# Semantic Audio-Visual Navigation Supplementary Material

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

In this supplementary material, we provide additional details about:

1. Video (with audio) for qualitative assessment of our agent's performance. Please listen with headphones to hear the spatial sound properly.
2. Implementation details and analysis of the baselines (Sec. 5)
3. Ablation of the policy network
4. Distribution of prediction accuracy over distance to goal (Sec. 5)
5. Analysis of semantic audio-visual navigation with distractors
6. On-policy location predictor training (Sec. 4.4)
7. Ablation with true goal category and location

## 1. Qualitative Video

The supplementary video demonstrates the audio simulation platform that we use and shows the comparison between our proposed model and the baselines as well as qualitative analysis for failure cases. Please listen with headphones to hear the binaural audio correctly.

## 2. Implementation Details and Analysis of the Baselines

**ObjectGoal RL.** We implement this baseline by first feeding the RGB-D observations into a CNN (similar CNN to $f_I(\cdot)$ in our model) and concatenating the visual features with a one-hot encoding of the target label. A one-layer GRU memory takes the concatenated feature as input and outputs a state vector of size 512. Similar to our work, this state representation is used by an actor-critic network to predict the action distribution and value of the current state. Furthermore, we use perfect stopping for this baselines since the model performs poorly with a learned stop action.

Although this baseline has the goal's ground truth label and perfect stopping, it fails quite often in reaching the goal. This shows knowing the category alone is insufficient to locate the particular object instance and to succeed in this task. The model needs to leverage both visual and acoustic cues to find the goal. This experiment also draws attention to the difference between the proposed semantic AudioGoal and the existing task of ObjectGoal.

**Gan et al. [4]** We compare to the model from Gan et al. [4] , which trains a goal location predictor in an offline fashion and uses a geometric planner for planning a path to the predicted goal location. We use the same amount of training data for our category predictor to train the goal predictor from [4]. The original model from [4] assumes a continuous periodic acoustic event and it cannot handle sporadic or short acoustic events like those considered in this work. To improve the existing model to perform in this task, we augment its goal location predictor with our update operation $f_\lambda$ (Sec. 4.2) for transforming the predicted location when the audio goal becomes silent.

In evaluation, our observations confirm those reported in [3]. Since the model does not leverage visual cues for reasoning about the goal location, it does not learn to associate visual and acoustic cues with scene observations and goal properties. Therefore, it is more prone to errors and the agent suffers from backtracking its steps quite often when the goal location prediction is inaccurate. The model achieves 15.9% success rate and 12.3% SPL on the *unheard sound* test split, compared to our SAVi model 24.8% success rate and 17.2% SPL. While [4] leverages external supervision for training the location predictor, this is not enough to solve this task efficiently because the agent needs to fully leverage the semantic and spatial cues from audio along with its visual perception to locate the sounding objects.

**AV-WaN [3].** While AV-WaN [3] reports large performance improvements over Chen et al. [2] on the standard AudioGoal task (see [3] for details), we do not observe similar margins between the two models here. Both models, AV-WaN [3] and Chen et al. [2], use RNNs to encode the

| | Success ↑ | SPL ↑ | SNA ↑ | DTG ↓ | SWS ↑ |
|---|---|---|---|---|---|
| **RNN Policy Network** | | | | | |
| Chen et al. [11] | 18.0 | 13.4 | **12.9** | 12.9 | 6.9 |
| SAVi w/ RNN+MLP (Ours) | **21.4** | **15.4** | 12.6 | **9.8** | **11.2** |
| **Transformer Policy Network** | | | | | |
| SMT [15] + Audio | 16.7 | 11.9 | 10.0 | 12.1 | 8.5 |
| SAVi w/ Transformer (Ours) | **24.8** | **17.2** | **13.2** | **9.9** | **14.7** |

Table 1: Ablation of our policy network with a typical RNN+MLP.

state representation; however, AV-WaN accumulates the observations at the waypoint prediction level while Chen et al. does so at each step. We speculate that this behavior creates large temporal gaps in the memory for AV-WaN, which makes it harder for the model to adapt to the more challenging task of semantic AudioGoal; the sound may stop at any moment, and the AV-WaN model may not be able to capture the last important acoustic cues in between waypoints. Our model outperforms both since it can keep track of a large set of observations and leverage this information at each step while navigating.

## 3. Ablation of the Policy Network

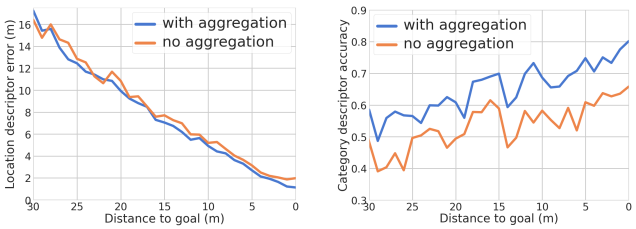
To analyze the impact of the transformer architecture of the policy network on our model performance, we include an additional ablation by replacing the transformer in our SAVi model with a typical RNN+MLP for action and value prediction (similar to [2]). Table 1 shows the results in the unheard sounds setting. We see that a significant part of the performance improvement comes from our goal descriptor network (GDN) contribution. Both models (SAVi w/ Transformer and SAVi w/ RNN+MLP) benefit significantly from having our GDN, with the transformer model leading to best performance since it allows the GDN to attend to longer observation sequences compared to the RNN.

## 4. Distribution of Goal Descriptor Accuracy

Figure 1 shows how the location descriptor error and the category descriptor accuracy change as the agent gets closer to the goal with and without temporal aggregation. The location error is measured as the Euclidean distance between the predicted and the ground truth goal location. The category accuracy is measured by whether the correct goal is predicted or not. We can see that the error of both predictions get lower as the agent gets closer to the goal location and the temporal aggregation leads to higher performance.

## 5. Analysis of Semantic Audio-Visual Navigation with Distractors

We have evaluated in the main paper the navigation performance of our model in the presence of acoustic distractors. The target and distractor sounds are disjoint in this setting and both are unheard at test time, which poses a great challenge for the agent to clearly separate the mixed audio signal. We believe this is a main factor in the performance drop seen by all models, though ours remains best (Table 2 in the main paper).




(a) (b)

Figure 1: Error analysis of the location predictions and the category predictions in the goal descriptor as a function of the agent's geodesic distance to goal.

| | Beeps | Music | Creak | Horn | Telephone |
|---|---|---|---|---|---|
| Chair | 0.26 | 0.28 | 0.20 | 0.20 | 0.24 |
| Cabinet | 0.25 | 0.25 | 0.14 | 0.12 | 0.23 |
| Counter | 0.28 | 0.47 | 0.34 | 0.25 | 0.41 |
| Sink | 0.03 | 0.07 | 0.03 | 0 | 0.07 |
| TV | 0.14 | 0.19 | 0.19 | 0.14 | 0.19 |

Table 2: Success rate of goals (rows) in the presence of various distractors (columns). We test our model with a single distractor type in each test run, and normalize the SR by the number of episodes for each goal type.

To further analyze the impact of acoustic dsitractors, we conduct an ablation of our model by changing the type of distractors at each test run. Table 2 shows a subset of the (goal, distractor) combinations. Indeed, when the distractor sound is sufficiently different from the goal (e.g., Music and Telephone), the model performs well, but when it is similar (e.g., Cabinet and Creak) or much louder (e.g., Horn) then it is harder for the model to extract a clear signal for the goal.

## 6. On-policy Location Predictor Training

As noted in the main paper, we find training the location predictor on-policy and online leads to higher accuracy compared to using a pretrained model. If we use an off-policy model in our approach (i.e., similar to the location predictor trained for Gan et al.), this version underperforms our model by 4.8% success rate and 4.7% SPL on the *unheard sound* test split.

## 7. Ablation with True Goal Category and Location

Our SAVi model learns to predict the goal descriptor (i.e., location and category) based on the heard acoustic cues while navigating. To show an upper bound performance for our goal descriptor network, we supply the model with the true goal category and location instead of the predictions. Our model achieves 65% SPL compared to the 24% SPL under the same setting but with predicted descriptors. Note that when the ground truth location of the goal is available at each step, the task boils down to the common PointGoal navigation [5, 1].

## References


[1] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020. 3

[2] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *ECCV*, 2020. 1, 2

[3] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 1

[4] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 1

[5] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 3