MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

RILA: Reflective and Imaginative Language Agent for Zero-Shot Semantic Audio-Visual Navigation

Yang, Zeyuan; Liu, Jiageng; Chen, Peihao; Cherian, Anoop; Marks, Tim K.; Le Roux, Jonathan; Gan, Chuang

TR2024-043 April 27, 2024

Abstract

We leverage Large Language Models (LLM) for zero- shot Semantic Audio Visual Navigation (SAVN). Existing methods utilize extensive training demonstrations for reinforcement learning, yet achieve relatively low success rates and lack generalizability. The intermittent nature of auditory signals further poses additional obstacles to infer- ring the goal information. To address this challenge, we present the Reflective and Imaginative Language Agent (RILA). By employing multi-modal models to process sensory data, we instruct an LLM-based planner to actively explore the environment. During the exploration, our agent adaptively evaluates and dismisses inaccurate perceptual descriptions. Additionally, we introduce an auxiliary LLM-based assistant to enhance global environmental comprehension by mapping room layouts and providing strategic insights. Through comprehensive experiments and analysis, we show that our method outperforms relevant base- lines without training demonstrations from the environment and complementary semantic information.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

RILA: Reflective and Imaginative Language Agent for Zero-Shot Semantic Audio-Visual Navigation

Zeyuan Yang[1]*, **Jiageng Liu**[2]*, **Peihao Chen**[3], **Anoop Cherian**[4],
Tim K. Marks[4], **Jonathan Le Roux**[4], **Chuang Gan**[5,6]
[1]Tsinghua University, [2]Zhejiang University, [3]South China University of Technology
[4]Mitsubishi Electric Research Laboratories (MERL), [5]UMass Amherst, [6]MIT-IBM AI Lab
yangzeyu21@mails.tsinghua.edu.cn

Abstract

We leverage Large Language Models (LLM) for zero-shot Semantic Audio Visual Navigation (SAVN). Existing methods utilize extensive training demonstrations for reinforcement learning, yet achieve relatively low success rates and lack generalizability. The intermittent nature of auditory signals further poses additional obstacles to inferring the goal information. To address this challenge, we present the **R**eflective and **I**maginative **L**anguage **A**gent *(RILA). By employing multi-modal models to process sensory data, we instruct an LLM-based planner to actively explore the environment. During the exploration, our agent adaptively evaluates and dismisses inaccurate perceptual descriptions. Additionally, we introduce an auxiliary LLM-based assistant to enhance global environmental comprehension by mapping room layouts and providing strategic insights. Through comprehensive experiments and analysis, we show that our method outperforms relevant baselines without training demonstrations from the environment and complementary semantic information.*

1. Introduction

Intelligent agents are anticipated to navigate intricate environments, leveraging both auditory and visual stimuli [31, 38]. Considering a scenario that a vase falls and breaks, a robot must swiftly pinpoint a target within a room, relying primarily on transient auditory cues. This need underpins our focus on the Semantic Audio-Visual Navigation (SAVN) task [11]. In SAVN, the target object within the scene emits intermittent sounds, which the agent must use, in conjunction with visual information, to find the object. In addition to the ambiguous goal information conveyed through sporadic sounds, intricate room layouts and complex navigation trajectories also present significant challenges [44], rendering the SAVN task notably difficult. Previous research [11] concentrated on the end-to-end training of reinforcement learning models, yielding inadequate performance despite the use of extensive training trajectories. Recent approaches enhance performance by integrating auxiliary modules [44] or employing oracle instructions [31, 38], which may not be feasible in real-world applications.

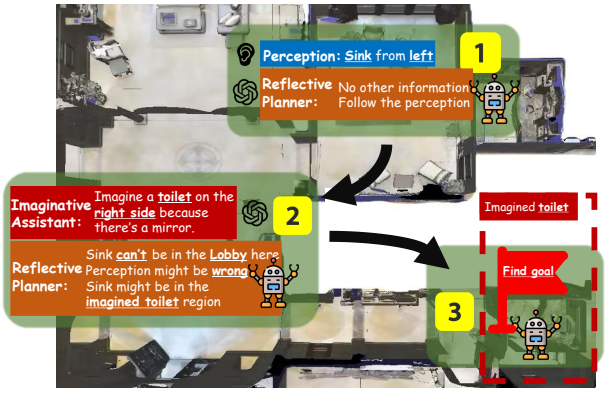

Figure 1. An illustration of our agent's strategy for semantic audio-visual navigation. The **Reflective Planner** initiates navigation by relying on perceptual information for exploration. When exploration leads to an incorrect region, it subsequently discounts the perceptual descriptions, redirecting its focus. Throughout this process, the **Imaginative Assistant** persistently contributes spatial insights and suggestions, thereby assisting in reasoning.

Large language models (LLMs) [35, 36] have shown remarkable progress [30, 47]. Beyond the promising performance on natural language tasks [37, 40], the integration of LLMs into embodied robotics applications has also resulted in substantial improvements [2, 15, 16, 48, 50]. Recent methods [55, 56] equip LLMs with multi-modal models [28, 29] that provide perception and feedback from the environment, either explicitly [49, 51] or implicitly [20, 24],

* Equal Contribution

in vision-and-language navigation tasks [4]. However, these applications also fail on SAVN due to their reliance on precise perception information and explicit goal descriptions. Consequently, realizing zero-shot SAVN, as anticipated for intelligent agents, remains a formidable challenge.

Therefore, we propose our ***R**eflective and **I**maginative **L**anguage **A**gent* (RILA), leveraging the inherent commonsense reasoning capabilities of LLMs to perform zero-shot SAVN. Practically, we design distinct perception models that process audio and visual signals, which further guide a frozen LLM in strategic planning. Through active exploration of the environment, our agent adaptively identifies and deprioritizes misleading goal descriptions. Furthermore, we introduce an LLM-based imaginative assistant, which extracts room layouts and provides high-level guidance. Incorporating this assistant enables our agent to achieve comprehensive environmental understanding and navigate toward the target object in a zero-shot manner. Fig. 1 provides an illustration of our agent's navigation.

To validate our approach, we conduct experiments within the SoundSpaces framework. Experimental results show that our method surpasses relevant baselines without reliance on training demonstrations or complementary modules. Notably, our agent exhibits a success rate exceeding 60% when paired with oracle perceptions, highlighting the strong planning capability of LLMs. Additionally, we conduct a thorough analysis of the bottleneck of the current task configuration. We summarize our contributions as follows:

- We propose RILA for zero-shot SAVN, exploiting the commonsense reasoning capabilities of LLMs to navigate effectively without precise goal descriptions.
- We introduce an imaginative assistant, designed to deduce the environment's room layout and provide comprehensive suggestions, thereby enhancing the navigation.
- Experiments substantiate that RILA surpasses previous baselines, which require training, in a zero-shot manner. We also conduct a thorough analysis of the SAVN task.

2. Related Work

2.1. Semantic Audio Visual Navigation

Semantic audio-visual navigation is defined in Habitat [33, 41] with the SoundSpaces dataset [10, 13]. Previous research [9, 52] extract features from RGB-D images and two-channel spectrograms using pre-trained encoders separately [3, 10], and then train an end-to-end policy network by reinforcement learning to predict the next action. However, these methods lack generalizability, failing in unsupervised scenes [44] despite necessitating extensive training demonstrations. Recent methods [31, 38] query for human instructions during the navigation. K-SAVEN [44] further constructs a knowledge graph to provide spatial comprehension. Instead of training on massive demonstrations, our method exploits the commonsense reasoning capabilities of LLMs to perform solve the task in a zero-shot manner.

2.2. Navigation with Large Language Models

LLMs have recently demonstrated impressive reasoning abilities across a range of tasks [21, 39], including embodied tasks [17]. Recent studies [43, 54] investigate visual-language navigation with LLMs. For instance, ESC [56] employs LLMs to deduce relationships between objects, thereby aiding navigation. Chen et al. [14] and Szot et al. [42], on the other hand, utilize visual foundation models to convert perceptions into natural language instructions. However, the application of LLMs in SAVN remains underexplored, especially since prior methods often rely on ground-truth goal descriptions. In contrast, RILA reflectively navigates toward the target, handling potentially misleading goal descriptions.

2.3. Layout Complementary

Spatial understanding, particularly regarding room layout, is crucial for comprehending complex environments. LGD[27] employs a room-type codebook to conceptualize room layouts from image clips. Text2Room [23], conversely, creates entire rooms guided by textual instructions. Recent LayoutGPT [18] taps into the visual planning capabilities of LLMs to produce plausible layouts for visual generation. In our work, RILA utilizes LLMs to progressively deduce the room layout and type, thereby achieving a global understanding of the environment.

3. Method

In this work, we consider solving the Semantic Audio Visual Navigation (SAVN) task [11] in a zero-shot manner, challenging agents to locate the sounding object within an intricate and unseen environment. Notably, the audio signals here are sporadic and often absent, posing a significant challenge to the agent's decision-making process. Instead of training on trajectories from the simulated environment or incorporating additional semantic information, we leverage the intrinsic commonsense reasoning capabilities of LLMs for navigation planning.

3.1. Overview

In this section, we provide an overview of our RILA framework, illustrated in Fig. 2. RILA consists of three parts: the perception module, the Imaginative Assistant, and the Reflective Planner, which we will introduce separately.

The perception module transforms the sensory data into natural language descriptions. Visual perceptions o_t^v are directly processed via a pre-trained visual-language model, which discerns and catalogs the observed objects, thereby facilitating the construction of a semantic top-down map.

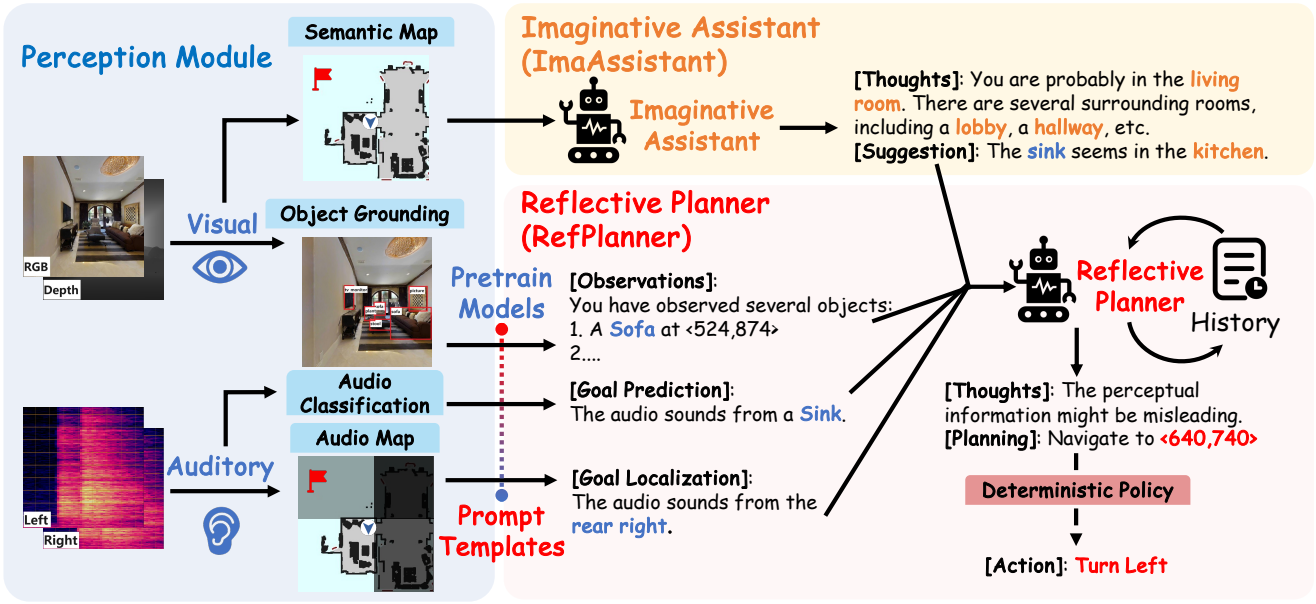

Figure 2. The architecture of our agent comprises three primary components. Firstly, the perception module transforms sensory inputs into text-based descriptions. Secondly, the Imaginative Assistant analyzes regional information and offers strategic guidance from a global perspective. Lastly, by integrating the two components, our Reflective Planner assesses perceptual data and navigates toward the target.

We develop distinct modules for auditory perceptions o_t^a to pinpoint the goal location and identify pertinent semantic cues, given the intermittent nature. Both perceptions are then synthesized into a text-based format for planning. A detailed description is illustrated in Section 3.2.

Extending beyond individual objects, we integrate an LLM-based Imaginative Assistant (ImaAsssistant) to deduce room layouts, thereby enriching the spatial comprehension of intricate environments. ImaAsssistant then utilizes the layout of both explored and partially observed areas to provide strategic planning guidance, aiding in navigation. A thorough explanation is provided in Section 3.3. By amalgamating insights from the perception module and ImaAsssistant, our Reflective Planner (RefPlanner) leverages inherent commonsense reasoning abilities to explore the environment and identify misleading auditory descriptions, circumventing the need for exact sound localization. Detailed explanations are shown in Section 3.4.

3.2. Audio Visual Perception

Following [44], we use the same pre-trained audio classification model M^a_{obj} to infer the target object. Considering the transient nature of audio signals, which presents a considerable obstacle in precise identification, we employ a progressive strategy. Upon an audio signal o_t^a at time step t, we make a prediction $M^a_{\text{obj}}(\{o^a_{1,\dots,t}\})$ by amalgamating the current audio with the accumulative history, facilitating a refined accuracy. The object $\hat{g}_t$ with the highest cumulative prediction score at time t is thus designated as the current goal object:

$$\hat{g}_t = \underset{g}{\text{argmax}} \left(\sum_{i=1}^{t} \mathbb{1}_{M^a_{\text{obj}}(\{o^a_{1,\dots,i}\})=g} \right), \tag{1}$$

where $\mathbb{1}$ denotes the indicator function. Guided by the prediction $\hat{g}_t$, we aim to further localize it, thereby improving the distinction of the target from analogous entities in the environment. Nonetheless, the complex reverberation of the simulation poses a significant challenge for localization, as evidenced by an error margin of about 8 meters [11].

Therefore, we partition the localization into independent estimations of distance and direction. To quantify sound distance, we collected 10,000 unheard auditory samples from the training environment to delineate the simulation's dimensional attributes. A pre-trained ResNet-18 model fine-tuned on this dataset demonstrates commendable accuracy in estimating distances. Predicting direction, however, is substantially more arduous.

Instead of ascertaining the precise angle, we shift to identify the binary directionality, greatly simplified by the dual-sensor configuration. Nonetheless, techniques such as Interaural Time Difference (ITD) [6, 22] and fine-tuned models fall short of the task, which is further discussed in Section 5. Consequently, we employ weighted predictions based on the Root Mean Square (RMS) intensity of auditory signals from the dual channels, denoted by R_l^t and R_r^t. Practically, we consider the audio source to be from the side with the larger RMS intensity. For each point p and time t,

the confidence C_p^t is calculated as:

$$C_p^t = \sum_{i=1}^{t} w_i^a \cdot \mathbb{1}_{RMS}(p, o_i^a), \tag{2}$$

where $\mathbb{1}_{RMS}(p, o)$ is an indicator function which is equal to 1 if p is located, with respect to the agent, in the side corresponding to the larger RMS intensity given observation o, and the weight w_t^a is calculated as $w_t^a = \frac{|R_l^t - R_r^t|}{\max(R_l^t, R_r^t)}$. Through iteratively accumulating the weighted predictions, we construct an audio map that facilitates an approximate localization of the goal.

To transform visual signals into linguistic representations, we employ the pre-trained GroundingDINO for both delineating bounding boxes and identifying the objects within the RGB observation, thereby furnishing a rudimentary environmental understanding. Besides, we separately prompt to detect the predicted goal object in case the target is missed. Simultaneously, a semantic top-down map is constructed from the Depth observations, with the map segmented into distinct regions demarcated by detected walls, enabling the assistant to provide a region-level comprehension. A more detailed illustration of our perception modules is further provided in Appendix.

3.3. Imaginative Assistant

Given the restricted information from the perception module, the planning relies mainly on discrete objects. However, a global environmental understanding substantially benefits planning, especially for distant goals requiring multi-room navigation. To address this, we integrate an auxiliary LLM-based Imaginative Assistant (ImaAsssistant), offering strategic suggestions to bolster navigation.

In practice, ImaAsssistant infers room layouts. By partitioning the semantic map into regions using the detected walls, we instruct ImaAsssistant to determine closed room types from observed objects. Yet, as a comprehensive exploration of a room rarely occurs, partially observed rooms are more frequently encountered. Therefore, we utilize the spatial imagination capabilities of LLMs to conceptualize the layout of these rooms, subsequently directing it to deduce room types by interior objects and adjacent rooms. We present below simplified versions of the prompts.

/* Task Description */
Please infer the room type and precise layout of the provided interested region.
/* Room Layouts */
Observed Rooms: living room, etc.
Partially Observed Room: wall_1, wall_2, etc.
Internal Objects: chair_1, chair_2, table, etc.

Through iterative deduction of both observed and partially observed rooms, RILA attains a comprehensive understanding of the environment, which yields additional insights beyond the scope of individual objects. To augment the planning, ImaAsssistant is further instructed to provide strategic navigation advice. Rather than specific waypoints, ImaAsssistant reasons about the potential goal locations, considering spatial layouts and semantic attributes. These insights enable ImaAsssistant to make suggestions that assist in selecting waypoints more effectively. A simplified version of the prompt template is presented below.

/* Task Description */
Given the room layout, infer where the **Counter** is.
Give your advice about which room to explore.
/* Information */
Current room: living room
Surrounding rooms: kitchen, hallway, etc.

3.4. Reflective Planner

By incorporating layouts and suggestions from ImaAsssistant, our LLM-based Reflective Planner (RefPlanner) harnesses the inherent commonsense reasoning capabilities in planning based on perceptions. At each time step t, audio and visual perceptions are formatted as *Goal Description* and *Observation*, respectively. Additionally, a *Task Description* is articulated at the outset. A simplified template for the perception prompt is as follows:

/* Task Description */
You are performing a navigation task.
/* Goal Description */
Navigate to the object that sounds like a **Counter**.
/* Observations */
You have observed the following objects.

With a natural language synopsis of the environment and the designated navigational objective, we commission RefPlanner to strategize high-level planning. Rather than specifying actions outright, we implement a heuristic method, frontier-based exploration (FBE), which discerns the junctures between explored and uncharted territories as potential waypoints for environmental reconnaissance. Instead of determining specific action, RefPlanner is directed to reason and select an exploration frontier based on current perceptions in a zero-shot manner. The navigation history of perceptions and reasonings is also provided. Practically, we implement a deterministic policy for decomposing the waypoint into action sequences. Utilizing a connected graph derived from the semantic top-down map, we apply Dijkstra's algorithm to determine the shortest path to the waypoint.

Moreover, as outlined in Section 3.2, the perception de-

scriptions, particularly the goal location, are often ambiguous and may lead to misconceptions, while an intelligent agent is anticipated to actively interact with the environment to make judgments about uncertain perceptions. Therefore, along with the localization confidence of the frontier from the perception module, we hint to RefPlanner about the potential inaccuracy, which empowers it to explore the environment adaptively and reflect the reliability of perception, thus enhancing its proficiency in locating the target object. The layouts and suggestions from ImaAsssistant are included as well. We present below a simplified version of the template used for the navigation prompt.

/* Agent Position */
You are at $\langle x, y \rangle$
/* Hint */
The perceptual confidence is not always accurate.
/* Frontier Candidates */
Frontier 1: $\langle x, y \rangle$ in the living room
Perceptual confidence: c
Surrounding objects: chair_1, chair_2, table, etc.
/* Suggestions */
The goal object may be in the kitchen.

As shown in Fig. 1, RefPlanner adaptively selects appropriate waypoints from a global perspective. When RefPlanner fails to find the target after exploring an area based on perceptions, it identifies perceptual inaccuracies and navigates using object characteristics. The full prompt scheme and a detailed example of the navigation are provided in Appendix. In practice, we implement all LLMs using the March 2023 version of `gpt-3.5-turbo`, leveraging the OpenAI LLM API service[1] with a temperature of 0.0.

4. Experiments

4.1. Experimental Setup

4.1.1 Datasets

We use SoundSpaces [10, 13] from Habitat [33, 41] environment to simulate navigation in 3D environments. We adopt the Matterport3D (MP3D) dataset for its ground-truth region layout labels and object labels. In particular, we evaluate our RILA on 1,000 test episodes within 10 unseen scenes with unheard sounds from 21 goal objects.

4.1.2 Baselines

We compare our model with several baselines:

- **AudioGoal** [10] uses a GRU state encoder to acquire the following action with an end-to-end RL policy network.
- **AV-WAN** [12] designs a waypoint predictor and leverages a local path planner to navigate to the waypoint.
- **SAVi** [11] incorporates a goal descriptor network to predict both the classification and location of the sounding object.
- **AVLEN** [38] adopts a hierarchical RL policy with goal predictor and memory unit, and queries oracle instructions from humans if necessary.
- **K-SAVEN** [44] proposes an end-to-end policy network with a knowledge graph constructed on the training data, presenting the relationship between regions and objects.

In addition, we incorporate two zero-shot methods based on foundation models to facilitate a more comprehensive comparison. The ground truth goal object is provided here.

- **ImageBind-LLM** [20] is a novel multi-modality model that aggregates ImageBind [19] and LLaMA-Adapter [53] and we use the perfect stop strategy.
- **ESC** [56] leverages LLMs and Probabilistic Soft Logic (PSL) [5] to choose a frontier for a visual-language navigation task. We provide our audio goal description.

4.1.3 Metrics

Following previous work [9, 38, 44], we report agent performance with the following metrics: Success Rate (SR), Success Rate weighted by Path Length (SPL), Success Rate weighted by Number of Actions (SNA), and Success When Silent (SWS), all in percentage (%). We also report the average Distance To Goal (DTG) in meters at episode end.

4.1.4 Implementation Details

Consistent with previous studies, the agent is provided with RGB and depth images at a resolution of 256 $\times$ 256. It also receives two-channel audio clips in the form of 65 $\times$ 26 spectrograms. The action space includes *MoveForward*, *TurnRight*, *TurnLeft*, and *Stop*, with a movement step set at 1 meter. Additionally, the agent obtains its GPS location at each time step. Detailed implementation details are provided in Appendix.

4.2. Experimental Results

The comparative results are presented in Table 1. We derive the results of major baselines from their respective papers. For ESC and ImageBind-LLM, we incorporate ground-truth audio descriptions for the SAVN task. Implementation details are provided in the Supplementary Material. According to Table 1, our agent surpasses baselines that utilize end-to-end reinforcement learning training, such as SAVi, in a zero-shot manner. Even when juxtaposed with baselines that utilize additional information, RILA achieves a higher success rate. Besides, we notice that Imagebind-LLM fails on the SAVN task, despite incorporating ground-truth audio descriptions, reflecting the limited performance

[1] https://platform.openai.com/docs/models

	Method	SR (%) ↑	SPL (%) ↑	SNA (%) ↑	DTG (m) ↓	SWS (%) ↑
Supervised	AudioGoal [10]	16.5	15.5	10.4	12.8	5.6
	AV-WAN [12]	17.2	13.2	12.7	11.0	6.9
	SAVi [11]	24.8	17.2	13.2	9.9	14.7
	AVLEN [38]	26.2	17.6	14.2	9.2	15.8
	K-SAVEN [44]	34.4	**23.4**	**21.7**	**6.6**	14.3
Zero-Shot	Imagebind-LLM† [20] + Audio*	2.4	1.5	1.1	22.6	1.4
	ESC [56] + Audio*	23.6	8.0	4.8	17.7	14.2
	Ours w/o Assistant	31.4	9.6	6.8	12.2	15.3
	Ours	**35.4**	11.8	8.7	11.4	**20.4**

Table 1. Comparison with relevant baselines on SoundSpaces **Matterport3D** test dataset. AVLEN incorporates extra oracle instructions. † denotes the perfect stop strategy and *Audio** indicates that the ground-truth audio description is provided. In contrast, our method requires no training trajectories or additional semantic information.

Method	SR (%) ↑	SPL (%) ↑	SWS (%) ↑
Random†	19.8	11.8	16.2
Nearest†	9.8	22.6	6.4
Llama-2 7B	39.4	22.2	35.4
Ours	60.8	39.6	56.6

Table 2. Ablation study on RefPlanner by replacing it with heuristic frontier selection methods and replacing the ChatGPT with Llama-2. † indicates using oracle stop.

of open-source multi-modality foundation models on complex embodied tasks. Notably, our approach significantly outperforms previous works in terms of SWS, with over 40% improvement over K-SAVEN. This underscores the exceptional efficacy of our method in scenarios involving long distances and intermittent sounds, thereby highlighting the potential of harnessing the commonsense reasoning abilities of LLMs for navigation in physical environments.

We observe a relatively lower SPL of our method, attributed to the fact that RILA requires holistic exploration of the environment to ascertain the target object due to the absence of end-to-end training. Additionally, given the vague nature of the goal descriptions, RILA adopts a more cautious strategy for navigation, often traversing longer distances before reaching the objective. For better illustration, we provide two cases of snapshots of the navigation process using RILA in Fig. 3. As demonstrated in the left case study, RILA initially explores the living room, guided by erroneous perceptual cues. Upon realizing the absence of the goal object, RILA shifts its navigation toward the bathroom, utilizing object characteristics to locate the toilet. This process highlights RILA's ability to effectively reflect on potentially misleading goal descriptions, a factor that inevitably results in a lower SPL. We posit that enhancing audio localization, perhaps through the well-established Neural Radiance Fields (NeRF) [34], could further improve the SPL. Moreover, as depicted in the right case of Fig. 3, when the RefPlanner encounters unexplored areas, the ImaAssistant supplies conjectural room layouts. The spatial insight directs the RefPlanner to explore the kitchen instead of the dining room in search of the sink, underscoring the ImaAssistant's utility. Overall, RILA demonstrates the capacity to adaptively navigate complex environments.

4.3. Ablation Study

Ablation on ImaAssistant. As shown in Table 1, the integration of ImaAssistant markedly improves performance, underscoring the impact of strategic guidance. We also observed considerable advancements in SWS, demonstrating the crucial role of comprehensive layout understanding for long-distance navigation in intricate settings.

Ablation on RefPlanner. We replace our frontier selection RefPlanner with two heuristic frontier-based exploration methods, namely Random which selects a frontier randomly, and Nearest which selects the nearest frontier. We also compare the ability of GPT-3.5 and Llama-2 for frontier selection by replacing GPT-3.5 in RefPlanner with Llama-2. To eliminate the effect from perception, we use ground-truth perceptions (*i.e.*, acoustic object, audio map) in these experiments. In the two heuristic approaches, we automatically execute the *Stop* action when the distance to the goal is less than 1m. As illustrated in Table 2, despite access to ground-truth perceptions, these heuristic methods exhibit poor performance. Notably, Llama-2 7B [45] also struggles to locate the goal object, indicating the lack of spatial reasoning ability of Llama-2 for navigation tasks.

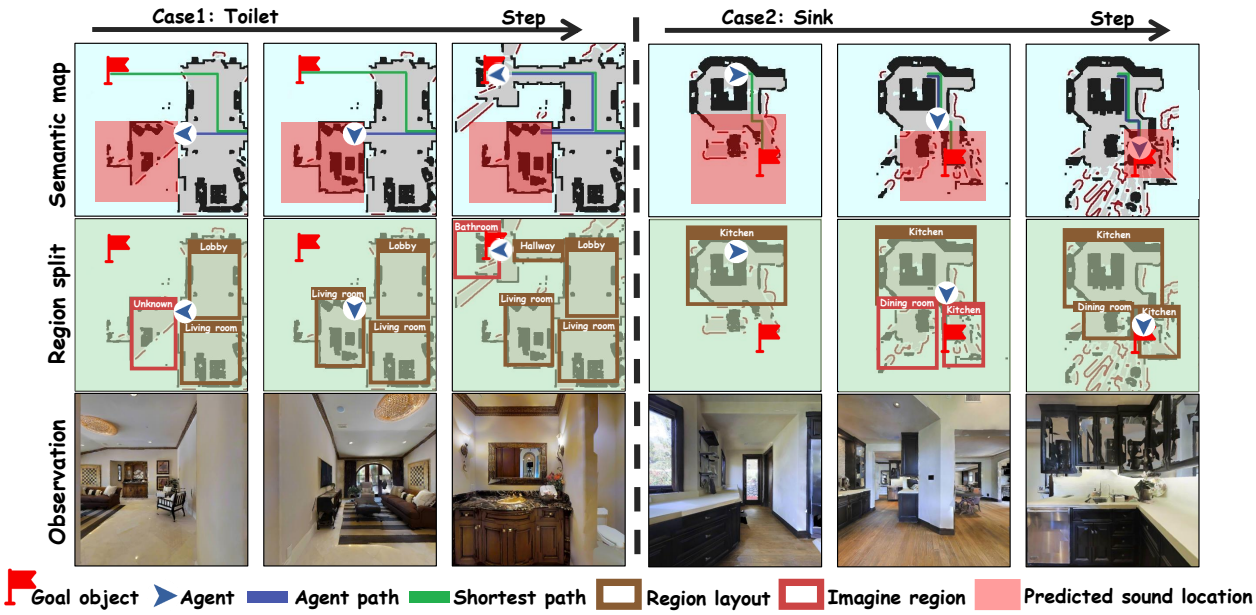

Figure 3. Visualization of two navigation trajectories, including region layouts and egocentric observations. The left case demonstrates how **RefPlanner** reflects on a misleading perception, whereas the right case illustrates that **ImaAssistant** makes imagination and suggestions, guiding **RefPlanner** in waypoint selection based on semantic relevance.

Perception	Accuracy (↑)
Object Recognition	83.9%
Audio Classification	93.0%
Audio Distance	83.8%
Audio Direction	73.7%

Table 3. Accuracy results of different perception modules. Object recognition accuracy represents the probability that the detected item is correctly classified. Audio distance prediction is deemed accurate within a 4-meter error range.

Ablation on Perception Module. Furthermore, we conducted a comprehensive evaluation of the perception modules across 500 episodes from 10 scenes. Results are shown in Table 3. GroundingDINO achieves an 85.0% recall rate on object recognition, indicating only a 15.0% error rate in goal object identification. For all recognized objects, the accuracy also reaches a notable 83.9%. Similarly, the audio classifier distinguishes among 21 classes with an accuracy rate of up to 93.0%. By progressively refining the prediction, RILA made correct predictions in almost all episodes. The accuracy of audio distance prediction is also commendable, reaching 83.8% within a 4-meter margin of error, and has an average distance error of 2.8 meters. Conversely, the accuracy of binary judgments on audio direction is limited to 73.7%, indicating a significant likelihood of error accumulation over steps. To investigate whether the direction judgment is impacted by complex reverberations in intricate environments, we further separately evaluate episodes based on whether the goal distance is less or more than 15 meters. Notably, accuracy reached 85.6% for shorter distances, in stark contrast to only 59.5% for longer distances. These findings underscore the difficulty of making binary direction determinations in SAVN, particularly over extended distances.

In conclusion, each component of RILA demonstrates competitive performance, with the exception of direction classification, which tends to be less reliable. To delve deeper into the capabilities of RILA, we present a comprehensive analysis in Section 5.

5. Analysis and Discussion

In this section, we focus on the following research questions: (i) Are LLMs adequate for completing complex navigation tasks? (ii) Does the sensory data provided by the SoundSpaces simulation offer clarity and sufficiency for effective navigation? (iii) Are there any inappropriate scenario settings within the current task configuration?

LLMs excel in intricate language-based navigation with inherent commonsense reasoning capabilities. By integrating ground-truth perceptual information, we investigate the navigational planning capabilities of LLMs. Rather than specifying precise goal locations, we provide only a rough

Method	SR ↑	SPL ↑	DTG ↓
Ours	30.2	9.0	11.8
+ GT Audio Semantic	30.2	11.2	11.6
+ GT Audio Localization	52.4	24.6	6.4
+ GT Visual Perception	62.0	39.2	4.8

Table 4. Comparison of incorporating different ground-truth perceptions on the validation dataset. Experiments in each row include the ground-truth information from all previous rows.

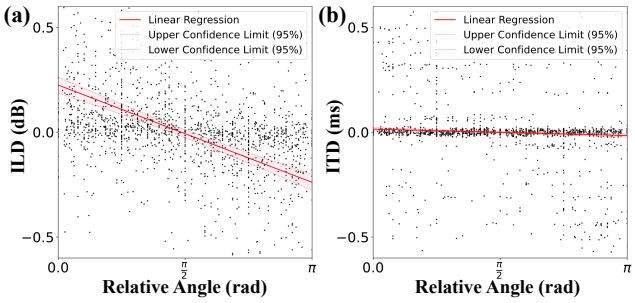

Figure 4. ILT and ITD of the sampled data points. We present the linear regression and the corresponding confidence intervals.

area. According to the results in Table 4, our agent achieves a success rate exceeding 60% with a DTG under 5 on the validation dataset. Failures typically arise from encountering similar objects in the target area or due to the inherent limitations of FBE in long-distance navigation. These findings further confirm the adequacy of LLMs' planning abilities for navigational tasks.

Besides, we observe that providing only ground-truth auditory data yields commendable performance. Conversely, the success rate markedly decreases in the absence of precise audio location information, consistent with the experimental results of the perception modules. Although RILA can effectively utilize potentially imprecise perceptual description, it remains vulnerable to misdirection caused by similar objects, thereby constraining the overall performance. These observations suggest that the current bottleneck in the SAVN task lies in sound source localization.

The auditory sensory data is inadequate for precise localization. To further investigate the audio localization, we sampled 4,000 dual-channel audio data points from the environment and computed two metrics: Interaural Level Difference (ILD) [46] and ITD. These metrics, crucial for sound source identification in dual-channel audio [1, 32], measure differences in sound intensity and arrival time, respectively. The results are depicted in Fig. 3, where the x-axis represents the sound source angle relative to the agent. Ideally, these metrics should display a pronounced negative correlation with the angle [25]. Our analysis reveals that while ILD demonstrates some negative correlation, serving as the basis for our direction classification, ITD does not effectively indicate the sound's relative direction. This underlines the constraints of the current audio input configuration [26], complicating precise localization based on auditory inputs. Detailed analysis is provided in Appendix.

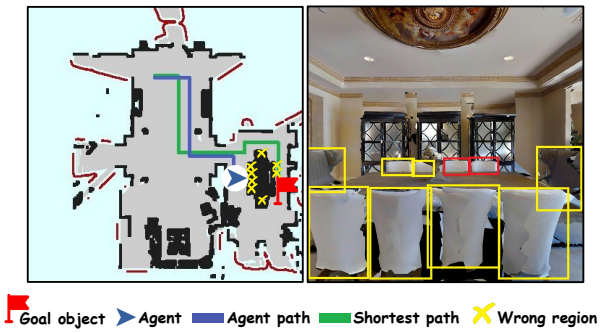

Figure 5. An example of an episode where the goal object is indistinguishable. In this case, the target is far from the agent and surrounded by similar, incorrect items.

Some cases could be further improved. Even in the absence of precise localization, semantic cues are expected to guide the agent to the target. However, our observations reveal situations where both audio localization is imprecise and semantic information fails to sufficiently differentiate between objects. For instance, as illustrated in Figure 5, the sounding object is distant from the agent, surrounded by numerous similar items, such as eight chairs in this case. In SAVN, where sounds are intermittent, the agent must semantically discern the correct stopping point. In this example, only two positions would lead to success. Lacking adequate reasoning cues, the agent resorts to random selection, leading to failure without exact goal location details. We postulate that these episodes could be improved by introducing distinct visual differences in target objects, such as overturning chairs, thus providing definitive cues for the agent to accurately identify the target.

6. Conclusion

In this work, we propose RILA, a reflective and imaginative agent for zero-shot semantic audio-visual navigation. By utilizing distinct models for sensory data processing, RILA guides an LLM-based reflective planner in active environmental exploration. Throughout this exploration process, RILA reflectively assesses and disregards erroneous sensory perceptions, especially the goal descriptions. Besides, we integrate an LLM-based auxiliary imaginative assistant, designed to generate room layouts and offer strategic guidance, thereby improving global understanding of the environment. Comprehensive experimental results demonstrate the efficacy of RILA.

RILA: Reflective and Imaginative Language Agent for Zero-Shot Semantic Audio-Visual Navigation

Supplementary Material

A. Implementation Details

In our experiments, we utilize the Matterport3D (MP3D) [8] environments within the SoundSpaces [10]. For the Imagebind-LLM [20] baseline, we involve directly providing the type of the goal object to construct the corresponding prompt. In contrast, for the ESC [56] baseline, we formulate the task instructions incorporating ground truth audio information. The performance results for other baselines are retrieved from their respective official papers. Unless otherwise specified, all experiments are conducted in a zero-shot manner on the test dataset.

B. Method

In this section, we provide detailed components of RILA.

B.1. Audio Perception

Audio Classification In this section, we provide the details of our audio classification model. We process original sounds from the Soundspace training set by segmenting them into one-second segments. These segments then undergo data augmentation through techniques such as time warping, time masking, and frequency masking. Additionally, each audio segment was enhanced using linear pitch modification and the Short Time Fourier Transform (STFT), collectively expanding our dataset to 30,000 samples. These enhanced segments were subsequently used for training a pre-trained Resnet18 model, obtained from torchvision.

Audio Localization Initially, we employ the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [7] method to directly ascertain the direction of audio sources. However, we encountered a limitation with GCC-PHAT, particularly in its performance on near-field models. This limitation manifests as an error margin of up to $\frac{\pi}{3}$ in our dual-channel audio setup, necessitating the adoption of specific strategies to determine the audio direction. Therefore, as discussed in Section 4, weighted predictions by RMS values are employed to ascertain the audio direction. It has been observed that occurrences of significant disparities in RMS values between audio channels are relatively infrequent. Consequently, the associated weights are often proportionately smaller. To more accurately represent the differences between dual-channel RMS values, we have adjusted the scaling of the weight by a factor of 0.4. This normalization enables us to derive more distinctive directional assessment weights, which are integral to the construction of the AudioMap.

In the process of predicting distances, we adopt a similar approach by randomly sampling 30,000 audio clips to train a Resnet18 model, which is aimed at capturing the scale of distances. Once a rough distance prediction is obtained, we apply a weighted approach to refine it, which involves expanding the predicted distance by a margin of 15%. Specifically, any distance falling below 85% or exceeding 115% of the predicted value is assigned a weight of 0. For distances that lie within this 15% boundary, we employ a linear decay weighting scheme, assigning the highest weight of 1 to the predicted distance itself. By effectively integrating predictions of both audio direction and distance, our perception modules accomplish a preliminary localization.

AudioMap Construction In our method, the AudioMap is constructed by integrating weighted predictions of both audio direction and distance. The audio direction predictions facilitate the partitioning of the map into distinct regions. Meanwhile, the distance predictions contribute to predicting regions with a circular, ring-shaped configuration. This integrative approach culminates in forming a confidence-based AudioMap, offering a comprehensive representation of audio spatial characteristics.

To enhance the interpretability of the AudioMap, we have visualized it as a grayscale image, which is dimensionally equivalent to the corresponding semantic map. In this visualization, each pixel's level of confidence is normalized to facilitate easier interpretation. The highest confidence regions are presented by white pixels, as shown in Figure 6.

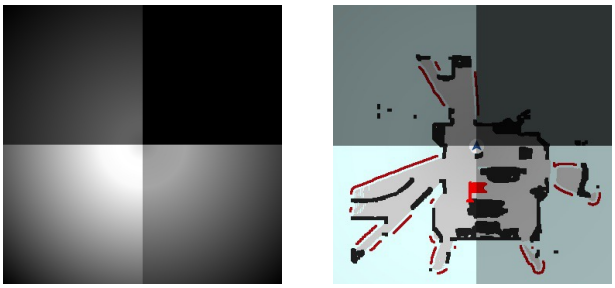

Figure 6. An example of our AudioMap. In the left figure, we display the direct AudioMap, where white pixels signify areas of high confidence. On the right, the figure showcases a composite image that merges the AudioMap with the semantic map, illustrating how they integrate and complement each other.

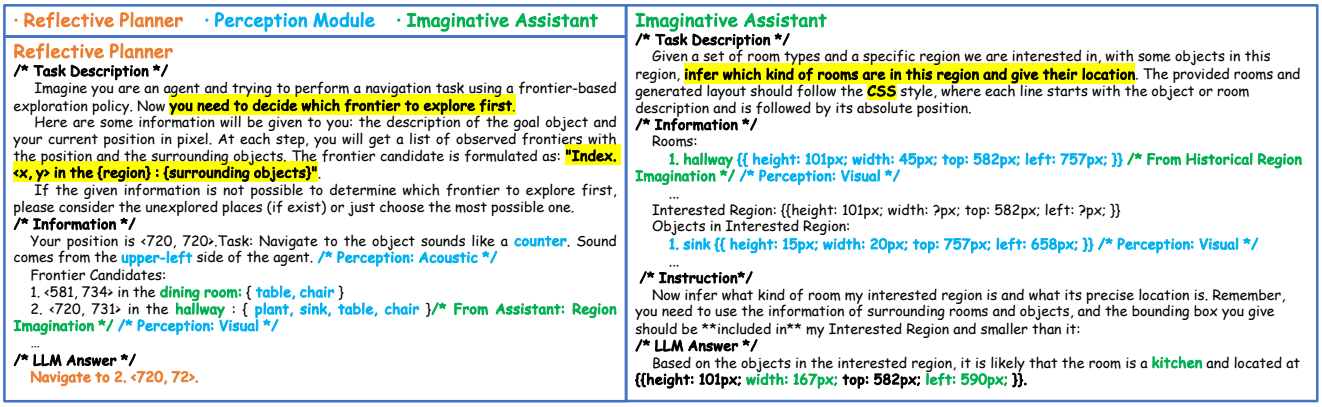

Figure 7. An example of the navigational prompts used by RILA. On the left side, we display the specific instructions provided to RefPlanner for selecting exploration frontiers. On the right side, the instructions given to ImaAssistant are shown, which guide it in inferring the environmental layout.

B.2. Visual Perception

GroundingDINO prompting We employ a two-stage strategy to differentiate between recognizing the goal and other objects, involving distinct recognition processes for general objects and goal prediction. General object recognition necessitates higher accuracy for imagining the region but has lower recall requirements. Hence, we formulate the prompt by presenting these objects, resulting in a certain level of missed recognition but with a lower error rate. On the other hand, goal prediction recognition demands a stronger emphasis on representation and higher recall. A prompt template is shown below. This two-stage strategy guarantees a significantly lower missed recognition rate.

> **/* Object Recognition */**
> There is a **Counter** (Goal Prediction) in:

Semantic Map Construction By utilizing the pixel data from the depth image and the camera's intrinsic parameters, we compute the spatial coordinates for each pixel. These coordinates are then amalgamated to create a point cloud, where each point is represented by three coordinates. In this structure, the z-coordinate, which denotes height, varies between 0 and 1. We apply a filtering process to this point cloud based on the height parameter. Points with a height exceeding 0.5 are identified as parts of obstacle regions, inferred from their horizontal coordinates. Conversely, points with a height less than or equal to 0.5 are classified as free regions. Areas falling outside the camera's field of view are designated as unknown regions. Subsequently, we project the map, initially scaled in meters, into a pixel-based image using a conversion ratio of 1:20, which means every 20 pixels in the image corresponds to 1 meter in the actual space, enabling us to construct a detailed semantic map.

Deterministic Navigation Policy In RILA, navigation toward a designated waypoint is governed by a deterministic policy. Given that the unit of forward movement is set at one meter, we have partitioned the semantic map into a discrete graph of dimensions 81×81, with each node encompassing an area of 20×20 pixels. A node is classified as visible within the graph if it contains more than $\frac{2}{3}$ of its pixels as either occupied or free. The connectivity between two nodes, represented by an edge, hinges on the absence of obstacles between the centers of these nodes. Subsequently, the shortest path is computed based on the accessibility of selected frontiers or the nearest reachable points, facilitating efficient navigation.

Region Layout Split For the region layout prediction in the Imaginative Assistant, we specifically segment the semantic map based on the locations of detected walls. This process adheres to more stringent criteria compared to the construction of occupied regions. Within the point cloud, we identify walls by locating consecutive segments that share the same horizontal position. This approach strikes a balance, effectively pinpointing walls while preserving the depth map's accuracy and avoiding excessive segmentation, which ensures that our Imaginative Assistant accurately delineates different areas, crucial for its functioning. Upon identifying walls within the point cloud, we proceed by extending and extracting continuous pixel segments until they intersect with other segments. These intersections are then established as definitive boundaries for various regions on the semantic map. Concurrently, we conduct an iteration over all detected objects, partitioning them into their respective regions based on the boundaries. This process effectively creates a semantic segmentation of the environment, laying down a structured framework for ImaAssistant. This segmented framework is instrumental for ImaAssistant in

understanding and predicting the spatial layout.

B.3. Imaginative Assistant

Following the delineation of a logical region layout with its associated objects, as identified by the walls, ImaAssistant proceeds to interpret the semantic details of these observed regions. This interpretation is guided by both the layout and semantic cues, which include information from the prompts containing bounding boxes and semantic CSS formats. In situations where the regions are only partially observed and lack complete enclosure, ImaAssistant engages its imaginative capabilities to infer and supplement these regions with reasonable bounding boxes. The synthesized information, encompassing both observed and imaginatively supplemented details, is subsequently relayed to RefPlanner. This integration into RefPlanner facilitates comprehensive exploration and strategy formulation for subsequent exploratory tasks, ensuring that RefPlanner has a holistic understanding of the environment for effective planning.

B.4. Reflective Planner

Frontier-based Exploration In our RILA framework, we employ a frontier-based strategy, central to which is RefPlanner in selecting the optimal frontier. This process comprises two main components: region suggestion and frontier planning. Region suggestion entails evaluating the potential of different regions for exploration in the next phase, based on the layout interpretations provided by ImaAssistant. Building on these suggestions, we compile a comprehensive list that includes all frontiers along with their associated regional semantics. Additionally, this list also integrates any supplemental objects located in the vicinity of these frontiers. Armed with this aggregated information, RefPlanner then proceeds to analyze and choose the most appropriate frontier for the upcoming exploration stage. To provide a clearer understanding of our approach, we illustrate a specific navigation instance of our agent in Figure 7, which showcases the detailed prompt template we employ.

C. Supplementary Experimental Result

C.1. Audio Perception

Audio Classification To analyze the audio samples, we apply STFT with specific parameters: a hop length of 160 samples and a window size of 512 samples. These parameters correspond to a time resolution of 0.032 seconds, considering a sample rate of 16,000 Hz. When processing one-second audio segments, this approach generates complex-valued matrices with a size of 257×101. Following the generation, we calculate their magnitudes and downsample these magnitudes, reducing the size of both dimensions to optimize the data for subsequent processing. Moreover, we sample 3344 one-second audio clips across 500 test episodes and compute the classification accuracy for 21 distinct goal objects respectively as shown in Table 5.

Object	Acc ↑	Count
bathtub	100.0	16
chair	99.7	652
counter	94.6	112
seating	100.0	12
sofa	100.0	124
toilet	85.7	28
bed	100.0	128
chest of drawers	92.0	88
cushion	85.9	376
picture	85.2	548
shower	91.7	12
stool	100.0	12
towel	98.8	84
cabinet	91.4	336
clothes	95.8	24
fireplace	75.0	4
plant	90.4	312
sink	100.0	104
table	99.3	300
tv monitor	76.4	72

Table 5. The accuracy results of audio classification for each specific object type within the test dataset. *Count* refers to the number of times each object appears within the test set.

Audio Localization We evaluate the difference in RMS values across 30,000 audio samples randomly selected from 500 episodes within the Soundspace test dataset. As mentioned in Section 4, when we deactivated the lowest level of weight, indicative of weak directional information, the accuracy in assessing left-right direction surpassed 73.7%.

C.2. Visual Perception

Object Recognition We evaluate object recognition with two metrics: recall and accuracy. Recall measures the proportion of ground truth objects that are successfully identified, while accuracy indicates the fraction of correctly identified objects among all recognized items. Furthermore, we make a distinction between goal objects and other objects to specifically assess the effectiveness of our prompt design. The evaluation results are detailed in Table 6. Notably, GroundingDINO demonstrates impressive results, achieving over 90% recall and over 80% accuracy in recognizing the predicted goal object. Additionally, our navigation process allows for the repeated observation of a single object at various stages, thereby ensuring reliable overall performance in object recognition.

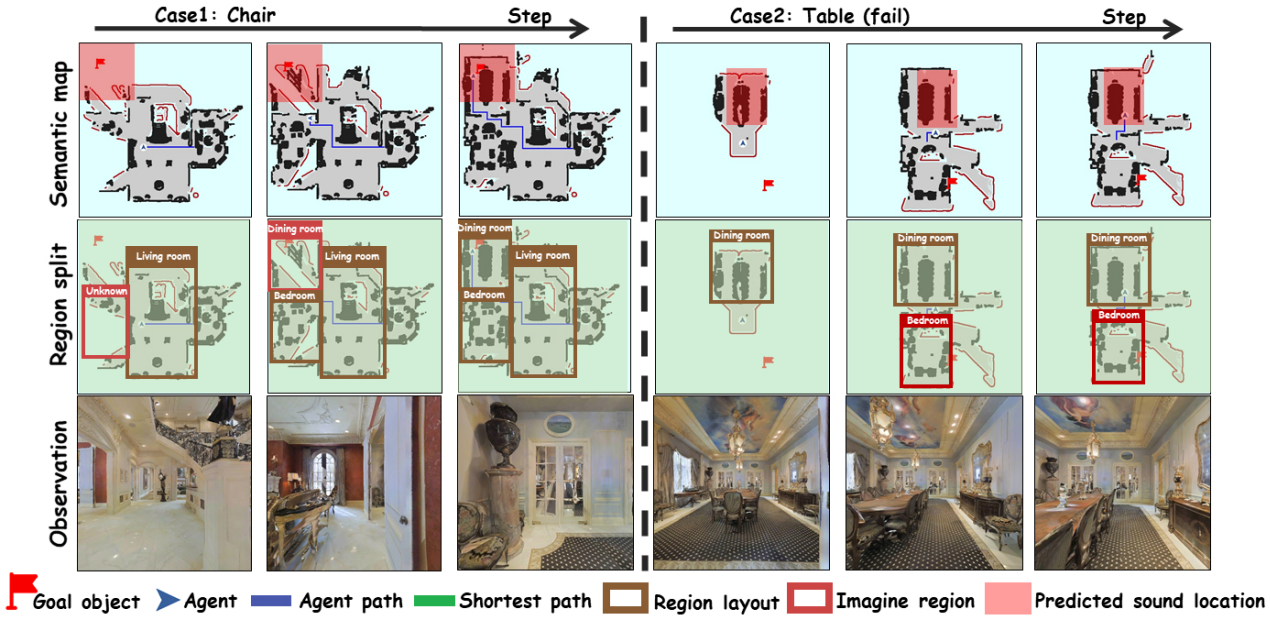

Figure 8. Two representative cases. The left figure illustrates a successful navigation process, whereas the right figure depicts a scenario where RILA navigates to an incorrect region, albeit with logically arranged layouts.

Object Type	Accuracy ↑	Recall ↑
Other Objects	85.0	62.2
Goal Prediction	83.9	91.6

Table 6. The accuracy and recall results of GoundingDINO in object recognition on the test dataset. *Goal Prediction* refers to the detection of the predicted goal object, while *Other Objects* encompasses the detection of all observed objects.

Method	SR (%) ↑	SPL (%) ↑	SWS (%) ↑
Random†	22.1	13.0	18.3
Nearest†	19.1	13.5	16.4
Llama-2 7B	24.8	11.9	22.3
Ours	35.4	11.8	11.4

Table 7. Ablation study on RefPlanner on the test dataset by replacing it with heuristic frontier selection methods and replacing the ChatGPT with Llama-2. † indicates using oracle stop.

Method	SR ↑	SPL ↑	DTG ↓
Ours	35.4	11.8	11.4
+ GT Audio Perception	51.0	23.4	7.3
+ GT Visual Perception	60.4	35.8	5.7

Table 8. Comparison of incorporating different ground-truth perceptions on the test dataset. Experiments in each row include the ground-truth information from all previous rows.

C.3. RefPlanner

In this section, we present supplementary experimental results of the ablation study. Table 7 illustrates the comparative analysis of various planning strategies on the test dataset, specifically utilizing the perception modules integrated within our framework. In contrast, Table 2 employs ground truth perceptions for its analysis. Table 7 indicates that RefPlanner effectively navigates to the target, which is in line with the results shown in Table 2.

Similarly, we evaluate RILA with ground truth perceptions, as presented in Table 8. Consistent with Table 4, RILA demonstrates exceptional planning performance when integrating with ground truth perceptions. This consistency underscores the current most significant limitation of RILA, its reliance on audio perception capabilities.

C.4. Other Results

Noisy Environments To simulate noisy environments, we adopt the distractor setting in the SAVN task. For low-light condition, we adjust the RGB inputs by reducing the brightness by half. These simulations affect primarily the Perception Module. Therefore, we provide a comparison in

Table 9, which indicates that these modules maintain competitive performance. Moreover, our agent naturally operates with potentially inaccurate perception, ensuring consistent performance in noisy settings.

(visual)	Default	Low-light
Object Recognition	83.9%	79.1%
(auditory)	Default	Noisy
Audio Classification	93.0%	82.9%
Audio Distance	83.8%	80.9%
Audio Direction	73.7%	75.2%

Table 9. Comparison of accuracy results of perception modules under regular and low-light environments.

More Scenes We further evaluate our methods in 10 unseen scenes from the val split. According to Table 10, our agent retains competitive performance. It is noteworthy that our agent operates in a zero-shot manner, which enables it to seamlessly generalize to varied unseen scenarios.

Scenarios	SR (%) ↑	SPL (%) ↑	SWS (%) ↑
Test (Default)	35.4	11.8	20.4
Val (10 unseen)	36.2	12.1	30.8

Table 10. Results on 10 unseen scenes from the val split.

C.5. Case Study

In this section, we provide two examples of RILA's navigation process, as depicted in Figure 8. The left figure demonstrates RILA's capability to accurately identify the correct region over long distances, utilizing visual cues and benefiting from spatial cognition. On the other hand, the right figure presents a typical instance of navigation failure. In this case, despite accurately inferring the layout, RILA erroneously navigates to the dining room in search of the table, based on semantic relationships, rather than heading to the bedroom, the intended target region. Overall, these examples indicate that, while generally effective, RILA's navigation can be subject to specific errors in decision-making.

References

[1] Neil Aaronson and William Hartmann. Testing, correcting, and extending the Woodworth model for interaural time difference. *The Journal of the Acoustical Society of America*, 135, 2014. 8

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as I can and not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[3] Ziad Al-Halah, Santhosh K. Ramakrishnan, and Kristen Grauman. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. *arXiv preprint arXiv:2202.02440*, 2022. 2

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. 2

[5] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *arXiv preprint arXiv:1505.04406*, 2017. 5

[6] Joshua G. W. Bernstein, Olga A. Stakhovskaya, Gerald I. Schuchman, Kenneth Kragh Jensen, and Matthew J. Goupell. Interaural time-difference discrimination as a measure of place of stimulation for cochlear-implant users with single-sided deafness. *Trends in Hearing*, 22, 2018. 3

[7] G.C. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, 1987. 1

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments, 2017. 1

[9] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. *arXiv preprint arXiv:2012.11583*, 2020. 2, 5

[10] 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*, pages 17–36. Springer, 2020. 2, 5, 6, 1

[11] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, pages 15516–15525, 2021. 1, 2, 3, 5, 6

[12] 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. 5, 6

[13] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. SoundSpaces 2.0: A simulation platform for visual-acoustic learning. *NeurIPS Datasets and Benchmarks Track*, 2022. 2, 5

[14] Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas H. Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-

language navigation. *arXiv preprint arXiv:2210.07506*, 2022. 2

[15] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. $\dot{a}$2 nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *arXiv preprint arXiv:2308.07997*, 2023. 1

[16] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An open simulation-to-real embodied AI platform. In *CVPR*, 2020. 1

[17] Vishnu Sashank Dorbala, James F. Mullen Jr. au2, and Dinesh Manocha. Can an embodied agent find your "cat-shaped mug"? LLM-guided exploration for zero-shot object navigation. *arXiv preprint arXiv:2303.03480*, 2023. 2

[18] Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023. 2

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all, 2023. 5

[20] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. ImageBind-LLM: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 1, 5, 6

[21] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023. 2

[22] David B. Hawkins, Lamar L. Young, and Cheryl Parker. An investigation of the interaural time difference threshold for speech. *Perception & Psychophysics*, 24:168–170, 1978. 3

[23] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting textured 3D meshes from 2D text-to-image models. In *ICCV*, 2023. 2

[24] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. *arXiv preprint arXiv:2307.12981*, 2023. 1

[25] Maike Klingel, Norbert Kopčo, and Bernhard Laback. Reweighting of binaural localization cues induced by lateralization training. *JARO: Journal of the Association for Research in Otolaryngology*, 22:551 – 566, 2020. 8

[26] Christine Köppl and Catherine Emily Carr. Maps of interaural time difference in the chicken's brainstem nucleus laminaris. *Biological Cybernetics*, 98:541–559, 2008. 8

[27] Mingxiao Li, Zehao Wang, Tinne Tuytelaars, and Marie-Francine Moens. Layout-aware dreamer for embodied referring expression grounding. *arXiv preprint arXiv:2212.00171*, 2022. 2

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[30] Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. WebGLM: Towards an efficient web-enhanced question answering system with human preferences. *arXiv preprint arXiv:2306.07906*, 2023. 1

[31] Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. Active sparse conversations for improved audio-visual embodied navigation. *arXiv preprint arXiv:2306.04047*, 2023. 1, 2

[32] Louise Loiselle, Michael Dorman, William Yost, Sarah Natale, and René Gifford. Using ILD or ITD cues for sound source localization and speech understanding in a complex listening environment by listeners with bilateral and with hearing-preservation cochlear-implants. *Journal of Speech Language and Hearing Research*, 59:1, 2016. 8

[33] 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. 2, 5

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 6

[35] OpenAI. Introducing ChatGPT, 2022. (Accessed on Jun 18, 2023). 1

[36] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[37] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics (ACL)*, pages 2086–2105, Dublin, Ireland, 2022. Association for Computational Linguistics. 1

[38] Sudipta Paul, Amit K Roy-Chowdhury, and Anoop Cherian. AVLEN: Audio-visual-language embodied navigation in 3d environments. In *NeurIPS*, 2022. 1, 2, 5, 6

[39] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR, 2023. 2

[40] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2023. 1

[41] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam,

Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021. 2, 5

[42] Andrew Szot, Max Schwarzer, Bogdan Mazoure, Harsh Agrawal, Walter Talbott, Katherine Metcalf, Natalie Mackraz, Devon Hjelm, and Alexander Toshev. Large language models as generalizable policies for embodied tasks. *arXiv preprint arXiv:2310.17722*, 2023. 2

[43] Yujin Tang, Wenhao Yu, Jie Tan, Heiga Zen, Aleksandra Faust, and Tatsuya Harada. SayTap: Language to quadrupedal locomotion, 2023. 2

[44] Gyan Tatiya, Jonathan Francis, Luca Bondi, Ingrid Navarro, Eric Nyberg, Jivko Sinapov, and Jean Oh. Knowledge-driven scene priors for semantic audio-visual embodied navigation. *arXiv preprint arXiv:2212.11345*, 2022. 1, 2, 3, 5, 6

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[46] Balemir Uragun and Ramesh Rajan. The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modelling. *BMC Neuroscience*, 14:114 – 114, 2013. 8

[47] Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, Qingqing Zhu, Zhenzhu Yang, Adam Nik, Qi Liu, Chenghua Lin, Shi Wang, Ruibo Liu, Wenhu Chen, Ke Xu, Dayiheng Liu, Yike Guo, and Jie Fu. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*, 2023. 1

[48] Justin Wasserman, Karmesh Yadav, Girish Chowdhary, Abhinav Gupta, and Unnat Jain. Last-mile embodied visual navigation. In *Conference on Robot Learning*, 2022. 1

[49] Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848*, 2023. 1

[50] Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, Aaron Gokaslan, Oleksandr Maksymets, Rishabh Jain, Ram Ramrakhya, Angel X Chang, Alexander Clegg, Manolis Savva, Eric Undersander, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2022. https://aihabitat.org/challenge/2022/, 2022. 1

[51] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. LLM-Grounder: Open-vocabulary 3D visual grounding with large language model as an agent. *arXiv preprint arXiv:2309.12311*, 2023. 1

[52] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *arXiv preprint arXiv:2111.14843*, 2023. 2

[53] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention, 2023. 5

[54] Zhen Zhang, Anran Lin, Chun Wai Wong, Xiangyu Chu, Qi Dou, and KW Au. Interactive navigation in environments with traversable obstacles using large language and vision-language models. *arXiv preprint arXiv:2310.08873*, 2023. 2

[55] Gengze Zhou, Yicong Hong, and Qi Wu. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023. 1

[56] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. ESC: Exploration with soft commonsense constraints for zero-shot object navigation. *arXiv preprint arXiv:2301.13166*, 2023. 1, 2, 5, 6