Cross-modal Map Learning for Vision and Language Navigation

Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan,
Eleni Miltsakaki, Dan Roth, Kostas Daniilidis
University of Pennsylvania
{ggeorgak,karls,kwanchoo,sohamdan,elenimi,danroth,kostas}@seas.upenn.edu
Project webpage: https://ggeorgak11.github.io/CM2-project/

Abstract

We consider the problem of Vision-and-Language Navigation (VLN). The majority of current methods for VLN are trained end-to-end using either unstructured memory such as LSTM, or using cross-modal attention over the egocentric observations of the agent. In contrast to other works, our key insight is that the association between language and vision is stronger when it occurs in explicit spatial representations. In this work, we propose a cross-modal map learning model for vision-and-language navigation that first learns to predict the top-down semantics on an egocentric map for both observed and unobserved regions, and then predicts a path towards the goal as a set of waypoints. In both cases, the prediction is informed by the language through cross-modal attention mechanisms. We experimentally test the basic hypothesis that language-driven navigation can be solved given a map, and then show competitive results on the full VLN-CE benchmark.

1. Introduction

For mobile robots to be able to operate together with humans, they must be able to execute tasks that are defined not in the form of machine-readable scripts but rather in the form of human instructions. A very basic but challenging task is going from A to B. While robots have been quite successful in executing this task using metric representations, it has been more challenging for robots to execute semantic tasks like "go to the kitchen sink" or follow instructions that describe a path and associate actions with natural language, defined as the Vision-and-Language Navigation (VLN) task [4, 32, 33]. In VLN, the robot is given instructions and has to reach a goal making use of images of the environment that it can acquire along the way.

The dominant approach for VLN tasks has been using end-to-end pipelines from images and instructions to actions [17, 23, 31, 32]. While they can be attractive due to their simplicity, they are expected to implicitly learn end-to-end all navigation components such as mapping, planning, and control, and thus often require considerable amounts of training data. This approach to designing navigation systems is in direct contrast to research on human spatial navigation, which has shown that humans and other species build map-like representations of the environment to accomplish way-finding [41, 51]. However, multiple findings have shown that the ability to build cognitive maps and acquire spatial knowledge deteriorates when humans exclusively use ready to drive or walk paths to a goal [6]. On the other hand, studies have shown that humans build better spatial representations when presented with landmark-based navigation instructions rather than full paths [54]. Such spatial representations enable the recall of landmarks on an egocentric map weeks after the experiment. While this does not prove that humans build a map during wayfinding when following semantic instructions, it is a strong indication that they can anchor landmarks and other semantics to a map that they easily recall. Research in learning of mapping and planning in computer vision and robotics [22] has also shown that an end-to-end system encompasses semantic maps that naturally emerge in the learning process.

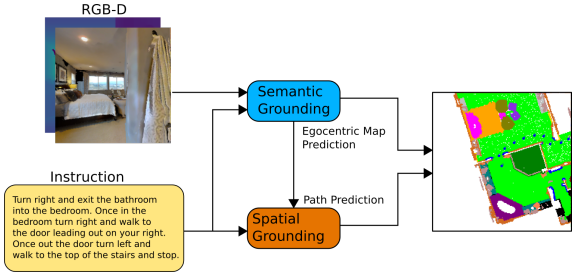

Figure 1. We approach the task of vision-and-language navigation as a two-stage procedure which learns to semantically and spatially ground the instruction on egocentric maps.

We propose *Cross-modal Map Learning* (CM^2), a novel navigation system for the VLN task in continuous environments, that learns a language-informed representation for both map and trajectory prediction by applying twice cross-

modal attention, hence CM^2. Our method decomposes the problem in the two paths of semantic and spatial grounding as illustrated in Figure 1. First, we use a cross-modal attention network to semantically ground the instruction through an egocentric map prediction task that learns to hallucinate information outside the field-of-view of the agent. This is followed by another cross-modal attention network that is responsible for spatially grounding the instruction by learning to predict the path on the egocentric map. Our analysis shows that through these two subtasks, the attended representations learn to focus on instruction-relevant objects and locations on the map.

The main difference between our method and existing image-language attentional mechanisms that generate actions is that in our approach, the robot is building a cognitive map that encodes the environmental priors and follows instructions based on this map. The motivation to use this representation is based on our finding that when the robot is given a local ground-truth ("correct") map of the environment, then the robot outperforms all approaches on the VLN task by a large margin. This map is still local, more like a crop of the blueprint than a global map of the environment, but can still hallucinate what is behind the walls so that it can align better the map with the language instruction. This differentiates us from approaches such as [12] that first build a topological map of the environment by exploring the whole scene and then executing the task having access to a global map. We further argue that by learning the layout priors through cross-modal attention, we can leverage the spatial and semantic descriptions from natural language and decrease the uncertainty over the hallucinated areas. As opposed to recent work [31] that outputs a single waypoint, we learn to predict the whole trajectory, while our waypoints are determined by the alignment between language and egocentric maps rather than the distance to the goal.

In summary, our contributions are as follows:

- A novel system for the VLN task that learns maps as an explicit intermediate representation.
- Semantic grounding of language to those maps by applying cross-modal attention when learning to predict semantic maps.
- Spatial grounding of instructions when learning to predict paths by applying cross-modal attention on semantic maps and language.
- An analysis over the learned representation that demonstrates the effectiveness of using egocentric maps for the VLN task.
- Competitive results in the VLN-CE [32] dataset against current state-of-the-art methods.

2. Related Work

Vision-and-Language Navigation. The problem of instruction following for navigation has drawn significant attention in a wide range of domains. These include Google Street View Panoramas [11], simulated environments for quadcopters [5], multilingual settings [33], interactive vision-dialogue setups [59], real world scenes [3], and realistic simulations of indoor scenes [4]. More relevant to our work is the literature on the Vision-and-Language Navigation (VLN) task initially defined in [4] on navigation graphs (R2R) in Matterport3D [8] dataset, and then converted for continuous environments in [32] (VLN-CE). Arguably, the biggest challenges in VLN are grounding the natural language to the visual input while keeping track which part of the instruction was completed. To address these issues, many methods rely on unstructured memory such as LSTM for visual-textual alignment [14, 17, 28, 37], or have dedicated progress monitor modules [37, 38]. Other approaches formulate instruction following as a Bayesian tracking problem [2], or learn to decompose and execute the instructions in short steps [58]. Another line of works [12, 21, 23, 26, 31, 39, 42, 43, 45] make use of attention mechanisms and adapt powerful language models such as BERT [15] and transformer networks [52] to the VLN task. For instance, Chen at al. [12] learn the association between instructions and nodes on a prebuild topological map of the environment, while Krantz et al. [31] learn to predict waypoints from panoramic images and investigate the prediction in different action spaces. In contrast to all these works, our method learns to associate the language and egocentric observations at the semantic level with 2D spatial representations followed by path prediction.

Cross-modal attention. The transformer architecture [52] has been extremely successful in language [15], speech [16] vision [30] and multimodal applications [27]. A key feature of the transformer architecture is the *attention* mechanism. Cross-modal transformers have been widely used for vision-language tasks such as visual-question answering and beyond, such as joint video and language understanding [49]. Additionally, there have been investigations into whether the multimodal transformers learn interpretable relations between the two modalities by analyzing the cross-modal attention heads, as studied in Visual BERT [34] and in a cross modal self attention network for referring image segmentation [55]. Prior works have trained cross-modal transformers in two ways: 1) *single-stream design* where the multimodal inputs (for example, word embeddings and image regions) are fed into a single transformer architecture. Examples of this are UNITER [13], VLBERT [48], VisualBERT [34]. 2) *multi-stream design* where the individual modalities are encoded separately via self-attention and then a cross-modal representation is learned by the transformer. Examples of this are LXMERT [50], ViLBERT [36], [57]. In this work, we adapt the multi-stream design for vision language navigation using egocentric maps. We also investigate the cross-modal attention heads and decoder

representation of the transformer for interpretable patterns.

Map Prediction in Navigation. Modular approaches using different types of spatial representations have been successful in multiple navigation tasks, whether they focused on occupancy [10, 18, 22, 29, 44] or semantic map prediction [7, 9, 19, 20, 35, 40]. For example, Gupta et al. [22] learn a differentiable mapper for predicting top-down egocentric maps that are trained end-to-end with a differentiable planner, while Cartillier et al. [7] learn to build top-down allocentric maps from egocentric RGB-D observations. Several recent works go beyond traditional mapping and learn to predict information outside the field-of-view of the agent [19,35,40,44]. The work of [44] learns to hallucinate occupancy layouts in indoor environments, while [19] extends the prediction to semantic classes and uses information gain objectives to increase the performance of the predictor. Our approach expands upon this last set of methods by presenting a language-informed model that attempts to hallucinate missing information using cues from both language and currently observed regions.

3. Approach

3.1. Problem setup

We address instruction-following navigation in indoor environments, where natural language instructions implicitly describe a specific path and goal location in the environment that an agent needs to follow. In particular, we consider the setup described in the Vision-and-Language Navigation in Continuous Environments (VLN-CE) [32] that was adapted from the Room-to-Room (R2R) [4] dataset from pre-specified navigation graphs to continuous 3D environments. VLN-CE uses the Habitat [47] simulator in the Matterport3D [8] scenes and offers more realistic settings and is much more challenging [32] than the original R2R. During a VLN-CE navigation episode, the agent has access to egocentric RGB-D observations at a resolution of 256×256 with a horizontal field-of-view of $90°$. In contrast to other recent methods [12, 31], we assume the agent observes a frame with limited field-of-view at each timestep (not panoramas). The action space is defined over a discrete set of actions consisting of `MOVE_FORWARD` by $0.25m$, `TURN_LEFT` and `TURN_RIGHT` by $15°$, and `STOP`, without actuation noise. Recently, the work of [31] demonstrated higher performance when continuous-space actions are considered, however we kept the action set discrete to remain consistent with prior work on VLN-CE.

3.2. Overview of our approach

We propose a method for Vision-and-Language Navigation involving path prediction over predicted semantic egocentric 2D maps. Our argument for this approach is threefold. First, an egocentric map offers a natural representation for grounding spatial and semantic concepts from natural language instructions. Second, a VLN method should take advantage of the knowledge over semantic and spatial layouts as they offer a strong prior over possible trajectories. Third, the language instruction provides a semantic description of a trajectory through the environment, which could be leveraged to improve map predictions.

Given the instruction, our method learns to predict the entire path defined as a set of waypoints on an egocentric local map at every step of the episode (Sec. 3.3). The agent then localizes itself on the current predicted path and chooses the following waypoint on the path as a short-term goal. This goal is then passed to an off-the-shelf local policy (DD-PPO [53]) which predicts the next navigation action. We assume that we have access to ground-truth pose as provided by the simulator to facilitate DD-PPO. We note that estimating the pose from noisy sensor readings is out of the scope of this work, and point to visual odometry methods [56] that can adapt DD-PPO agents to such a setting.

To obtain the egocentric map we define a language-informed two-stage semantic map predictor that learns to hallucinate the semantics in the unobserved areas (Sec. 3.4). An overview of our method is shown in Figure 2. In the following two paragraphs we briefly describe the common input encoding procedures between different components of our method.

Instruction Encoding. We use a pretrained Bidirectional Encoder Representations from Transformers (BERT) [15] model, which is a multi-layer transformer [52], to extract a feature vector for each word in the instruction. The overall feature representation for the instruction $X' \in \mathbb{R}^{M \times d'}$ is passed through a fully-connected layer to obtain the final representation $X \in \mathbb{R}^{M \times d}$, where M is the number of words in the instruction, $d' = 768$ is the default feature dimension of BERT, and $d = 128$ is the feature dimension we use throughout our method. During training we only finetune the last layer of BERT.

Egocentric Map Encoding. Our network encodes an input egocentric semantic map $s \in \mathbb{R}^{h' \times w' \times c}$ with a truncated ResNet18 [24], where h', w', c are height, width, and the number of semantic classes, respectively as $Y = Enc(s)$. The ResNet18 initially produces a feature representation $Y' \in \mathbb{R}^{h \times w \times d}$ ($h = \frac{h'}{16}$, $w = \frac{w'}{16}$), which is then reshaped to $Y \in \mathbb{R}^{N \times d}$ ($N = h \times w$). One of these modules encodes the ground projected RGB-D observations for the map predictor (Sec. 3.4) and a separate module is used to encode the predicted semantic map for the path predictor (Sec. 3.3).

3.3. Cross-modal attention for path prediction

The cross-modal attention for path prediction module takes as input the instruction representation X and the egocentric map encoding Y and formulates the path prediction problem as a waypoint localization task. In order to learn

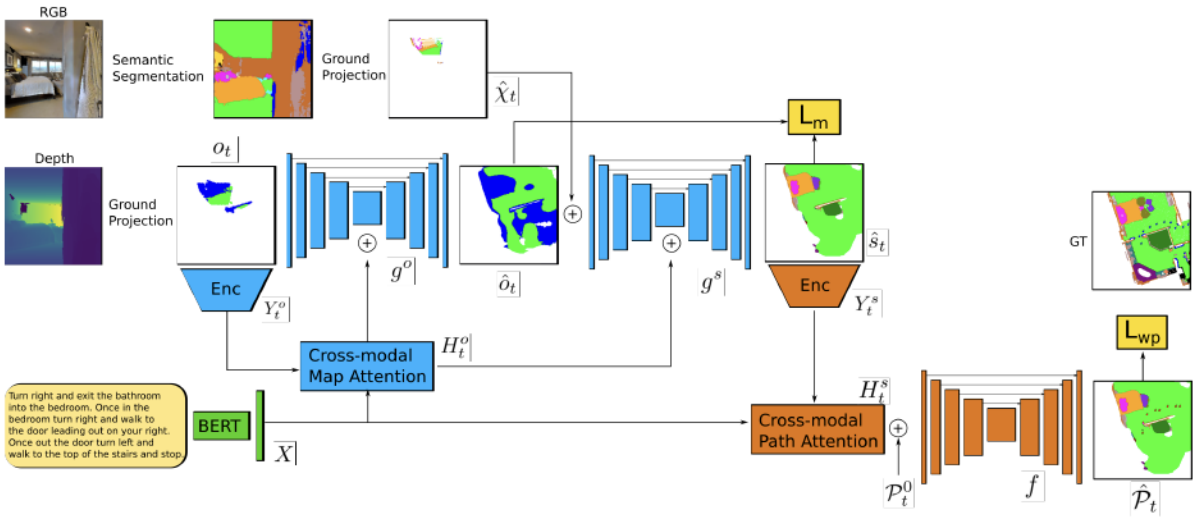

Figure 2. We propose an approach to predict egocentric semantic maps and paths described by natural language instructions. At the core of our method are two cross-modal attention modules that learn language-informed representations to facilitate both the hallucination of semantics over unobserved areas as well as the prediction of a set of waypoints that the agent needs to follow to reach the goal. The components colored in blue refer to the map prediction part (Sec. 3.4) of our model, the ones in orange correspond to the path prediction (Sec. 3.3), and the yellow boxes are the losses.

a grounded representation of the natural language instruction on the egocentric map, we define a cross-modal attention module following the architecture of the self-attention transformer model [52]. While it is common to concatenate the representations of the two modalities and then use self-attention (such as VisualBERT [34]) we follow the example of LXMERT [50] and treat the egocentric map separately as the query and the instruction as the key and value. The idea is that during an episode the language instruction remains the same while the egocentric map changes at each time-step and is used to query the model for the path.

Specifically, given the egocentric map feature representation $Y_t^s = Enc(s_t)$ at time t during a VLN episode, and the instruction features X, we use the scaled dot-product attention:

$$Q = Y_t^s W_q, K = XW_k, V = XW_v \tag{1}$$

$$H_t^s = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2}$$

where W_q, W_k, and $W_v \in \mathbb{R}^{d\times d}$ are learned parameter matrices, and $H_t^s \in \mathbb{R}^{N\times d}$ is the attended representation over the egocentric semantic map regions. In practice, this architecture [52] first applies self-attention to each modality followed by the cross-modal attention.

We define the path as a set of 2D waypoints $\{p_t^i\}_{i=1}^k$ situated on the egocentric map. The first and last waypoints always represent the starting position and the final goal position respectively. During training we sample the remaining waypoints from the ground-truth path with respect to the instruction. These are used to construct ground-truth heatmaps $\mathcal{P}_t \in \mathbb{R}^{k\times u\times v}$ based on a 2D Gaussian centered at each waypoint with $\sigma = 1$, where u, v are the height and width of the heatmap. We predict the entire path at every time-step given the entire instruction. This can cause ambiguity with regards to the waypoint placements towards the agent's current pose, since the agent has no knowledge of the amount of the path covered at a given time-step. In other words, if the agent is half-way through the path then the model should learn to predict both backward and forward waypoints along the path, as opposed to predicting only forward waypoints at the beginning of the episode. We mitigate this issue in two ways. First, the path prediction is conditioned on the starting position heatmap $\mathcal{P}_t^0$ relative to the current agent's pose. Second, we add an auxiliary loss that trains the model to predict a probability $\hat{\xi}_t^i$ for each waypoint whether it has already been traversed. We empirically found this auxiliary loss to help the learning process.

The waypoint predictor model is defined as an encoder-decoder UNet [46] f, that takes as inputs the instruction attended representation of the egocentric map regions H_t^s and the starting position $\mathcal{P}_t^0$:

$$\hat{\mathcal{P}}_t, \hat{\xi}_t = f\left(H_t^s, \mathcal{P}_t^0\right). \tag{3}$$

We train the waypoint prediction with the following loss:

$$L_{wp} = \sum_{i=1}^{k} b_t^i ||\hat{\mathcal{P}}_t^i - \mathcal{P}_t^i||_2^2 - \lambda_\xi \xi_t^i \log \hat{\xi}_t^i \tag{4}$$

where b_t^i is a binary indicator whether the particular waypoint i is visible on the egocentric map at time t, and λ_ξ

weighs the auxiliary loss.

3.4. Cross-modal attention for map prediction

We design a language-informed semantic map predictor for obtaining the egocentric semantic map s_t from RGB-D observations. Given the often limited field-of-view of embodied agents, we are interested in hallucinating the semantic information in regions where the agent cannot directly observe. While different versions of this procedure were attempted in the past [18, 19, 44], our key contribution is to learn the layout priors by leveraging the spatial and semantic descriptions from the instructions.

The map prediction is defined as a semantic segmentation task over the top-down egocentric map. Our model first takes as input the depth observation which is ground-projected to an egocentric grid $o_t \in \mathbb{R}^{h' \times w' \times 3}$ containing the classes *occupied*, *free*, and *void*. For the ground-projection we first unproject the depth to a 3D point cloud using the camera intrinsic parameters and then map each 3D point to an $h' \times w'$ grid following the procedure described here [25]. Note that o_t is an incomplete representation of the occupancy map around the agent, where all areas outside the field-of-view are considered unknown.

We define a cross-modal attention module similar to the one in Sec. 3.3, where the feature representation $Y_t^o = Enc(o_t)$ is determined as the query, while the instruction features X are used as key and value. Following Eq. 1 and 2 (where Y_t^s is replaced by Y_t^o) we get the attended representation H_t^o over the incomplete egocentric map o_t. The prediction model includes two encoder-decoder UNet [46] models g^o, g^s stacked together:

$$\hat{o}_t = g^o\left(o_t, H_t^o\right) \qquad \hat{s}_t = g^s\left(\hat{o}_t, H_t^o, \hat{\chi}_t\right) \tag{5}$$

where $\hat{\chi}_t \in \mathbb{R}^{h' \times w' \times c}$ is a ground-projected semantic segmentation of the RGB frame. Note that H_t^o is concatenated at the bottlenecks of both g^o, g^s models. The model is trained with a pixel-wise cross-entropy loss on the occupancy and the semantic classes:

$$L_m = -\sum_{q \in (s,o)} \sum_k \sum_c q_{k,c} \log \hat{q}_{k,c} \tag{6}$$

where k iterates over the number of pixels in the map and $q_{k,c}$ is the ground-truth label for pixel k. The ground-truth semantic maps are created from the available 3D semantic information in Matterport3D. The network that produces $\hat{\chi}$ is another UNet which is pre-trained separately from the rest of the model.

Overall learning objective. During training we add up all the losses from the path and map prediction modules:

$$L = \lambda_{wp} L_{wp} + \lambda_m L_m \tag{7}$$

where the λs denote the corresponding loss weights, and perform a single backward pass through the entire model.

3.5. Controller

The method described so far outputs the path as a set of 2D waypoints $\{p_t^i\}_{i=1}^k$ on an egocentric map from an RGB-D observation. In order to follow this path towards the goal, at each time-step we designate a waypoint as a short-term goal, following:

$$\zeta = 1 + \arg\min_i \Delta(\hat{p}_t^i, \varrho_t) \tag{8}$$

where Δ is the euclidean distance, $\hat{p}_t^i$ corresponds to the mode of the predicted waypoint heatmap $\hat{\mathcal{P}}_t^i$, and ϱ_t is the agent's pose at time t. This effectively determines the closest predicted waypoint to the agent and selects the next one in the sequence as the short-term goal p_t^ζ. In order to reach the short-term goal, we use the off-the-shelf deep reinforcement learning model DD-PPO [53] that is trained for the PointNav [1] task. DD-PPO receives the current depth observation and p_t^ζ and outputs the next navigation action for the agent. Finally, at any time during the episode the agent may decide on the `STOP` action when it's within a certain radius τ (m) of the final goal (last predicted waypoint) and the confidence of the goal in the predicted heatmap is above a threshold γ.

4. Experiments

We conduct our experiments in the VLN-CE [32] dataset, which offers 16,844 path-instruction pairs over 90 visually realistic scenes in the Matterport3D [8] dataset. We follow the typical evaluation scenario and report results in scenes which were observed (*val-seen*) and not observed (*val-unseen*) during training. An episode is considered successful if the `STOP` decision is taken within $3m$ of the goal position, and the agent has a fixed-time budget of 500 steps to complete an episode. As mentioned before, the agent has access to egocentric RGB-D observations with a horizontal field-of-view of 90°. We perform three sets of experiments. First, we compare against other methods on the VLN-CE dataset including the held-out test set of the VLN-CE challenge (Sec. 4.1), followed by an ablation study (Sec. 4.2). Finally, we provide visual examples of the learned representation (Sec. 4.3). We use two main variations of our method. ***CM²*** refers to our full pipeline that predicts both the egocentric map and path from RGB-D inputs, while ***CM²-GT*** refers to using the ground-truth egocentric map as input, effectively only performing path prediction. All egocentric maps used are local 192×192 with each pixel corresponding to $5cm \times 5cm$. The map covers a square 9.6 meters on a side, leaving most of the scene unobserved. We provide code, trained models and instructions to reproduce our results: https://github.com/ggeorgak11/CM2. Implementation details along with additional experimental results are included in the supplementary material.

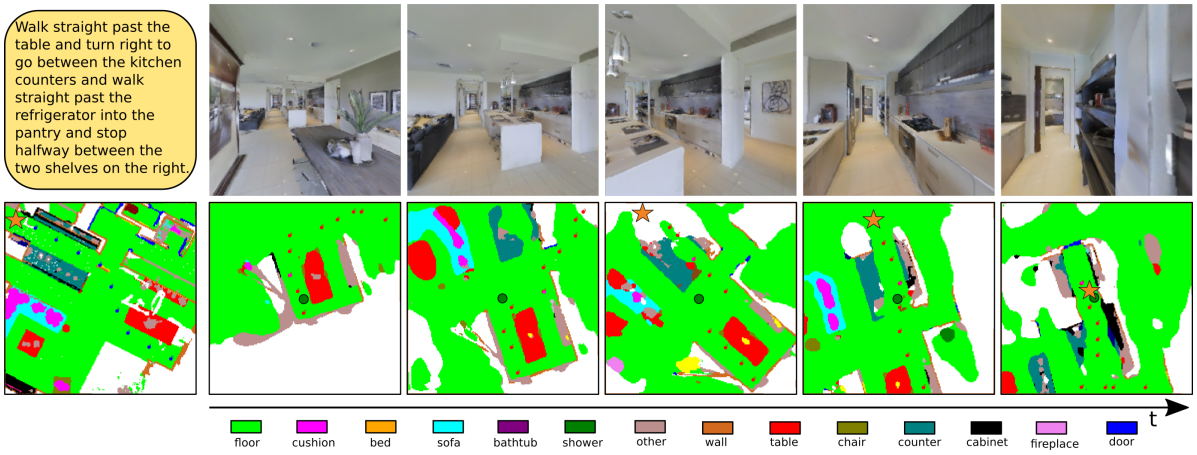

Figure 3. Navigation example using our method *CM²* on a scene from *val-unseen*. The top row shows the RGB observations of the agent, while bottom shows the path prediction on the egocentric maps (the agent is in the middle looking upwards shown as the green circle). The red waypoints represent our path prediction at the particular time-step. Observe that the goal, shown as an orange star, is neither visible nor within the egocentric map at the beginning of the episode. The ground-truth map and path are depicted in the bottom left corner.

	Val-Seen					Val-Unseen				
	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑
Seq2Seq+PM+DA+Aug [32]	9.37	7.02	46.0	33.0	31.0	9.32	7.77	37.0	25.0	22.0
AG-CMTP* [12]	-	6.60	**56.2**	35.9	30.5	-	7.9	39.2	23.1	19.1
R2R-CMTP* [12]	-	7.10	45.4	36.1	31.2	-	7.9	38.0	26.4	22.7
CMA+PM+DA+Aug [32]	9.26	7.12	46.0	37.0	35.0	8.64	7.37	40.0	32.0	30.0
WPN-DD* [31]	**9.11**	6.57	44.0	35.0	32.0	**8.23**	7.48	35.0	28.0	26.0
LAW [45]	9.34	6.35	49.0	40.0	**37.0**	8.89	**6.83**	**44.0**	**35.0**	**31.0**
CM^2 (Ours)	12.05	**6.10**	50.7	**42.9**	34.8	11.54	7.02	41.5	34.3	27.6
WPN-CC* [31]	10.29	6.05	51.0	40.0	35.0	10.62	6.62	43.0	36.0	30.0
HPN-C* [31]	8.71	5.17	53.0	47.0	45.0	7.71	6.02	42.0	38.0	36.0
CM^2-GT (Ours)	12.60	4.81	58.3	52.8	41.8	10.68	6.23	41.3	37.0	30.6

Table 1. Evaluation on VLN-CE dataset. All methods marked with * use panoramic images. CM^2-GT is the same as CM^2, but uses ground truth local maps, rather than predicting them. HPN-C and WPN-CC use a more expressive action space than the rest of the methods. AG-CMTP and R2R-CMPT allow the agent to explore each scene before the experiment begins. Our method is the most successful on *val-seen* while it is competitive on *val-unseen*.

4.1. VLN-CE Evaluation

Here we evaluate the performance of our method on the continuous vision-and-language navigation task against current state-of-the-art methods. The metrics reported are the following: Trajectory length **TL** (m), navigation error from goal **NE** (m), oracle success rate **OS** (%), success rate **SR** (%), and success weighted by path length **SPL** (%). More details on these metrics can be found in [1, 4].

We compare our method against the following works:

Krantz et al. [32]: Two baselines are used from here. First, *Seq2Seq+PM+DA+Aug* is a simple sequence-to-sequence baseline that uses a recurrent policy to predict the action directly from the visual observations. Second, *CMA+PM+DA+Aug* utilizes cross-modal attention between instruction and RGB-D observations. Both methods use off-the-shelf techniques for Progress Monitor (PM), DAgger (DA), and synthetic data augmentation.

Chen et al. [12]: This work uses cross-modal attention between the instruction and a topological map to compute a global navigation plan. To construct the topological map, the authors assume the agent can explore the environment before the execution of the navigation episode. Each node in the topological map corresponds to a panoramic image. We compare against *AG-CMTP* and *R2R-CMTP* which use the method's generated map and the maps from the Room2Room [1] dataset respectively.

Raychaudhuri et al. [45]: This method (*LAW*) updates the training setup of *CMA+PM+DA+Aug* [32] by adjusting the supervision to use the nearest waypoint on the path rather than the goal location.

Krantz et al. [31]: We compare against the Waypoint Prediction Network (*WPN*) and the Heading Prediction Network (*HPN*) which are end-to-end models that predict rel-

ative waypoints directly from natural language instructions and panoramic RGB-D inputs. The models differ with respect to the waypoint prediction space. *WPN-CC* considers continuous values for distance and direction, *WPN-DD* considers discrete values, and *HPN-C* uses a constant value for distance and continuous for direction. Our method is analogous to *WPN-DD*, since our waypoint prediction is on the discrete 2D space of maps. Investigating more expressive waypoint prediction spaces is out of the scope of our work.

Quantitative results are shown in Table 1 and a navigation example can be seen in Figure 3. On *val-seen* our method *CM²* outperforms all other baselines except *WPN-CC* and *HPN-C* (which use more expressive waypoint prediction spaces) on navigation error and success rate, while it is competitive on SPL. In particular, we show better results than *WPN-DD* which uses panoramic images (4× larger field-of-view), and was trained on 200M steps of experience (285× more data) [31]. This is a characteristic of end-to-end methods that need to learn all navigation components such as mapping, planning, and control in a single network and thus require large amounts of data. In contrast, aligning language to egocentric maps proves to be much more sample efficient, as our model was trained with only 0.7M training samples. Regarding our comparison to [12], *AG-CMPT* performs better only on oracle success rate, while our *CM²* method has a noticeably higher success rate. However, this baseline has a prior scene exploration phase, which is not counted in the task step limit, that acquires knowledge of scene topology to use during the navigation episode. In comparison, our *CM²-GT*, which also has knowledge over the map, performs better on all metrics. We are also competitive against *CMA+PM+DA+Aug* and *LAW* that use cross-modal attention mechanisms between the instruction and the RGB-D frames. The latter also employs a more sophisticated reward function that forces the agent to stay on the path and trains on an augmented dataset with over ten times as many trajectories. We outperform both in success rate on *val-seen* and we have almost the same performance with *LAW* on *val-unseen*. Finally, when the input to our method is the ground-truth egocentric semantic map (*CM²-GT*) we observe a significant increase in success rate in *val-seen*. Although the map is local and the goal location is usually not visible, this performance gain further justifies our choice of using cross-modal attention on egocentric maps.

VLN-CE Leaderboard We submitted our CM² on the held-out test-unseen set containing 3.4K episodes in unseen environments used for the VLN-CE challenge. Table 2 shows the leaderboard as accessed on Mar 8th 2022. Our method is leading in terms of OS, SR, and NE among those that use standard observation (no panoramas) and action spaces (discrete), and is 4th overall on OS SR, and NE.

Team Name	TL	NE	OS	SR	SPL
CWP-VLNBERT*	13.3	5.9	51	42	36
CWP-CMA*	11.9	6.3	49	38	33
WaypointTeam*	8.0	6.6	37	32	30
CM²	13.9	7.7	39	31	24
TJA*	10.4	8.1	42	29	27
VIRL_Team	8.9	7.9	36	28	25

Table 2. Results on the VLN-CE challenge leaderboard. Methods marked with * use either panoramic images and/or a non-standard action space.

	IoU (%)	F1 (%)	PCW (%)
CM²-w/o-MapAttn	21.2	33.2	71.1
CM²	**28.3**	**42.2**	**76.5**

Table 3. Effect of map attention on map and waypoint prediction.

Val-Seen	TL	NE	OS	SR	SPL
CM²-GT, $\tau = 1.5$	**10.18**	5.01	53.6	49.5	45.1
CM²-GT, $\tau = 1.0$	11.48	4.94	56.4	51.9	43.8
CM²-GT, $\tau = 0.5$	12.60	4.81	58.3	52.8	41.8
CM²-GT-384, $\tau = 0.5$	12.89	**4.52**	**66.4**	**58.4**	**46.7**

Table 4. Effect of map size and stop distance threshold on VLN.

4.2. Ablation Study

In this experiment we provide an analysis over our model and aim to answer the following questions:

How important is the cross-modal map attention? The cross-modal map attention, shown in Figure 2, is the attention module that learns the semantic grounding and influences the semantic map prediction. We are interested in quantifying its contribution towards the map and path prediction and define the baseline *CM²-w/o-MapAttn* that does not include the cross-modal map attention module and therefore is not aware of the language instruction. We compare against our method *CM²* on the popular semantic segmentation metrics of Intersection over Union (IoU) and F1 score, and on the Percentage of Correct Waypoints (PCW) that evaluates the quality of the path prediction. PCW counts a predicted waypoint as correct if it is within $1.92m$ (on the 192×192 maps) of the ground-truth waypoint. Results are reported in Table 3. *CM²* has higher performance on IoU, F1, and PCW by 7.1%, 9.0%, and 5.4% respectively. These results show that the cross-modal map attention extracts useful information from language that improves the prediction of the semantic map and the path. Examples over map predictions are shown in Figure 4.

What is the effect of the stop decision threshold? We vary the stop decision distance threshold τ (m) used by the controller and observe the performance on the VLN-CE metrics in Table 4. This experiment is carried out on *val-seen* using *CM²-GT*. When $\tau = 1.5$, success rate drops by 3.3% because the agent chooses to stop more aggressively thus it is more likely to choose `STOP` outside the goal radius. On the

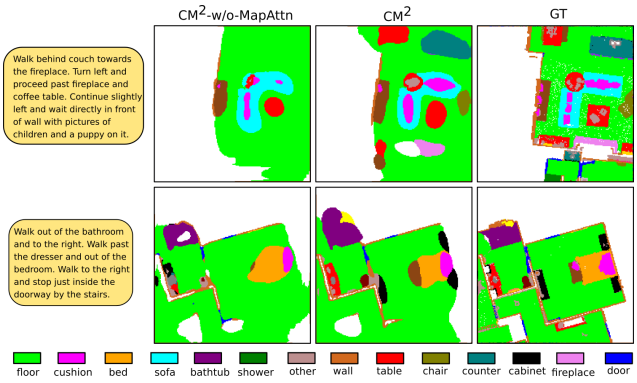

Figure 4. Semantic map predictions with and without cross-modal map attention.

other hand, SPL gained 3.3% since stopping earlier reduces the path length. This result signifies a trade-off between success rate (SR) and SPL based on the value of τ that can adjust the agent's behavior.

What is the effect of egocentric map size? All experimental evaluation of our work (*CM²*, *CM²-GT*) uses 192×192 egocentric maps. Given that each cell in the map corresponds to $5cm \times 5cm$, this translates to a distance from the center of the map (where the agent is situated) to each side of 4.8m. With the mean euclidean distance between the start position and goal being around 8m across *val-seen* and *val-unseen* episodes, this means that for the majority of the episodes the goal is not located within the egocentric map at the beginning. In order to see how much this affects performance, we train our path predictor again *CM²-GT-384* with maps of size 384×384 (9.6m between the agent and the sides of the map) and compare to our original method in Table 4. Doubling the map size increases SR by 5.6%, OS by 8.1%, and SPL by 4.9% demonstrating that the larger maps have significant impact on the navigation performance.

4.3. Validation of Semantic and Spatial Grounding

Finally, we provide evidence that the learned representations can be semantically and spatially grounded on the egocentric maps. Specifically, we visualize (Figure 5) two feature representations from the cross-modal path attention module: 1) The attention decoder output $H_t^s \in \mathbb{R}^{N \times d}$ which we max-pool over the feature dimension d and reshape N back to its encoded map dimensions of $h \times w$ to get a spatial heatmap. This representation is shown to focus around goal locations and along paths. 2) The cross-modal attention ($Softmax\left(\frac{QK^T}{\sqrt{d}}\right)$) between the map regions and the words in the instruction with dimensionality $N \times M$ from which we can visualize the attention heatmap for a specific word token over the map. This demonstrates that the cross-modal attention learns to associate instruction tokens to semantic objects on the map.

Figure 5. Top: Visualization of attention decoder output H_t^s which tends to focus on areas corresponding to goal locations. The agent's location is denoted with a green circle and the goal with an orange star. Bottom: Cross-modal attention between map and specific word tokens.

5. Conclusion

We presented a new method for the Language-and-Navigation task that solves the problem by first predicting the egocentric semantic map and then estimating the trajectory, defined by the instruction, on the 2D map. This is facilitated by two cross-modal attention modules that learn to semantically and spatially ground the natural language on the egocentric map. We showcased the effectiveness of our method with competitive results on the VLN-CE dataset and demonstrated that grounding the language on the maps allows for good VLN performance with a fraction of the data that the end-to-end methods require. Furthermore, we qualitatively show that our method learns meaningful intermediate representations.

Acknowledgements. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0080, as well as by the ARL DCIST CRA W911NF-17-2-0181, NSF TRIPODS 1934960, and NSF CPS 2038873 grants.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 5, 6

[2] Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. Chasing ghosts: instruction following as bayesian state tracking. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 371–381, 2019. 2

[3] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pages 671–681. PMLR, 2021. 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1, 2, 3, 6

[5] Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Conference on Robot Learning*, pages 505–518. PMLR, 2018. 2

[6] Annina Brügger, Kai-Florian Richter, and Sara Irina Fabrikant. How does navigation system behavior influence human behavior? *Cognitive research: principles and implications*, 4(1):1–22, 2019. 1

[7] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 964–972, 2021. 3

[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. *International Conference on 3D Vision (3DV)*, 2017. 2, 3, 5

[9] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems 33*, 2020. 3

[10] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *International Conference on Learning Representations*, 2020. 3

[11] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019. 2

[12] Kevin Chen, Junshen K. Chen, Jo Chuang, Marynel Vazquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11276–11286, June 2021. 2, 3, 6, 7

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[14] Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. Evolving graphical planner: Contextual global planning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 33:20660–20672, 2020. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3

[16] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018. 2

[17] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3318–3329, 2018. 1, 2

[18] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. *International Conference in Robotics and Automation (ICRA)*, 2022. 3, 5

[19] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. *International Conference on Learning Representations (ICLR)*, 2022. 3, 5

[20] Georgios Georgakis, Yimeng Li, and Jana Kosecka. Simultaneous mapping and target driven navigation. *arXiv preprint arXiv:1911.07980*, 2019. 3

[21] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643, 2021. 2

[22] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2017. 1, 3

[23] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13137–13146, 2020. 1, 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[25] Joao F Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018. 5

[26] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 2

[27] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 2

[28] Haoshuo Huang, Vihan Jain, Harsh Mehta, Jason Baldridge, and Eugene Ie. Multi-modal discriminative model for vision-and-language navigation. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 40–49, 2019. 2

[29] Kapil Katyal, Katie Popek, Chris Paxton, Phil Burlina, and Gregory D. Hager. Uncertainty-aware occupancy map prediction using generative networks for robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5453–5459, 2019. 3

[30] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021. 2

[31] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15162–15171, 2021. 1, 2, 3, 6, 7

[32] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 5, 6

[33] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2020. 1, 2

[34] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, 2020. 2, 4

[35] Yiqing Liang, Boyuan Chen, and Shuran Song. SSCNav: Confidence-aware semantic scene completion for visual semantic navigation. *International Conference on Robotics and Automation (ICRA)*, 2021. 3

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23, 2019. 2

[37] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *International Conference on Learning Representations (ICLR)*, 2019. 2

[38] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6732–6740, 2019. 2

[39] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pages 259–274. Springer, 2020. 2

[40] Medhini Narasimhan, Erik Wijmans, Xinlei Chen, Trevor Darrell, Dhruv Batra, Devi Parikh, and Amanpreet Singh. Seeing the un-scene: Learning amodal semantic maps for room navigation. *European Conference on Computer Vision. Springer, Cham*, 2020. 3

[41] J. O'Keefe and L. Nadel. *The hippocampus as a cognitive map*. Oxford University Press, 1998. 1

[42] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15942–15952, 2021. 2

[43] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1655–1664, 2021. 2

[44] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. *European Conference on Computer Vision*, pages 400–418, 2020. 3, 5

[45] Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4018–4028, 2021. 2, 6

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 5

[47] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9339–9347, 2019. 3

[48] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-

linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[49] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[50] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 2, 4

[51] E.C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55:189–208, 1948. 1

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3, 4

[53] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *International Conference on Learning Representations (ICLR)*, 2019. 3, 5

[54] Anna Wunderlich and Klaus Gramann. Landmark-based navigation instructions improve incidental spatial knowledge acquisition in real-world environments. *biorxiv:10.1101/789529*, 2020. 1

[55] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511, 2019. 2

[56] Xiaoming Zhao, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. The surprising effectiveness of visual odometry techniques for embodied pointgoal navigation. *International Conference on Computer Vision (ICCV)*, 2021. 3

[57] Chen Zheng, Quan Guo, and Parisa Kordjamshidi. Cross-modality relevance for reasoning on language and vision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7642–7651, 2020. 2

[58] Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. Babywalk: Going farther in vision-and-language navigation by taking baby steps. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2539–2556, 2020. 2

[59] Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. Self-motivated communication agent for real-world vision-dialog navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1594–1603, 2021. 2