Audio Visual Language Maps for Robot Navigation

Chenguang Huang[1]*, Oier Mees[1], Andy Zeng[2], and Wolfram Burgard[3]

[1] University of Freiburg, Germany,
[2] Google Research, USA,
[3] University of Technology Nuremberg, Germany.

Abstract. While interacting with the world is a multi-sensory experience, many robots continue to predominantly rely on visual perception to map and navigate in their environments. We propose AVLMaps, a 3D spatial map representation that stores cross-modal information from audio, visual, and language cues. AVLMaps fuse features from pre-trained multimodal foundation models into a multi-layer representation. This enables robots to index goals in the map based on multimodal queries, such as textual descriptions, images, or audio snippets of landmarks. AVLMaps allow for zero-shot multimodal spatial goal navigation and perform better than alternatives in ambiguous scenarios. These capabilities extend to mobile robots in the real world. Videos and code are available at https://avlmaps.github.io.

Keywords: multimodal semantic mapping, language-based navigation, open-vocabulary indexing

1 Introduction

Humans efficiently integrate multiple sensing modalities to navigate the physical world. Acoustic signals, such as the sound of breaking glass or a buzzing microwave, provide valuable complementary information. This is evident in the utility it offers to the visually impaired for navigation. In contrast, current mobile robots heavily rely on visual, LiDAR, or ultrasound perception in human-centered environments. How to effectively incorporate audio signals as an additional sensing modality for cross-modal reasoning in robotics tasks remains a relevant research question.

To address this issue, we propose Audio-Visual-Language Maps (AVLMaps), a unified 3D spatial map representation that stores cross-sensing information from audio, visual, and language modalities. AVLMaps are constructed from image and audio observations by computing dense features from multimodal foundation models trained on Internet-scale data [15,11]. An AVLMap enables indexing of landmark locations using multimodal queries, such as textual descriptions, images, or audio snippets, facilitating language-based goal-driven navigation without model fine-tuning, e.g., "go in between the sound of the breaking glass and {the image of a refrigerator}" as in Fig. 1. By including audio information, AVLMaps allow robots to accurately disambiguate goal locations using sound cues in scenarios with multiple similar objects, e.g., "go to the table where you heard someone coughing", when there are multiple tables in the scene.

*The corresponding author's email: huang@informatik.uni-freiburg.de

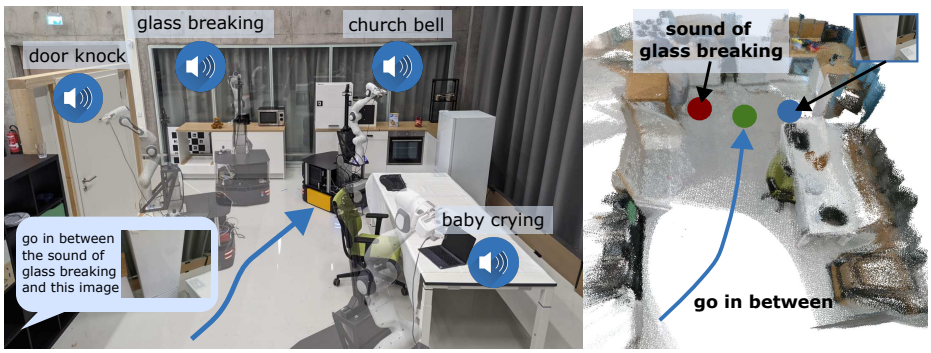


Fig. 1. AVLMaps provide an open-vocabulary 3D map representation for storing cross-modal information from audio, visual, and language cues. When combined with large language models, AVLMaps consume multimodal prompts from audio, vision and language to solve zero-shot spatial goal navigation by effectively leveraging complementary information sources to disambiguate goals.

2 Related Work

The combination of traditional SLAM techniques and advancements in vision-based semantic understanding has resulted in the enhancement of 3D maps with semantic information. Previous approaches focused on abstracting the map at the object level [17]. However, these methods are limited to predefined semantic classes. Recent works have demonstrated the integration of visual-language features into occupancy maps, allowing open-vocabulary object indexing with natural language and freeing the maps from fixed semantic categories [12,5]. However, these approaches focus solely on visual perception, disregarding complementary information sources like acoustic signals.

Recent advances in simulation applications [23] have fueled research on multimodal navigation in two main directions: (i) vision-and-language navigation (VLN) [1] where an agent needs to follow a natural language instruction towards the goal with visual input, and (ii) audio-visual navigation (AVN) [7] in which an agent should navigate to the sound source based on information from a binaural sensor and vision. Despite different degrees of success in both directions [10,8], less attention has been paid to solving the navigation problem involving vision, language, and audio at the same time. The most relevant concept to our knowledge is from AVLEN [19], which extends the AVN with a further query step, introducing a language instruction that helps with navigating to the sound source. In addition, most of the existing methods on AVN focus on approaching the sound without understanding its semantics.

Recent trends have shown that pre-trained models [21,3] serve as powerful tools for robotic tasks including object detection and segmentation [9,14], robot manipulation [16,18], and navigation [6,12]. Most related to our work are approaches like VLMaps [12], NLMap [6], and ConceptFusion [13], all of which combine pre-trained visual-language models with a 3D reconstruction of

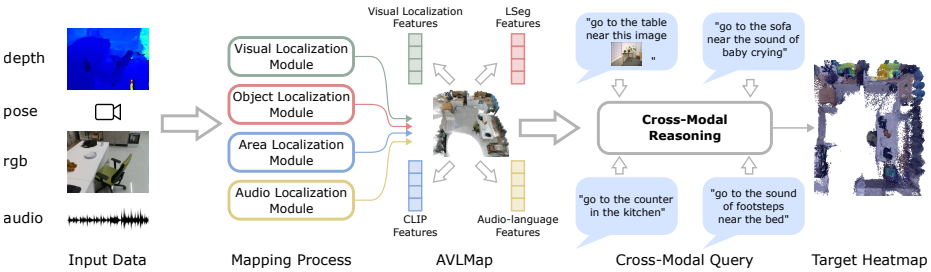


Fig. 2. System overview. AVLMaps are constructed from RGB-D, audio, and odometry inputs, converting raw data into visual localization features, visual-language features, and audio-language features. During inference time, each module's output is unified with cross-modal reasoning, allowing users to query spatial location with multimodal information.

the scene, enabling landmark indexing with natural language and downstream language-based planning tasks.

In contrast to previous methods, AVLMaps integrate audio, visual, and language cues into a 3D map, enabling the agent to navigate to various multimodal goals and effectively disambiguate them, enabling a robot to navigate to multimodal goals specified with either goal image or natural language like "go to the sound of baby crying", "go to the table" or multimodal prompts such as "go to the {image of a table} where the sound of the microwave was heard".

3 Method

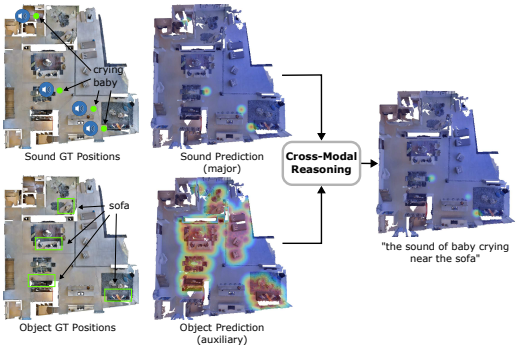
We aim to create an audio-visual-language map that can directly localize objects, areas, audio, and visual goals using natural language or target images. We propose AVLMaps by combining 3D reconstruction libraries with pre-trained visual-language and audio-language models. We also suggest a cross-modal reasoning approach to disambiguate locations referring to targets from different modalities. Fig. 2 shows the system pipeline.

3.1 Building an Audio Visual Language Map

Given an RGB-D video stream with an audio track and odometry information, we utilize four modules to build a multimodal feature database as AVLMaps.

Our **Visual Localization Module** follows a hierarchical scheme to localize a query image in the map. It involves computing global NetVLAD features [2] and local SuperPoint descriptors for images [22], finding a candidate reference image through nearest neighbor search, establishing key point correspondences using SuperGLUE [22], obtaining 3D-2D correspondences, and estimating the query camera pose relative to the reference camera using the Perspective-n-Point method.

Our **Object Localization Module** uses an open-vocabulary segmentation method (e.g., LSeg [15]) to generate pixel-level features from the RGB image and associates them with back-projected depth pixels in 3D reconstruction. During inference, it encodes a target text query [21], computes the cosine similarity



(a) Cross-Modal Reasoning.

```
# move to the middle of the sound of cat meowing and
the image: /path/to/image.png
img = robot.load_image("/path/to/image.png")
sound_map = robot.get_major_map(sound="cat meowing")
img_map = robot.get_major_map(img=img)
pos1 = robot.get_max_pos_3d(sound_map)
pos2 = robot.get_max_pos_3d(img_map)
pos = (pos1 + pos2) / 2
robot.move_to(pos)
# move to the window next to the sound of
glass breaking
obj_map = robot.get_major_map(obj="window")
sound_map = robot.get_map(sound="glass breaking")
fuse_map = obj_map * sound_map
pos = robot.get_max_pos_3d(fuse_map)
robot.move_to(pos)
...

# move in between the image ./006899.png and the
backpack near the sound of glass breaking
img = robot.load_image("./006899.png")
img_map = robot.get_major_map(img=img)
obj_map = robot.get_major_map(obj="backpack")
sound_map = robot.get_map(sound="glass breaking")
fuse_map = obj_map * sound_map
pos1 = robot.get_max_pos_3d(img_map)
pos2 = robot.get_max_pos_3d(fuse_map)
pos = (pos1 + pos2) / 2
robot.move_to(pos)
```

(b) Prompts for LLM.

Fig. 3. As in 3a, the key idea of cross-modal reasoning is converting the predictions from different modalities into heatmaps, and then fusing them with element-wise multiplication, effectively using complementary multimodal information to resolve ambiguous prompts. 3b shows the prompts to GPT-3 to generate executable code for multimodal goal navigation (prompt in gray, input task commands in green, and generated outputs are highlighted).

scores between all point-wise and language features, and selects the top-scoring points in the map as the indexing result.

The **Area Localization Module** builds a sparse topological CLIP feature map to recognize coarse visual concepts like "kitchen area". During inference, given a language concept, we compute the language features with the CLIP language encoder [21] and image-to-language cosine similarity scores to predict the location with confidence values.

The **Audio Localization Module** partitions the audio clip from the audio stream input into several segments using silence detection and computes audio-lingual features for each segment with AudioCLIP [11]. During inference, given a language description, it computes matching scores between the language and all audio segments. The odometry associated with the top-scoring segment is the predicted location.

3.2 Cross-Modality Reasoning

A key advantage of our method is its capability to disambiguate goals with additional information, even from different modalities. Given a specific query, each module introduced in the previous section returns predicted spatial locations on the map in the form of 3D voxel heatmaps. A heatmap can be denoted as $\mathcal{H} \in [0,1]^{\bar{H}\times\bar{W}\times\bar{Z}}$, where $\bar{H}$, $\bar{W}$ and $\bar{Z}$ represent the size of the voxel map and the value in each element represents the probability of being the target position. $\mathbf{p} = (x,y,z)^T, \{x,y,z \in \mathbb{Z} \mid 1 \le x \le \bar{H}, 1 \le y \le \bar{W}, 1 \le z \le \bar{Z}\}$ is a voxel position in the map $\mathcal{H}$.

Visual Localization Heatmap. In the visual localization module, the predicted global camera location is denoted as $\mathbf{p}_v = (x_v, y_v, z_v)^T$. In the heatmap $\mathcal{H}_v$, we define the probability at $\mathbf{p}_v$ as 1.0, and the probability linearly decays

around this location according to the distance on the top-down map:

$$\mathcal{H}_v(\mathbf{p}) = \max(1.0 - \epsilon \cdot dist_{xy}(\mathbf{p}, \mathbf{p}_v), 0) \tag{1}$$

$$dist_{xy}(\mathbf{p}, \mathbf{q}) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \tag{2}$$

where ϵ is the decay rate, and $dist_{xy}(\mathbf{p}, \mathbf{q})$ denotes the distance between 3D vectors $\mathbf{p}$ and $\mathbf{q}$ on the xy-plane.

Object Localization Heatmap. The object localization results are a list of points, denoted as $\{\mathbf{p}_{oi} = (x_{oi}, y_{oi}, z_{oi}) \mid i = 1, \ldots, N\}$ where N is the total number of points for the target object. We define the probabilities for all these locations as 1.0 in heatmap $\mathcal{H}_o$, and the probability linearly decays around these locations based on the Euclidean distance:

$$d_{min}(\mathbf{p}) = \min\{dist(\mathbf{p}, \mathbf{p}_{oi}) \mid i = 1, \ldots, N\} \tag{3}$$

$$\mathcal{H}_o(\mathbf{p}) = \max(1.0 - \epsilon \cdot d_{min}(\mathbf{p}), 0) \tag{4}$$

where $d_{min}(\mathbf{p})$ denotes the minimal distance between $\mathbf{p}$ and all object points $\{\mathbf{p}_{oi} \mid i = 1, \ldots, N\}$, $dist(\mathbf{p}, \mathbf{q})$ denotes the Euclidean distance between $\mathbf{p}$ and $\mathbf{q}$.

Area Localization Heatmap. The area localization results are a list of position-confidence pairs, denoted as $\{(\mathbf{p}_{ai}, s_{ai}) \mid i = 1, \ldots, M\}$ where M is the total number of frames in the input RGB-D stream. The scores s_{ai} are normalized between 0 and 1. We define the probability for each point $\mathbf{p}_{ai}$ on the heatmap $\mathcal{H}_a$ as its score s_{ai}, and the probability linearly decays around the point on the xy-plane direction:

$$\mathcal{H}_a(\mathbf{p}) = \max(\max\{s_{ai} - \epsilon \cdot dist_{xy}(\mathbf{p}, \mathbf{p}_{ai}) \mid i = 1, \ldots, M\}, 0) \tag{5}$$

where the max operator for the curly brackets means taking the highest probability when a location is inside the affected regions for several $\mathbf{p}_{ai}$.

Audio Localization Heatmap. The audio localization results are similar to those of the area localization module. The position-score pairs are denoted as $\{(\mathbf{p}_{si}, s_{si}) \mid i = 1, \ldots, K\}$ where K is the total number of sound segments in the input video stream. The heatmap $\mathcal{H}_s$ is defined as:

$$\mathcal{H}_s(\mathbf{p}) = \max(\max\{s_{si} - \epsilon \cdot dist_{xy}(\mathbf{p}, \mathbf{p}_{si}) \mid i = 1, \ldots, K\}, 0) \tag{6}$$

Cross-Modal Reasoning. The main idea of cross-modal reasoning is shown in Fig. 3a. We treat the predictions from four modules as four modalities. When there are several queries referring to different modalities, we compute the respective heatmaps first and then perform element-wise multiplication assuming conditional independence among all heatmaps:

$$\mathcal{H}_{target} = \mathcal{H}_1 \odot \mathcal{H}_2 \odot \ldots \odot \mathcal{H}_L \tag{7}$$

where $\odot$ is the element-wise multiplication operator, and L is the total number of referred modalities. We extract the position on the target heatmap $\mathcal{H}_{target}$ that has the highest probability as the predicted location.

When we compute the heatmaps, we design that there is always a primary heatmap while others are auxiliary ones. To illustrate this, consider the query "navigate to the chair near the sound of crying". In this case, the target object we intend to approach is "the chair", so its corresponding heatmap is designated as the primary heatmap, while the heatmap for "the sound of crying" serves as an auxiliary heatmap. Conversely, when the query is "navigate to the sound of crying near the chair", the roles will be reversed in the results. We set the decay rate for the primary heatmap higher (e.g., 0.1 in this work) since we want to know the exact location of the target while tuning the decay rate for the auxiliary heatmap lower (e.g., 0.01) as having a broader effect area to narrow down major targets is desirable.

3.3 Multimodal Goal Navigation from Language

In the setting of multimodal goal navigation from language, the agent is given language descriptions of targets from different modalities (e.g., sound, image, and object) and is required to plan paths to them. While most of the previous navigation methods focus mainly on a specific type of goal, we unify these tasks with the help of large language models (LLMs). Specifically, we use an LLM to interpret the natural language commands and synthesize API calls combined with simple logic structures in the form of executable python code [16,12,18]. For heatmap generation, we implement interfaces ***get_major_map****(obj=None, sound=None, img=None)* and ***get_map****(obj=None, sound=None, img=None)*. They take the object name, sound name, or image as input and output heatmaps indicating the locations of targets. The ***get_major_map*** generates heatmaps with higher decay rate while ***get_map*** with lower decay rate. To support the image prompt, we add an image path in the language query like "the image img_path.png" and use LLMs to call the image loading API. Some examples of prompts and queries are shown below (prompt in gray, input task commands in green, and generated outputs are highlighted):

4 Experiments

This section presents our experiments conducted in both simulation and real-world environments. We begin by describing the simulation setup in Sec. 4.1. Next, we show the results of our experiments on multimodal goal navigation in Sec. 4.2. In addition, we present the results of our experiments on cross-modal goal indexing and navigation in Sec. 4.3 and Sec. 4.4. Finally, we discuss the details of our real-world experiments in Sec. 4.5.

4.1 Simulation Setup

Exerimental setup. We use the Habitat simulator [23] with the Matterport3D dataset [4] for the evaluation of multimodal navigation tasks. For mapping purposes, we manually collect RGB-D video streams in the simulator across ten different scenes and add random audio tracks to the videos to simulate the audio sensing modality. All audio comes from the validation fold (Fold-1) of the ESC-50 dataset [20]. In navigation tasks, the robot has four actions to take: **forward 0.1 meters**, **left/right 5 degrees**, and **stop**. In sequential goal setting,

the robot is required to navigate to a sequence of goals and take the **stop** action when it reaches each subgoal. The subgoal is considered successfully reached when the stop position is less than one meter away from the ground truth position.

Tasks collection. In multimodal goal navigation tasks in Sec. 4.2, we consider three kinds of goals: image goals, object goals, and sound goals. For image goals, we randomly sample positions and orientations on the top-down map and render images as targets. For object goals, we access the metadata (e.g., bounding boxes and semantics) from the Matterport3D dataset and sample a list of categories in each scene as queries. For sound goals, we randomly sample sound classes of audio merged with the mapping videos as targets, treating the video frame positions as the ground truth. In cross-modal goal indexing tasks in Sec. 4.3, we collect three types of datasets:

- **Visual-Object cross-modal indexing** We manually select image-object pairs on the top-down map for localizing "an object X near the image Y".
- **Area-Object cross-modal indexing** We access the region and object metadata (e.g., bounding boxes and semantics) from the Matterport3D dataset to automatically generate a list of object-region pairs. This dataset is for localizing "an object X in the area of Y".
- **Object-Sound cross-modal indexing** We manually insert several sounds of the same kind into a scene and select for each sound location a nearby object for disambiguation. The query is "a sound X near the object Y".

In cross-modal goal navigation in Sec. 4.4, we randomly sample starting pose in 10 scenes and treat the visual-object and object-sound cross-modal goals in Sec. 4.3 as navigation goals.

4.2 Multimodal Goal Navigation

Sound goal navigation. We first test AVLMaps in sound goal navigation tasks. We collect 200 sequences of sound goals in 10 different scenes. In each sequence, there are 4 sound categories that require the robot to reach. The results are shown in Table 1. We generate AudioCLIP [11] features with our audio localization module and match all audio with the target sound category in the embedding space, similar to a text-to-audio retrieval setup. Then the agent plans a path to the audio position. We tested different ranges of sound categories inserted into the map. The full list of sound categories in each major class can be found in the link[4]. The results show that our agent manages to recognize sound goals and navigate with a 77.5% success rate.

Visual and object goals navigation. We then test AVLMaps with visual and object goal navigation tasks. The agent is given an image and two object categories in the language in one sequence of tasks and asked to navigate to

[4] https://github.com/karolpiczak/ESC-50

the image goal and two object goals in sequence. In 200 sequences of tasks in 10 scenes, the success rate is reported in Table 2. The results show that our method enables the agent to navigate to goals from different modalities.

Table 1. The success rate (%) of sound goal navigation with AVLMaps.

Tasks	No. Subgoals in a Row				Independent
	1	2	3	4	Subgoals
Domestic Sound	59.5	33.0	15.5	7.0	62.5
+ Human Sound	69.5	47.0	36.5	23.0	72.38
+ Animal Sound	74.5	58.5	45.5	33.0	77.5

Table 2. The success rate (%) of multimodal goal navigation with AVLMaps. The agent is required to navigate to one visual goal and two object goals in sequence.

Tasks	No. Subgoals in a Row			Independent
	1	2	3	Subgoals
AVLMaps (Ours)	71.5	40.5	25.0	47.4

4.3 Cross-Modal Goal Indexing

Area-Object goal indexing. In this setup, we use an area description to disambiguate the object goal. We collected 100 indexing tasks in 10 scenes. Each task consists of an object category and a region category (e.g., "living room", "kitchen", "dining room", "bathroom" etc.). The agent needs to predict the correct object location which is inside the region. The top-1 recall with different distance tolerance is reported in Table 3. We can notice that VLMaps [12] struggles to find the goal in the correct region because VLMaps integrates visual-language features from the encoder fine-tuned on the instance segmentation dataset, improving its segmentation performance on common objects while dropping its ability to recognize more general concepts like regions. In contrast, our area localization module integrates pre-trained CLIP features into the map without fine-tuning, enabling it to recognize general concepts including regions, and thus the indexing results are improved.

Object-Sound goal indexing. In this setting, we use object goals to disambiguate sound goals. We collected 119 indexing tasks, each of which consist of a sound category and a nearby object category. Each sound category in a scene can be heard at more than 1 location, introducing ambiguity to the localization scenario. The recall is reported in Table 4. With the combination of object and audio localization modules, our method largely increases the recall rate for localizing the correct sound goal position in ambiguous scenarios.

Visual-Object goal indexing. In visual-object goal indexing tasks, visual clues are used to resolve ambiguity. Given an object category and an image, our method can localize the correct object near the image position with over 60% of recall for 0.5 meters distance tolerance, as is shown in Table 5.

Table 3. The recall (%) of area-object cross-modal indexing.

Method	Recall@1 (%)				Average min. distance (m)
	<0.5m	<1m	<1.5m	<2m	
baseline (VLMaps)	5.56	7.78	13.33	17.78	8.22
+ ConceptFusion	12.22	13.33	16.67	21.11	7.60
+ CLIP sparse map (Ours)	**15.56**	**24.44**	**31.11**	**35.56**	**6.17**
+ GT region map	37.78	44.44	55.56	61.11	2.62

Table 4. The recall (%) of object-sound cross-modal indexing.

Method	Recall@1 (%)				Average min. distance (m)
	<0.5m	<1m	<1.5m	<2m	
baseline (wav2clip [24])	8.40	10.08	10.92	14.29	8.52
baseline (AudioCLIP [11])	26.05	35.29	36.97	42.01	5.04
VLMaps + wav2clip	24.37	30.25	33.61	38.66	6.27
VLMaps + AudioCLIP (Ours)	**53.78**	**65.55**	**67.23**	**70.59**	**2.74**

4.4 Multimodal Ambiguous Goal Navigation

We collected 119 sequences of ambiguous goal navigation tasks. In each task, the agent is required to navigate to an ambiguous sound goal (e.g., "the sound X near object Y") and an ambiguous object goal (e.g., "the object X near sound Y") sequentially. We consider two single-modality baselines: VLMaps [12] and AudioCLIP [11] and one multimodal baseline. The multimodal baseline uses VLMaps as the object localization module, wav2clip [24] as the audio localization module and the same visual localization module as our method. The results are shown in Table 6. We observe that AVLMaps navigate to cross-modal goals with 24.2% higher success rate to ambiguous sound goals and with 2.1% higher success rate to ambiguous object goals compared to the alternative multimodal baseline.

4.5 Real World Experiment

Environment setup. We control a mobile robot to record RGB-D videos in a room with multiple ambiguous goals such as tables, chairs, and paper boxes. Then we artificially add sounds to the RGB-D video when the robot moves to certain locations. After collecting the data, we run the AVLMaps mapping offline. For navigation tasks, we provide the AVLMap and the language instruction as input. The robot parses the instruction (Sec. 3.3) and executes the generated python code for goal indexing and planning. We use the ROS navigation package for global and local planning.

Multimodal Spatial Goal Reasoning and Navigation with Natural Language. We design 20 language-based multimodal navigation tasks, asking the robot to navigate to sounds, images, and objects. We report an overall success rate of 50%. We also design an evaluation consisting of ten multimodal spatial goals. The agent needs to reason across objects, sounds, images, and spatial

Table 5. The recall (%) of visual-object cross-modal indexing.

Method	Recall@1 (%)				Average min. distance (m)
	<0.5m	<1m	<1.5m	<2m	
VLMaps w/o vis loc	7.55	9.43	11.32	11.94	11.22
VLMaps w/ vis loc (Ours)	**62.26**	**66.67**	**70.44**	**72.32**	**3.11**

Table 6. The success rate (%) of multimodal ambiguous goal navigation.

Method	No. Subgoals in a Row		Sound Goals	Object Goals
	1	2		
VLMaps [12]	-	-	-	27.1
AudioCLIP [11]	-	-	16.9	-
VLMaps + wav2clip	22.0	12.7	22.0	53.4
VLMaps + AudioCLIP (Ours)	**46.2**	**28.6**	**46.2**	**55.5**

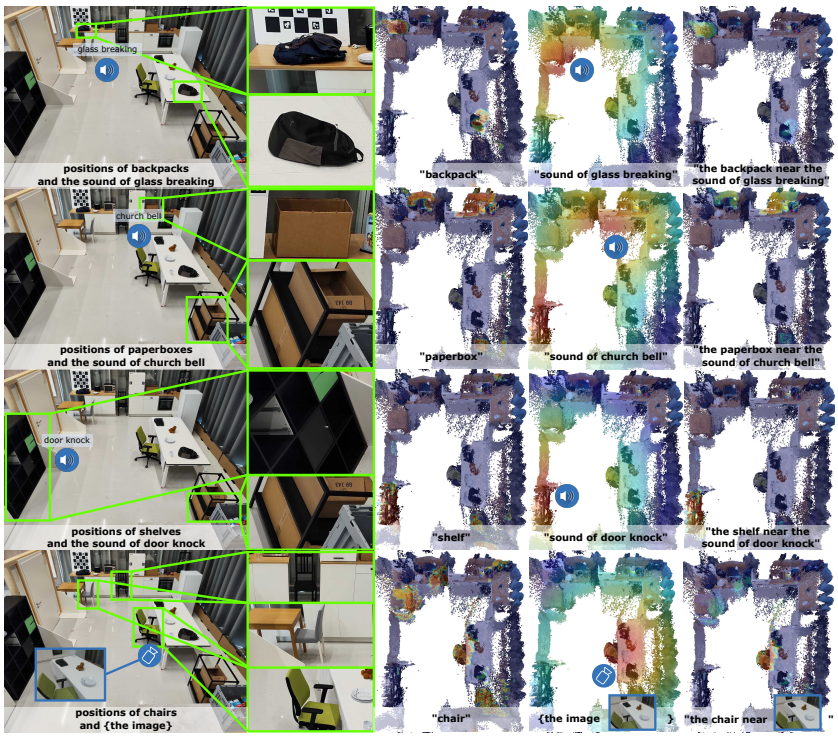


Fig. 4. Visualization of heatmaps in AVLMaps for multimodal goal reasoning for ambiguous object goals. From left to right: scene overview (objects/sounds/images locations), close-up view of ambiguous objects, predicted heatmap for the object target, predicted heatmap for the audio/visual target, and the predicted heatmap for the multimodal goal. The heatmap is shown in the JET color scheme.

concepts. An example is "navigate in between the backpack near the sound of glass breaking and {the image of a fridge}". In the end, six out of ten tasks were successfully finished. Fig. 4 shows the cross-modal reasoning results in real-world environments. More results can be found on our Website.

5 Conclusion

In this paper, we presented AVLMaps, a unified 3D spatial map representation that effectively incorporates cross-modal information from audio, visual, and language cues. By leveraging multimodal prompts, AVLMaps enable zero-shot spatial goal navigation and improve target indexing accuracy compared to baselines, particularly in cases with ambiguous goals. However, AVLMaps do have limitations. They are sensitive to audio noise and assume a static environment throughout their lifespan. Future work will explore integrating lifelong learning capabilities into the agent to further automate multimodal spatial learning.

Acknowledgements

This work has been supported partly by the German Federal Ministry of Education and Research under contract 01IS18040BOML and the BrainLinks-BrainTools center. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - 428605208

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3674–3683 (2018)
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5297–5307 (2016)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
4. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV) (2017)
5. Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M.S., Stone, A., Kappler, D.: Open-vocabulary queryable scene representations for real world planning. arXiv preprint arXiv:2209.09874 (2022)
6. Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M.S., Stone, A., Kappler, D.: Open-vocabulary queryable scene representations for real world planning. arXiv preprint arXiv:2209.09874 (2022)
7. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pp. 17–36. Springer (2020)

8. Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S.K., Grauman, K.: Learning to set waypoints for audio-visual navigation. arXiv preprint arXiv:2008.09622 (2020)
9. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2021)
10. Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I., Schmid, C.: Airbert: In-domain pretraining for vision-and-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1634–1643 (2021)
11. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 976–980. IEEE (2022)
12. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK (2023)
13. Jatavallabhula, K.M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Li, S., Iyer, G., Saryazdi, S., Keetha, N., Tewari, A., et al.: Conceptfusion: Open-set multimodal 3d mapping. arXiv preprint arXiv:2302.07241 (2023)
14. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2021)
15. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022). URL https://openreview.net/forum?id=RriDjddCLN
16. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. arXiv preprint arXiv:2209.07753 (2022)
17. McCormac, J., Clark, R., Bloesch, M., Davison, A., Leutenegger, S.: Fusion++: Volumetric object-level slam. In: 2018 international conference on 3D vision (3DV), pp. 32–41. IEEE (2018)
18. Mees, O., Borja-Diaz, J., Burgard, W.: Grounding language with visual affordances over unstructured data. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK (2023)
19. Paul, S., Roy-Chowdhury, A.K., Cherian, A.: Avlen: Audio-visual-language embodied navigation in 3d environments. arXiv preprint arXiv:2210.07940 (2022)
20. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015–1018 (2015)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
22. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)
23. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
24. Wu, H.H., Seetharaman, P., Kumar, K., Bello, J.P.: Wav2clip: Learning robust audio representations from clip. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4563–4567. IEEE (2022)