# SoundSpaces: Audio-Visual Navigation in 3D Environments

Changan Chen*[1,4], Unnat Jain*†[2,4], Carl Schissler[3], Sebastia Vicenc Amengual Gari[3], Ziad Al-Halah[1], Vamsi Krishna Ithapu[3], Philip Robinson[3], and Kristen Grauman[1,4]

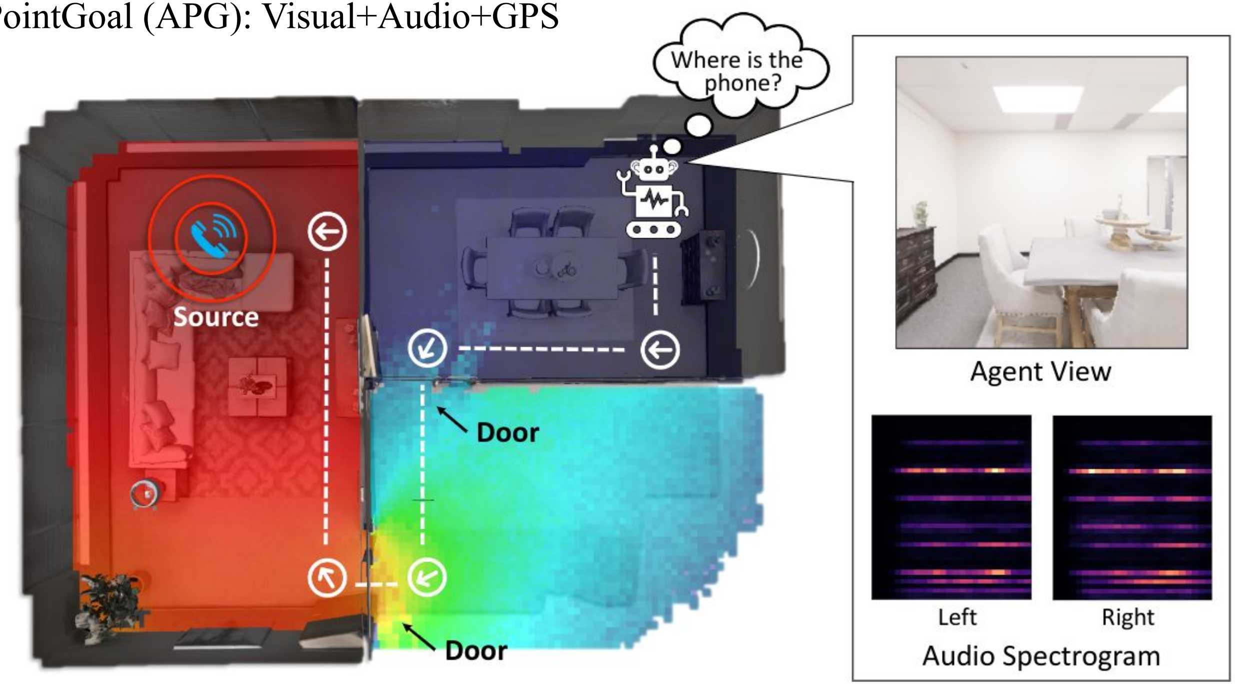
[1]UT Austin, [2]UIUC, [3]Facebook Reality Labs, [4]Facebook AI Research

ECCV 2020

https://arxiv.org/pdf/1912.11474.pdf

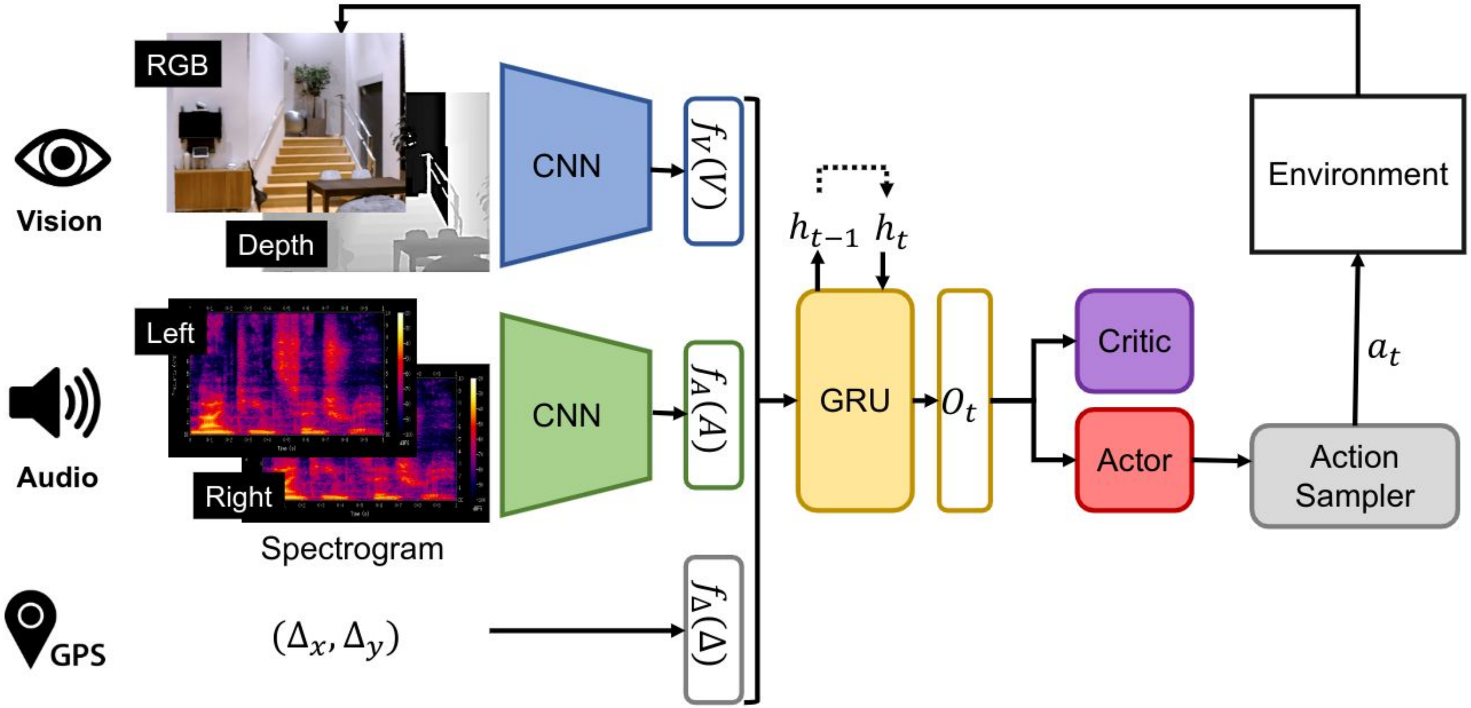
# Audio-Visual Navigation

- AudioGoal (AG): Visual+Audio
- AudioPointGoal (APG): Visual+Audio+GPS




https://arxiv.org/pdf/1912.11474.pdf

# Pipeline




https://arxiv.org/pdf/1912.11474.pdf

# Audio helps navigation

SPL:

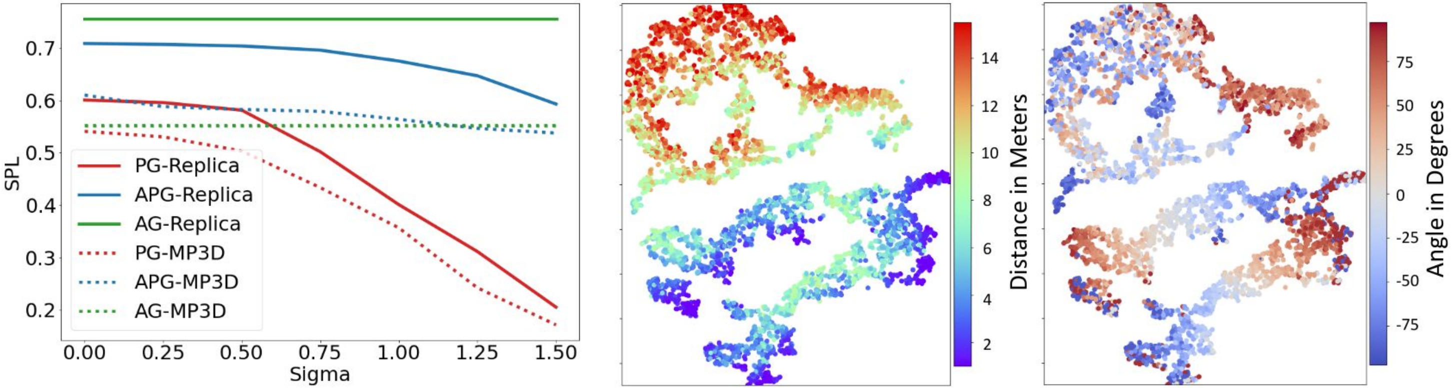
$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_i}{\max(p_i, \ell_i)}. \quad (1)$$

Table 2: Adding sound to sight and GPS sensing improves navigation performance significantly. Values are success rate normalized by path length (SPL); higher is better.

| | | Replica | | Matterport3D | |
|---|---|---|---|---|---|
| | | PointGoal | AudioPointGoal | PointGoal | AudioPointGoal |
| Baselines | Random | 0.044 | 0.044 | 0.021 | 0.021 |
| | Forward | 0.063 | 0.063 | 0.025 | 0.025 |
| | Goal follower | 0.124 | 0.124 | 0.197 | 0.197 |
| Varying visual sensor | Blind | 0.480 | **0.681** | 0.426 | **0.473** |
| | RGB | 0.521 | **0.632** | 0.466 | **0.521** |
| | Depth | 0.601 | **0.709** | 0.541 | **0.581** |

https://arxiv.org/pdf/1912.11474.pdf

# Audio gives similar or even better cues than displacements




(a) From perfect to noisy GPS

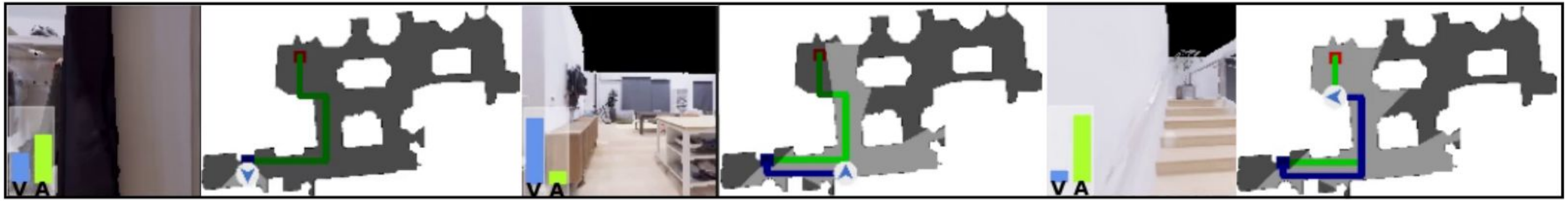
(b) t-SNE of AudioGoal audio features

Fig. 5: **Audio as a learned spatial sensor.** (a) Navigation accuracy with increasing GPS noise. Unlike existing PointGoal agents, our AudioGoal agent does not rely on GPS, and hence is immune to GPS noise. (b) t-SNE projection of audio features, color coded to reveal their correlation with the goal location (left) and direction (right), *i.e.*, source is far (red) or near (violet), and to the left (blue) or right (red) of the agent.

https://arxiv.org/pdf/1912.11474.pdf

# Impact of each modality on action selection


Turn Left Turn Right Stop




Turn Left Move Forward Turn Right

https://arxiv.org/pdf/1912.11474.pdf

# Effect of different sound sources

- same sound: sound source of ‘telephone’ test in unseen scenes.
- varied heard sounds: 78 sounds heared in training and test in unseen scenes.
- varied unheard sounds: 18 unheared sounds test in unseen scenes.

Table 3: Navigation performance (SPL) when generalizing to unheard sounds. Higher is better. Results are averaged over 7 test runs; all standard deviations are $\leq 0.01$.

| Dataset | | | *Same sound* | | *Varied heard sounds* | | *Varied unheard sounds* | |
|---|---|---|---|---|---|---|---|---|
| | | *PG* | *AG* | *APG* | *AG* | *APG* | *AG* | *APG* |
| Replica | Blind | 0.480 | 0.673 | 0.681 | 0.449 | 0.633 | 0.277 | 0.649 |
| | RGB | 0.521 | 0.626 | 0.632 | 0.624 | 0.606 | 0.339 | 0.562 |
| | Depth | 0.601 | 0.756 | 0.709 | 0.645 | 0.724 | 0.454 | 0.707 |
| Matterport3D | Blind | 0.426 | 0.438 | 0.473 | 0.352 | 0.500 | 0.278 | 0.497 |
| | RGB | 0.466 | 0.479 | 0.521 | 0.422 | 0.480 | 0.314 | 0.448 |
| | Depth | 0.541 | 0.552 | 0.581 | 0.448 | 0.570 | 0.338 | 0.538 |

https://arxiv.org/pdf/1912.11474.pdf

# 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

https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.pdf

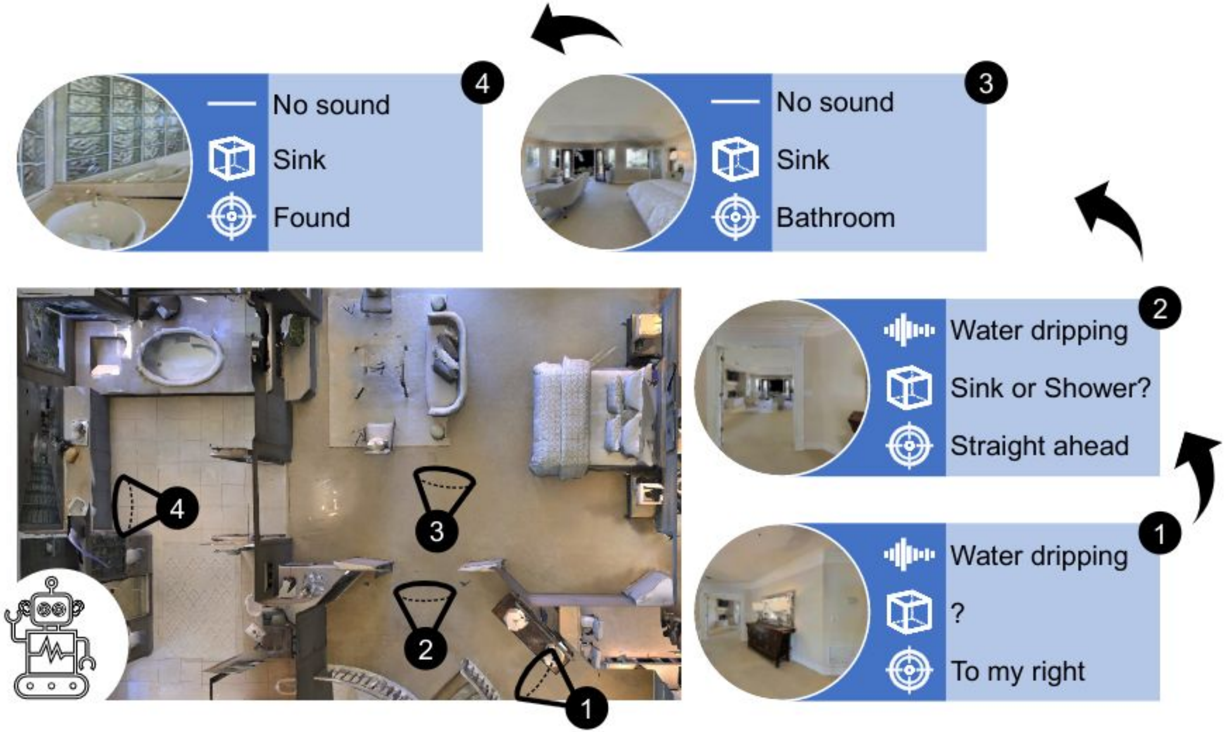
# Limitation

1. Prior work assumes the target object constantly makes a steady repeating sound.

2. In current realistic 3D environment simulators, the sound emitting target has neither a visual embodiment nor any semantic context.

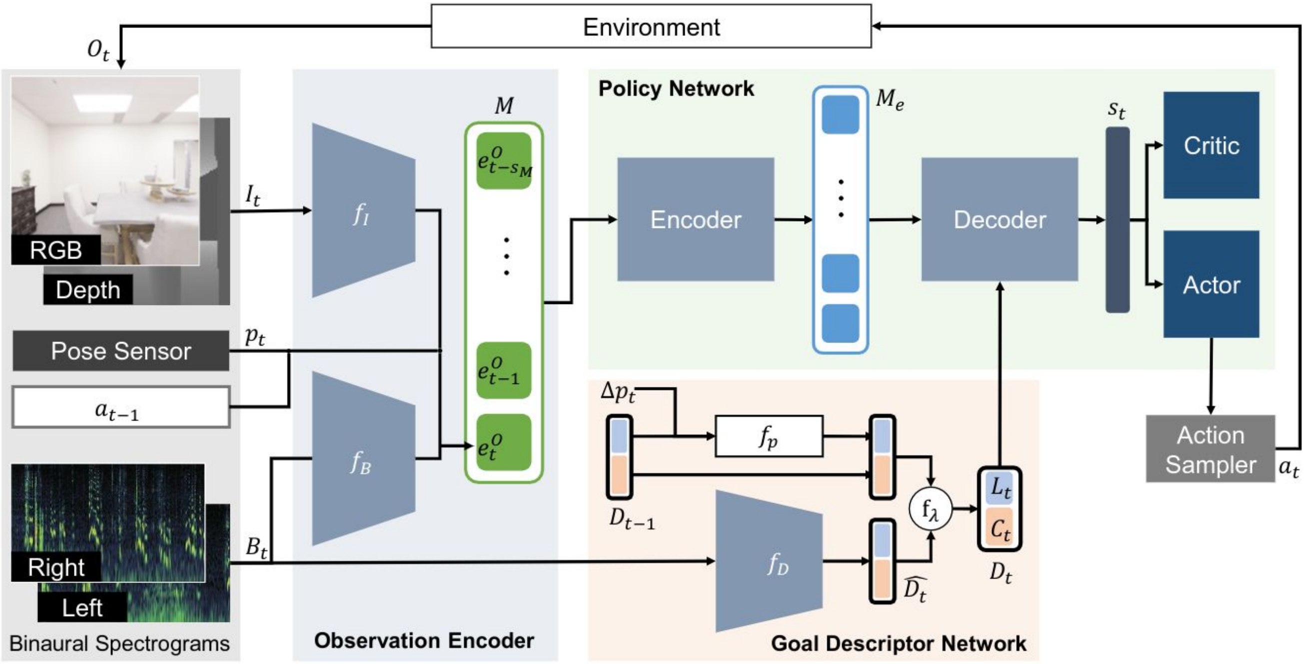
https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.pdf

# Semantic Audio-Visual Navigation




https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.pdf

# Pipeline




https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.pdf

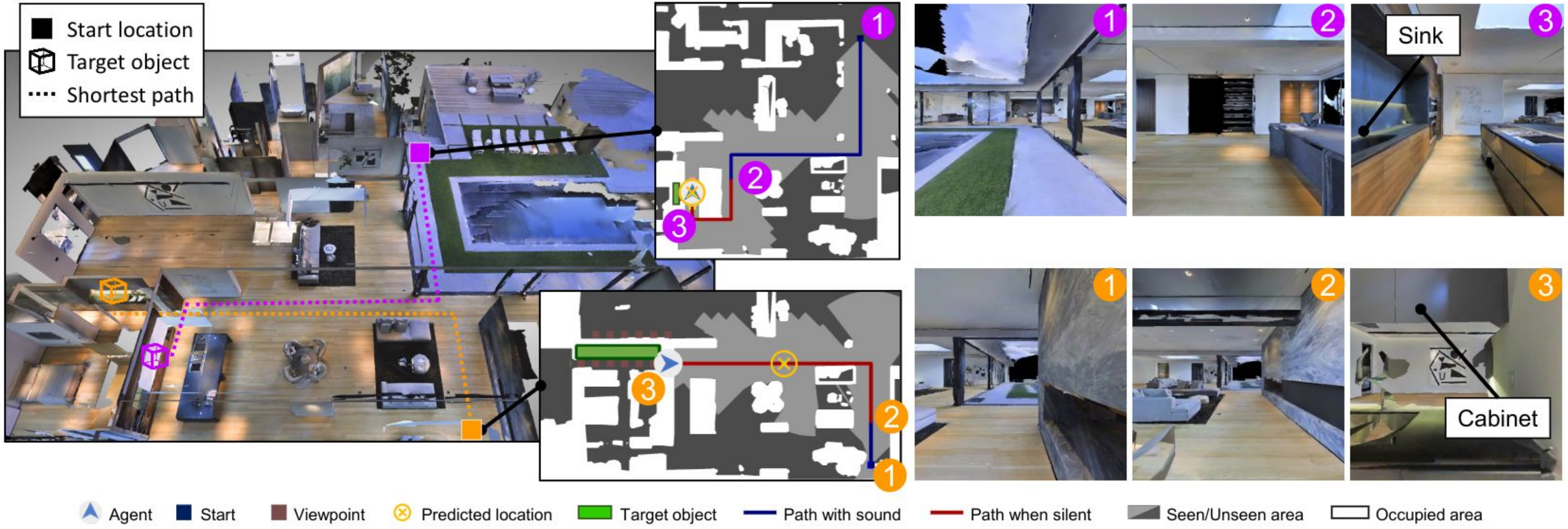
# SAVi outperforms all other models

- SNA: success weighted by inverse number of actions.
- DTG: average distance to goal when episodes are finished.
- SWS: fraction of successful episodes when the agent reaches the goal after sound stops.

| | *Heard Sounds* | | | | | *Unheard Sounds* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Success ↑ | SPL ↑ | SNA ↑ | DTG ↓ | SWS ↑ | Success ↑ | SPL ↑ | SNA ↑ | DTG ↓ | SWS ↑ |
| Random | 1.4 | 3.5 | 1.2 | 17.0 | 1.4 | 1.4 | 3.5 | 1.2 | 17.0 | 1.4 |
| ObjectGoal RL | 1.5 | 0.8 | 0.6 | 16.7 | 1.1 | 1.5 | 0.8 | 0.6 | 16.7 | 1.1 |
| Gan et al. [19] | 29.3 | 23.7 | **23.0** | 11.3 | 14.4 | 15.9 | 12.3 | 11.6 | 12.7 | 8.0 |
| Chen et al. [11] | 21.6 | 15.1 | 12.1 | 11.2 | 10.7 | 18.0 | 13.4 | 12.9 | 12.9 | 6.9 |
| AV-WaN [12] | 20.9 | 16.8 | 16.2 | 10.3 | 8.3 | 17.2 | 13.2 | 12.7 | 11.0 | 6.9 |
| SMT [15] + Audio | 22.0 | 16.8 | 16.0 | 12.4 | 8.7 | 16.7 | 11.9 | 10.0 | 12.1 | 8.5 |
| SAVi (**Ours**) | **33.9** | **24.0** | 18.3 | **8.8** | **21.5** | **24.8** | **17.2** | **13.2** | **9.9** | **14.7** |

Table 1: Navigation performance on the SoundSpaces Matterport3D dataset [11]. Our SAVi model has higher success rates and follows a shorter trajectory (SPL) to the goal compared to the state-of-the-art. Equipped with its explicit goal descriptor and having learned semantically grounded object sounds from training environments, our model is able to reach the goal more efficiently—even after it stops sounding—at a significantly higher rate than the closest competitor (see the SWS metric).

# Navigation trajectories




https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.pdf

# Other experiments

- Location predictor has a comparatively larger impact on the model's performance.
- Aggregation stabilizes the goal descriptor prediction.
- The proposed model is able to cope with long silence to reach goals.

| | Success ↑ | SPL ↑ | SNA ↑ | DTG ↓ | SWS ↑ |
|---|---|---|---|---|---|
| $C_t$-only | 20.5 | 13.5 | 11.6 | 9.8 | 11.0 |
| $L_t$-only | 23.9 | 16.2 | **13.5** | **9.3** | 13.8 |
| w/o aggregation | 21.9 | 14.3 | 11.1 | 9.7 | 13.4 |
| Full model | **24.8** | **17.2** | 13.2 | 9.9 | **14.7** |

Table 3: Ablation experiment results.

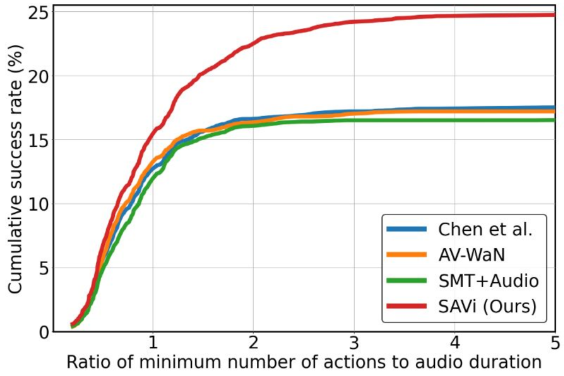



Figure 4: Cumulative success rate vs. silence percentage.

https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.pdf

Thanks!