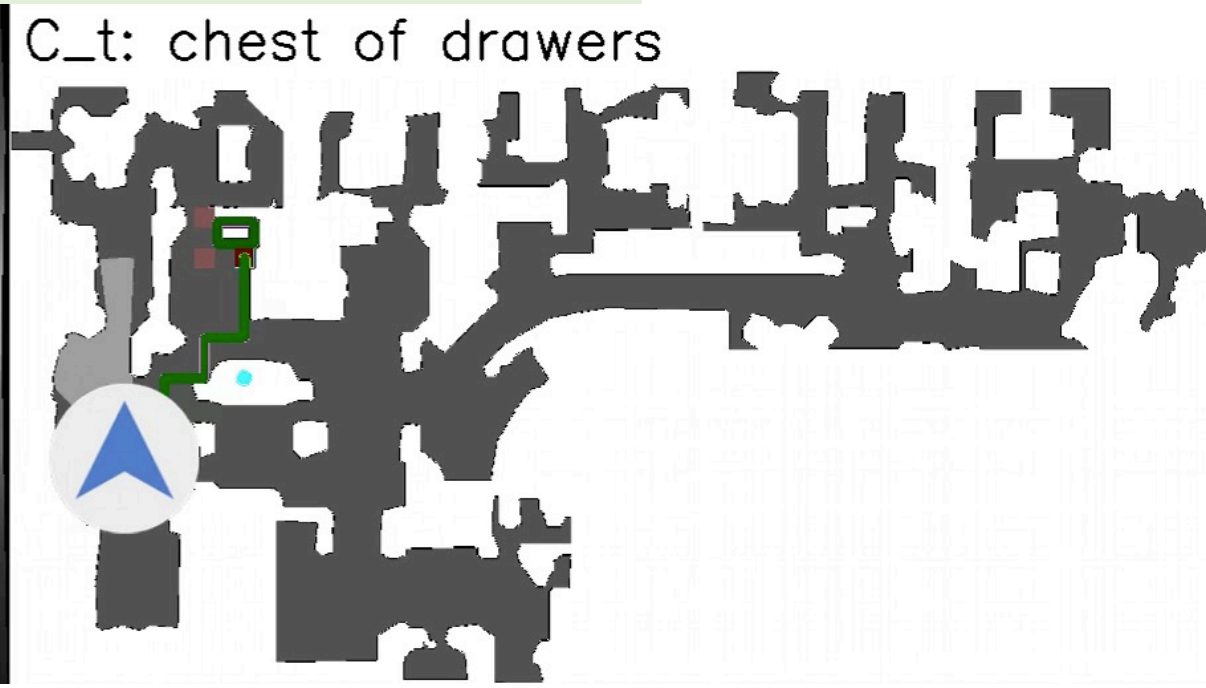
- SAVi strongly outperforms all existing methods
- Generalizing to unheard sounds is much more challenging



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 sound stops before the agent reaches the goal, but the agent identifies its chest-of-drawers sound and uses its visual perception to locate the target object

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

Code and audio simulation data available at:
http://vision.cs.utexas.edu/projects/semantic-audio-visual-navigation

SoundSpaces Challenge at Embodied AI Workshop, CVPR 2021:
https://soundspaces.org/challenge