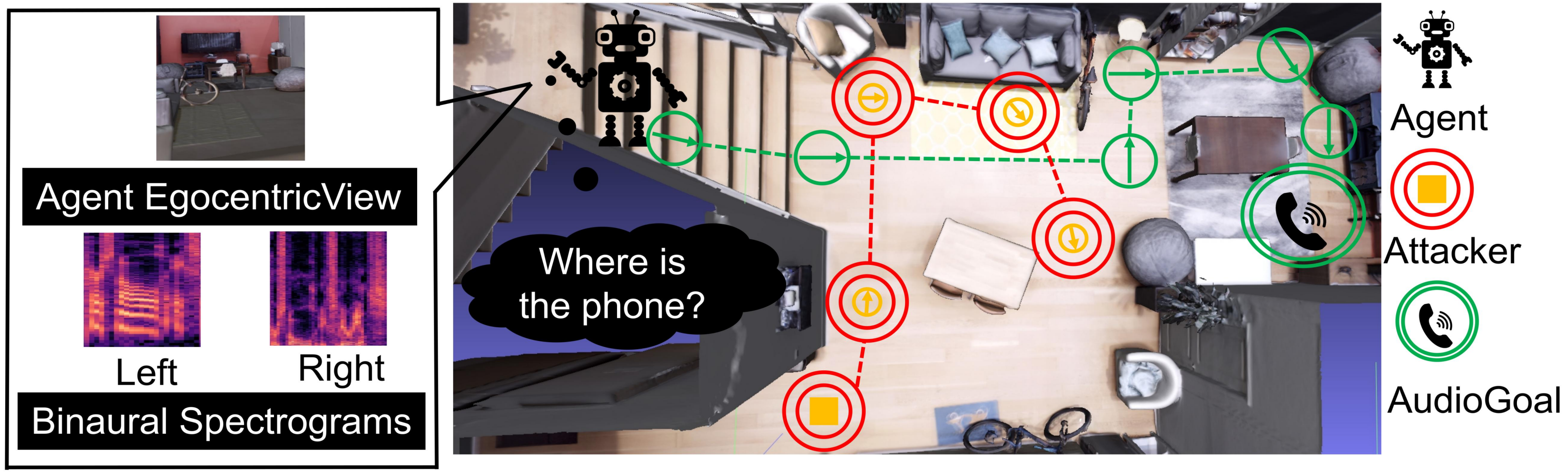
Sound Adversarial Audio-Visual Navigation

Yinfeng Yu[1,3], Wenbing Huang[2], Fuchun Sun[*1],

Changan Chen[4], Yikai Wang[1,5], Xiaohong Liu[1]

[1]Beijing National Research Center for Information Science and Technology (BNRist),
State Key Lab on Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University
[2]Institute for AI Industry Research (AIR), Tsinghua University
[3]College of Information Science and Engineering, Xinjiang University
[4]UT Austin [5] JD Explore Academy, JD.com



This work aims to do an adversarial sound intervention for robust audio-visual navigation.

∗ Corresponding author: Fuchun Sun.

Motivation

SoundSpaces[1] is focus on audio-visual navigation problem in the acoustically clean or simple environment.

However, there are many situations different from the setting of SoundSpaces, which there are some non-target moving sounding objects in the scene:
For example, a kettle in the kitchen beeps to tell the robot that the water is boiling, and the robot in the living room needs to navigate to the kitchen and turnoff the stove; while in the living room, two children are playing a game, chuckling loudly from time to time.

The limitation of existing navigation agent/training environments:
they cannot model non-target moving sounding objects.

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

Challenges

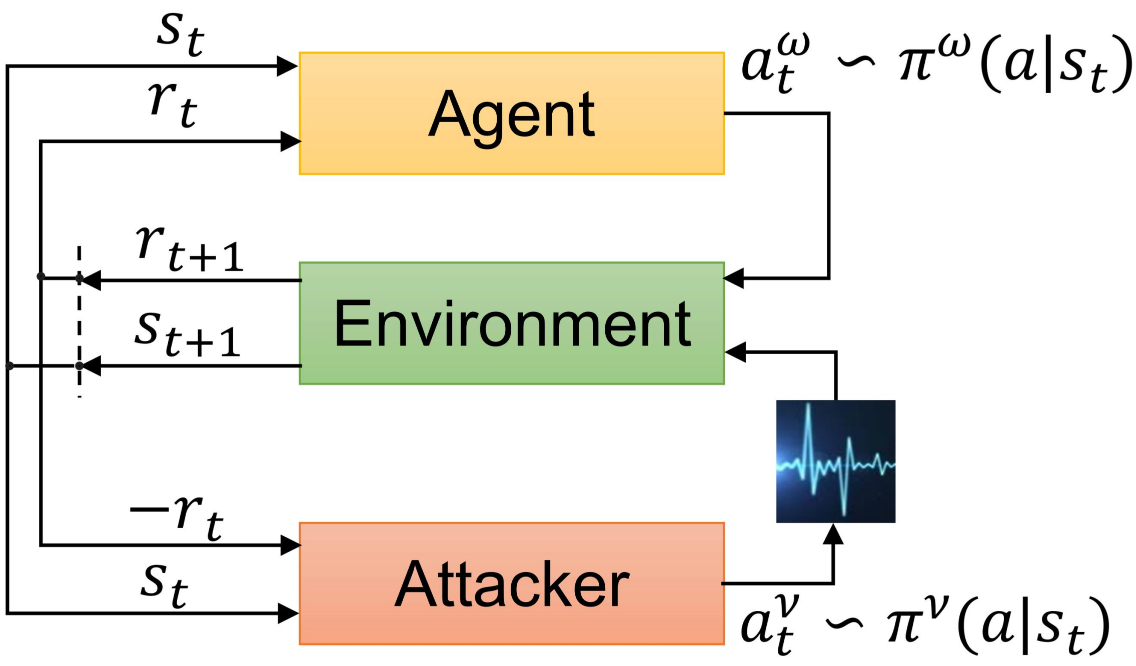
Challenge 1:
How to model non-target moving sounding objects in a simulator or reality? There is no such setting that existed!

Challenge 2:
Can an agent still find its way to the destination in an acoustically complex environment?

Our ideas

How to model?

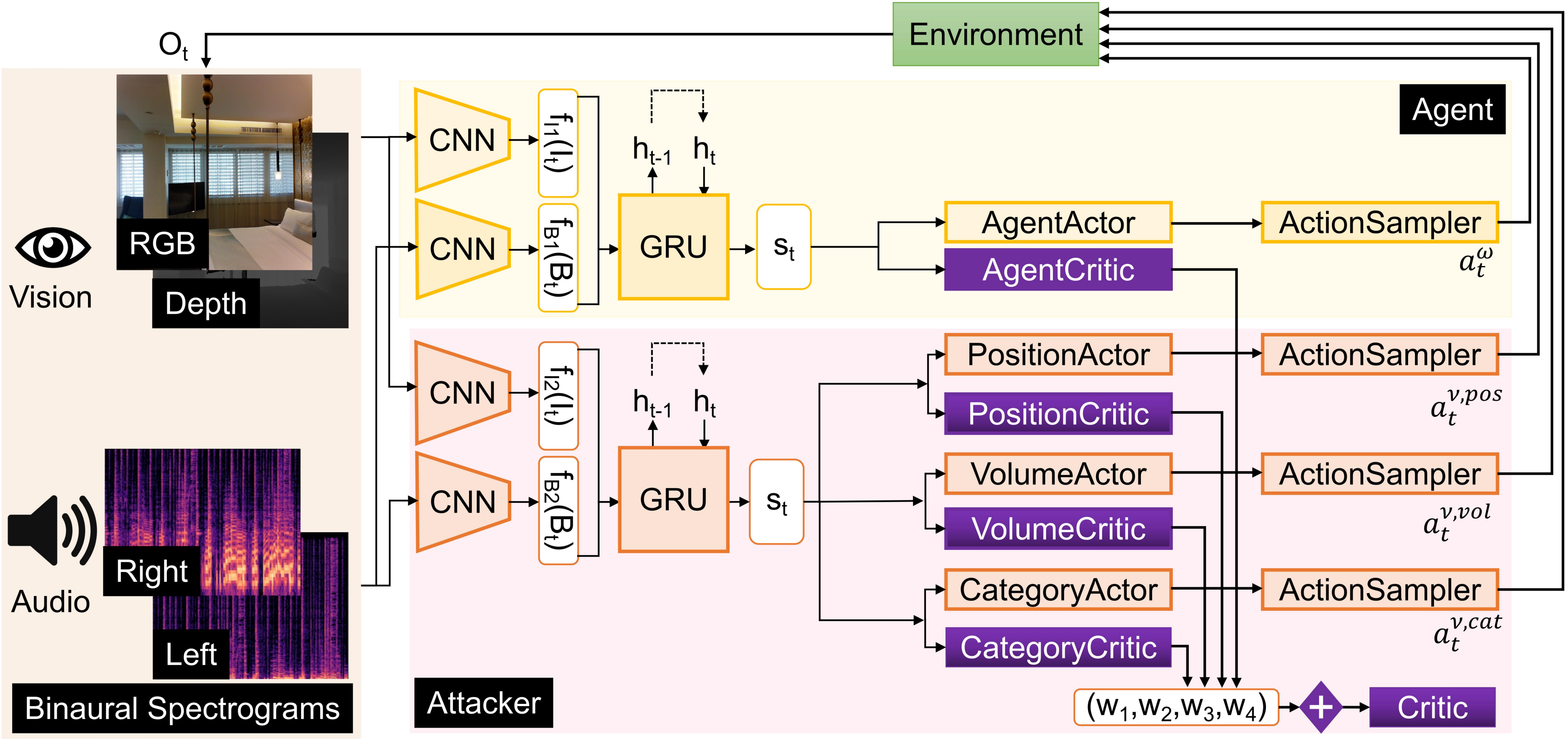
- **Worst case strategy**: Regard non-target sounding objects as deliberately embarrassing the robot. We called them sound attackers.
- **Simplify**: Only consider the simplest situation, one sound attacker.
- **Zero-sum game**: One agent, one sound attacker.



Contributions

- We construct a sound attacker for audio-visual navigation that aims to improve the agent's robustness.
- We develop a joint training paradigm for the agent and the attacker.

Neural network model



The agent and the sound attacker first encode observations and learn state representation s_t respectively. Then, s_t are fed to actor-critic networks, which predict the next action a_t^{ω} and a_t^{ν}.

Experiment Setup

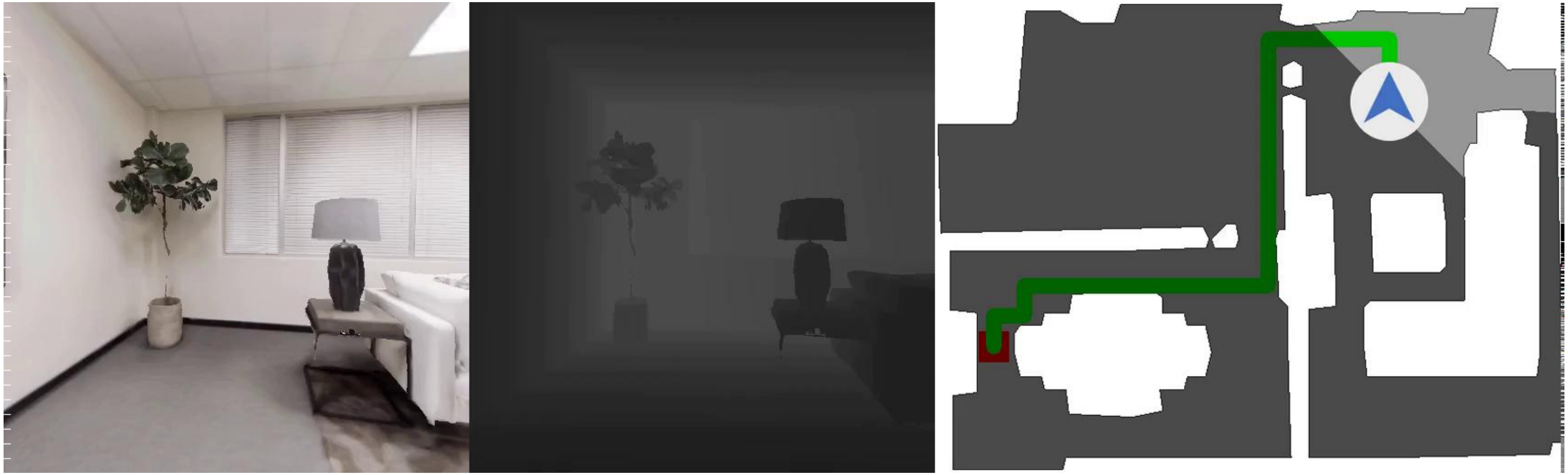
Baselines:
- **Random**: A random policy that uniformly samples one of three actions.
- **AVN**[1]: An audiovisual embodied navigation trained in an environment without sound intervention.
- **SA-MDP**[2]: We adopt its idea but only intervene state of the sound input and do not process the visual information.

Metrics :
- **SPL:** Success weighted by Path Length
- **R_{mean}:** stands for average episode reward of agent
- **SSPL:** Soft Success weighted by Path Length
- **SR:** Success Rate
- **DTG:** average Distance To Goal
- **NDTG:** Normalized average Distance To Goal

[1] C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020
[2] H. Zhang, H. Chen, et al., Robust deep reinforcement learning against adversarial perturbations on state observations, NeurIPS 2020

Demos

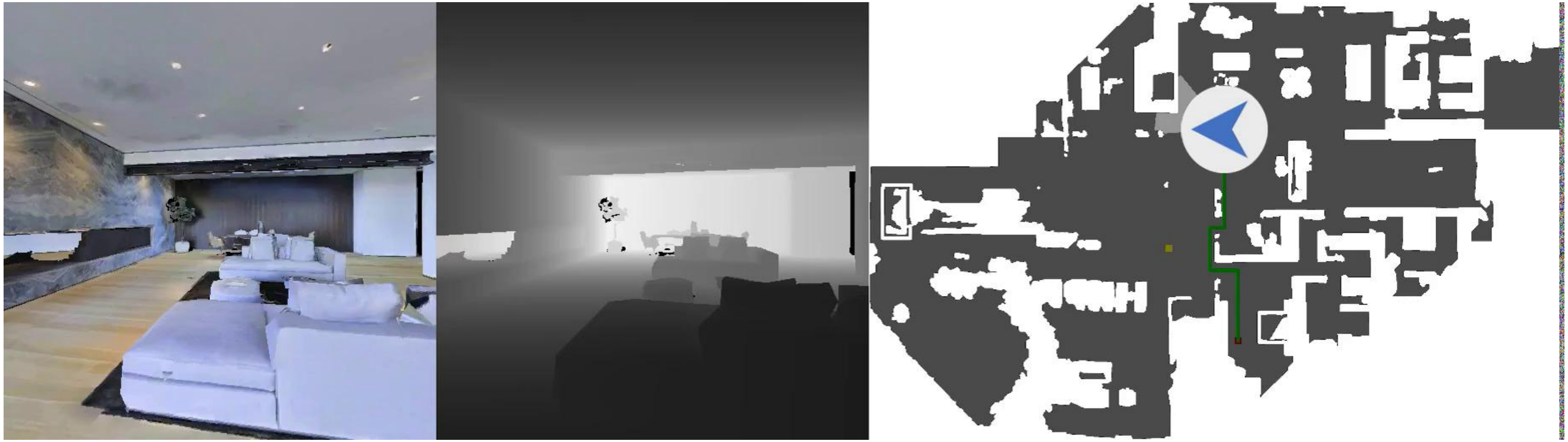
This is a demo on Replica in the clean environment.



4x Speed

Demos

This is a demo on Matterport3D in the complex environment.



4x Speed

Experiment Results

Our model is better than others in both clean and complex environments from the following tables .

Performance comparison of different models on Replica.

Method	Clean env.	PVC.
Random	0.000/-4.7	0.000/-4.5
AVN	0.721/15.1	0.389/8.0
SA-MDP	0.590/10.2	0.368/7.2
SAAVN	**0.742/16.6**	**0.552/10.6**

Performance comparison of different models on Matterport3D.

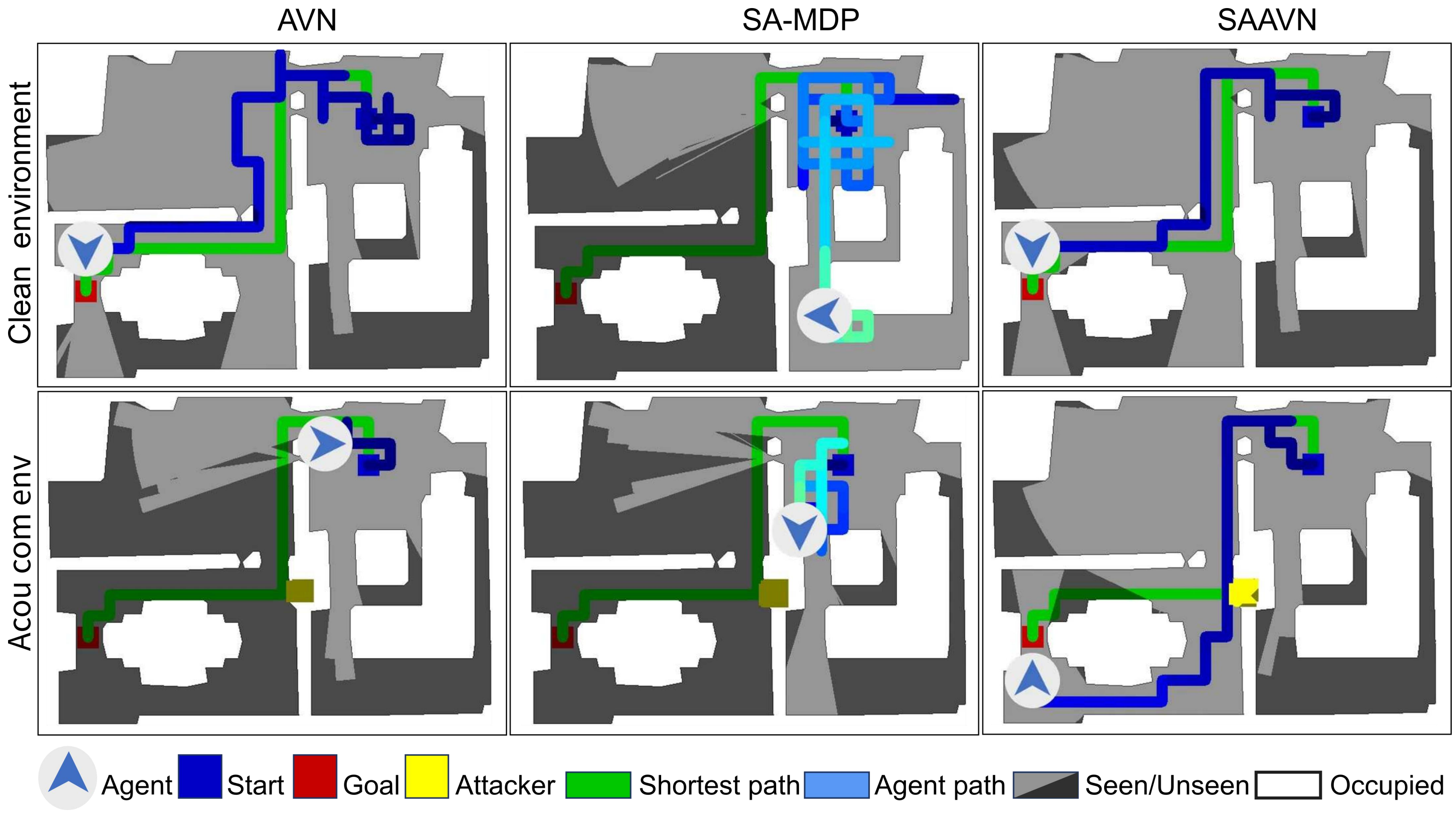
Method	Clean env.	PVC.
Random	0.000/-5.0	0.000/-5.0
AVN	0.539/18.1	0.397/15.3
SAAVN	**0.549/18.7**	**0.478/17.3**

an acoustically complex environment

Comparison of different models under SPL (↑)/R_{mean} (↑) metrics.

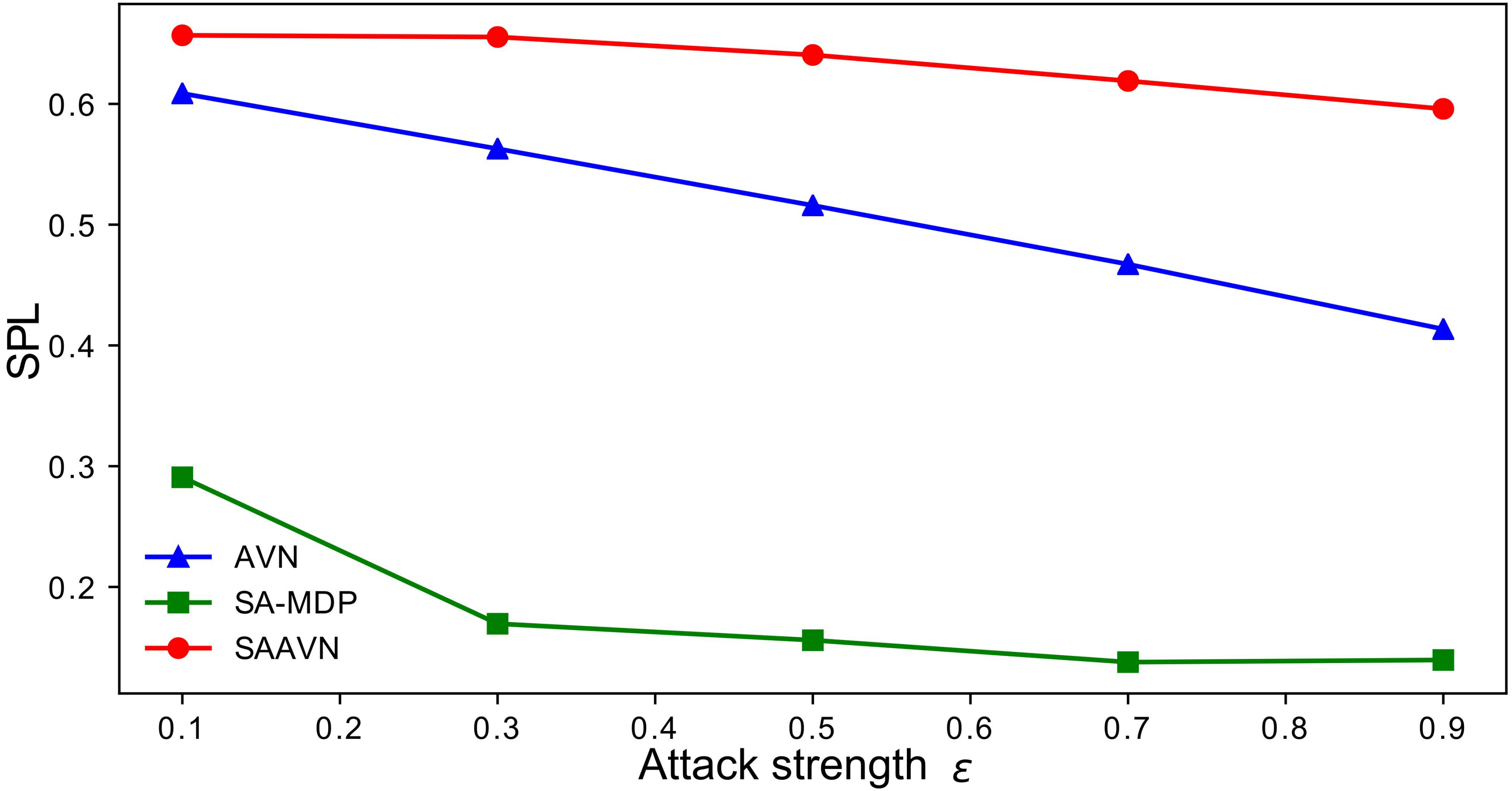
Navigation Trajectories

- AVN: can complete the task in a clean environment but fails in an acoustically complex environment.
- SA-MDP: was unable to complete the job successfully in all two settings.
- SAAVN (ours): reaches the goal most efficiently.



Robustness

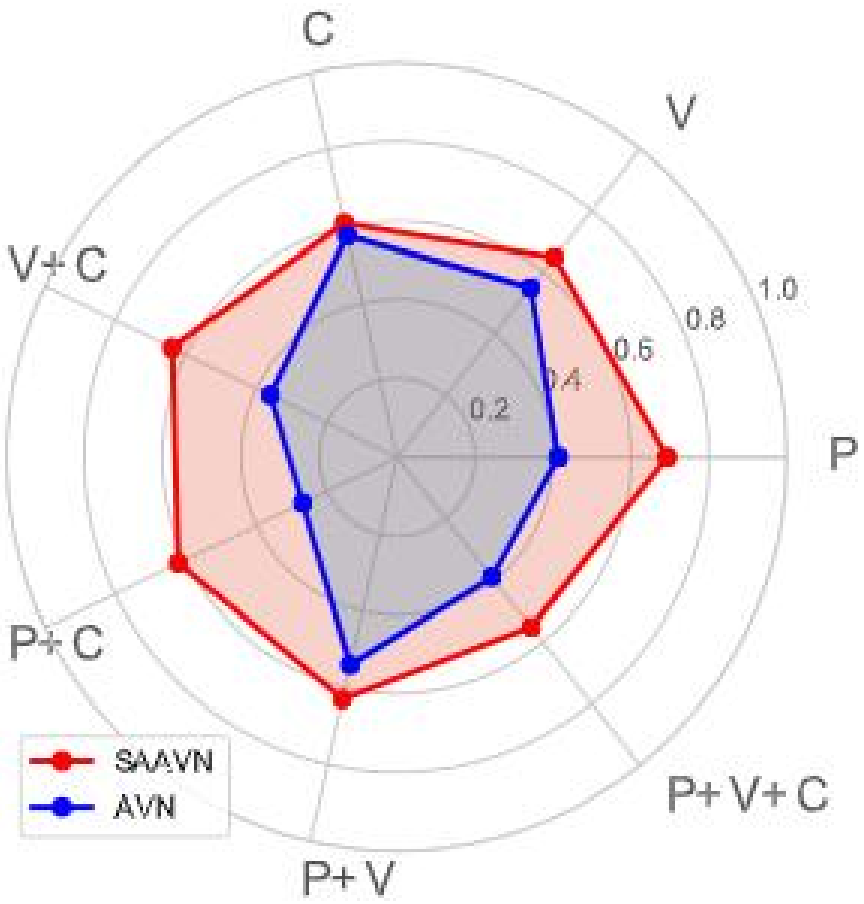
Our method's performance decreases more slowly under different attack strength, which fully demonstrates that our approach helps improve the robust performance of the algorithm.



Ablation study

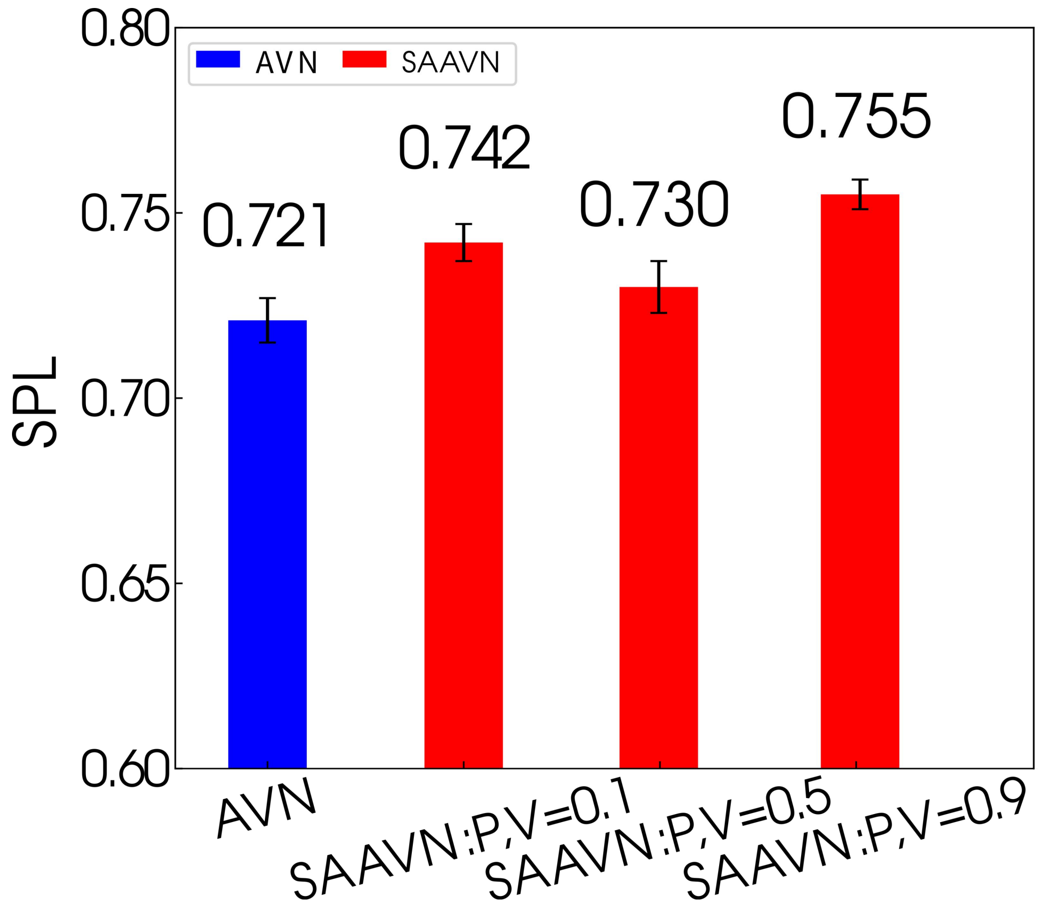
Ablation study for the attack policy regarding the position, volume, and sound category.

Our model's performance is better than AVN in all environments.



Ablation study

The relationship between SPL and the volume of the sound attacker is not straightforward. It depends on other factors, including the position and sound category.



Conclusion

- We proposed a game with a sound attacker.
- We designed various policies for the sound attack.
- We rationally model the gap between research and application.

For more details, please refer to the original paper.

Sound Adversarial Audio-Visual Navigation

Yinfeng Yu[1,3], Wenbing Huang[2], Fuchun Sun[*1],
Changan Chen[4], Yikai Wang[1,5], Xiaohong Liu[1]

The project and code can be viewed at the following website:

https://yyf17.github.io/SAAVN