Where Are You? Localization from Embodied Dialog

Meera Hahn[1] **Jacob Krantz**[2] **Dhruv Batra**[1,3] **Devi Parikh**[1,3]
James M. Rehg[1] **Stefan Lee**[2] **Peter Anderson**[1*]

[1]Georgia Institute of Technology [2]Oregon State University [3]Facebook AI Research (FAIR)

{mhahn30, dbatra, parikh, rehg}@gatech.edu
{krantzja, leestef}@oregonstate.edu
pjand@google.com

Abstract

We present WHERE ARE YOU? (WAY), a dataset of ∼6k dialogs in which two humans – an Observer and a Locator – complete a cooperative localization task. The Observer is spawned at random in a 3D environment and can navigate from first-person views while answering questions from the Locator. The Locator must localize the Observer in a detailed top-down map by asking questions and giving instructions. Based on this dataset, we define three challenging tasks: Localization from Embodied Dialog or LED (localizing the Observer from dialog history), Embodied Visual Dialog (modeling the Observer), and Cooperative Localization (modeling both agents). In this paper, we focus on the LED task – providing a strong baseline model with detailed ablations characterizing both dataset biases and the importance of various modeling choices. Our best model achieves 32.7% success at identifying the Observer's location within 3m in unseen buildings, vs. 70.4% for human Locators.

1 Introduction

Imagine getting lost in a new building while trying to visit a friend who lives or works there. Unsure of exactly where you are, you call your friend and start describing your surroundings (*'I'm standing near a big blue couch in what looks like a lounge. There are a set of wooden double doors opposite the entrance.'*) and navigating in response to their questions (*'If you go through those doors, are you in a hallway with a workout room to the right?'*). After a few rounds of dialog, your friend who is familiar with the building will hopefully know your location. Success at this cooperative task requires goal-driven questioning based on your friend's understanding of the environment, unambiguous answers communicating observations via language, and active perception and navigation to investigate the environment and seek out discriminative observations.

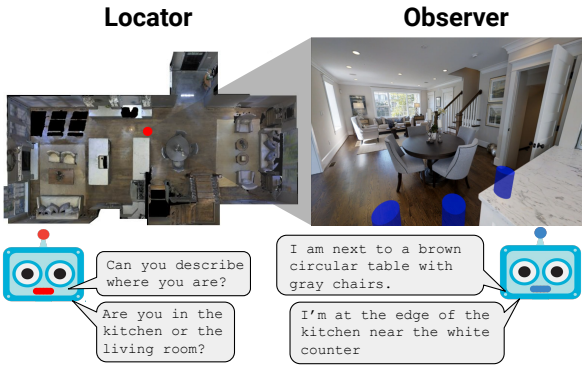


Figure 1: LED Task: The Locator has a top-down map of the building and is trying to localize the Observer by asking questions and giving instructions. The Observer has a first person view and may navigate while responding to the Locator. The turn-taking dialog ends when the Locator predicts the Observer's position.

In this work we present WHERE ARE YOU? (WAY), a new dataset based on this scenario. As shown in Fig. 1, during data collection we pair two annotators: an Observer who is spawned at random in a novel environment, and a Locator who must precisely localize the Observer in a provided top-down map. The map can be seen as a proxy for familiarity with the environment – it is highly detailed, often including multiple floors, but does not show the Observer's current or initial location. In contrast to the "remote" Locator, the Observer navigates within the environment from a first-person view but without access to the map. To resolve this information asymmetry and complete the task, the Observer and the Locator communicate in a live two-person chat. The task concludes when the Locator makes a prediction about the current location of the Observer. For the environments we use the Matterport3D dataset (Chang et al., 2017) of

*Now at Google.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 806–822,
November 16–20, 2020.

90 reconstructed indoor environments. In total, we collect ∼6K English dialogs of humans completing this task from over 2K unique starting locations.

The combination of localization, navigation, and dialog in WAY provides for a variety of modeling possibilities. We identify three compelling tasks encapsulating significant research challenges:

– Localization from Embodied Dialog. LED, which is the main focus of this paper, is the state estimation problem of localizing the Observer given a map and a partial or complete dialog between the Locator and the Observer. Although localization from dialog has not been widely studied, we note that indoor localization plays a critical role during calls to emergency services (Falcon and Schulzrinne, 2018). As 3D models and detailed maps of indoor spaces become increasingly available through indoor scanners (Chang et al., 2017), LED models could have the potential to help emergency responders localize emergency callers more quickly by identifying locations in a building that match the caller's description.

– Embodied Visual Dialog. EVD is the navigation and language generation task of fulfilling the Observer role. This involves using actions and language to respond to questions such as *'If you walk out of the bedroom is there a kitchen on your left?'* In future work we hope to encourage the transfer of existing image-based conversational agents (Das et al., 2017a) to more complex 3D environments additionally requiring navigation and active vision, in a step closer to physical robotics. The WAY dataset provides a testbed for this.

– Cooperative Localization. In the CL task, both the Observer and the Locator are modeled agents. Recent position papers (Baldridge et al., 2018; McClelland et al., 2019; Bisk et al., 2020) have called for a closer connection between language models and the physical world. However, most reinforcement learning for dialog systems is still text-based (Li et al., 2016) or restricted to static images (Das et al., 2017b; De Vries et al., 2017). Here, we provide a dataset to warm-start and evaluate goal-driven dialog in a realistic embodied setting.

Our main modeling contribution is a strong baseline model for the LED task based on LingUnet (Misra et al., 2018). In previously unseen test environments, our model successfully predicts the Locator's location within 3 meters 32.7% of the time, vs. 70.4% for the human Locators using the same map input, with random chance accuracy at 6.6%. We include detailed studies highlighting the importance of data augmentation and residual connections. Additionally, we characterize the biases of the dataset via unimodal (dialog-only, map-only) baselines and experiments with shuffled and ablated dialog inputs, finding limited potential for models to exploit unimodal priors.

Contributions: To summarize:

1. We present WAY, a dataset of ∼6k dialogs in which two humans with asymmetric information complete a cooperative localization task in reconstructed 3D buildings.
2. We define three challenging tasks: Localization from Embodied Dialog (LED), Embodied Visual Dialog, and Cooperative Localization.
3. Focusing on LED, we present a strong baseline model with detailed ablations characterizing both modeling choices and dataset biases.

2 Related Work

Image-based Dialog Several datasets grounding goal-oriented dialog in natural images have been proposed. The most similar settings to ours are Cooperative Visual Dialog (Das et al., 2017a,b), in which a question agent (Q-bot) attempts to guess which image from a provided set the answer agent (A-bot) is looking at, and GuessWhat?! (De Vries et al., 2017), in which the state estimation problem is to locate an unknown object in the image. Our dataset extends these settings to a situated 3D environment allowing for active perception and navigation on behalf of the A-bot (Observer), and offering a whole-building state space for the Q-bot (Locator) to reason about.

Embodied Language Tasks. A number of 'Embodied AI' tasks combining language, visual perception, and navigation in realistic 3D environments have recently gained prominence, including Interactive and Embodied Question Answering (Das et al., 2018; Gordon et al., 2018), Vision-and-Language Navigation or VLN (Anderson et al., 2018; Chen et al., 2019; Mehta et al., 2020; Qi et al., 2020), and challenges based on household tasks (Puig et al., 2018; Shridhar et al., 2020). While these tasks utilize only a single question or instruction input, several papers have extended the VLN task – in which an agent must follow natural language instructions to traverse a path in the environment – to dialog settings. Nguyen and Daumé III (2019) consider a scenario in which the agent can query an oracle for help while complet-

ing the navigation task. However, the closest work to ours is Cooperative Vision-and-Dialog Navigation (CVDN) (Thomason et al., 2019). CVDN is a dataset of dialogs in which a human assistant with access to visual observations from an oracle planner helps another human complete a navigation task. CVDN dialogs are set in the same Matterport3D buildings (Chang et al., 2017) and like ours they are goal-oriented and easily evaluated. The main difference is that we focus on localization rather than navigation. Qualitatively, this encourages more descriptive utterances from the first-person agent (rather than eliciting short questions). Our work is also related to Talk the Walk (de Vries et al., 2018) which presented a dataset for a similar task in an outdoor setting using a restricted, highly-abstracted map which encouraged language that is grounded in the semantics of building types rather than visual descriptions of the environment.

Table 1 compares the language in WAY against existing embodied perception datasets. Specifically, size, length and the density of different parts of speech (POS) are shown. Vocab size was determined by the total number of unique words. We used the (Loper and Bird, 2002) POS tagger to calculate the POS densities over the text in each dataset. We find that WAY has a higher density of adjectives, nouns, and prepositions than related datasets suggesting the dialog is more descriptive than the text in existing datasets.

Localization from Language. While localization from dialog has not been intensively studied, localization from language has been studied as a sub-component of instruction-following navigation agents (Blukis et al., 2018; Anderson et al., 2019; Blukis et al., 2019). The LingUnet model – a generic language-conditioned image-to-image network we use as the basis of our LED model in Section 4 – was first proposed in the context of predicting visual goals in images (Misra et al., 2018). This also illustrates the somewhat close connection between grounding language to a map and grounding referring expressions to an image (Kazemzadeh et al., 2014; Mao et al., 2016).

It is important to note that localization is often a precursor to navigation – one which has not been addressed in existing work in language-based navigation. In both VLN and CVDN, the instructions are conditioned on specific start locations – assuming the speaker knows the navigator's location prior to giving directions. The localization tasks of the WAY dataset fill this gap by introducing a dialog-based means to localize the navigator. This requires capabilities such as describing a scene, answering questions, and reasoning about how discriminative potential statements will be to the other agent.

3 WHERE ARE YOU? Dataset

We present the WHERE ARE YOU? (WAY) dataset consisting of 6,154 human embodied localization dialogs across 87 unique indoor environments.

Environments. We build WAY on Matterport3D (Chang et al., 2017), which contains 90 buildings captured in 10,800 panoramic images. Each building is also provided as a reconstructed 3D textured mesh. This dataset provides high-fidelity visual environments in diverse settings including offices, homes, and museums – offering numerous objects to reference in localization dialogs. We use the Matterport3D simulator (Anderson et al., 2018) to enable first-person navigation between panoramas.

Task. A WAY episode is defined by a starting location (i.e. a panorama p) in an environment e. The Observer is spawned at p_0 in e and the Locator is provided a top-down map of e (see Fig. 1). Starting with the Locator, the two engage in a turn-based dialog $(L_0, O_0, \ldots L_{T-1}, O_{T-1})$ where each can pass one message per turn. The Observer may move around in the environment during their turn, resulting in a trajectory $(p_0, p_1, \ldots, p_T)$ over the dialog. The Locator is not embodied and does not move but can look at the different floors of the house at multiple angles. The dialog continues until the Locator uses their turn to make a prediction $(\hat{p}_T)$ of the Observer's current location (p_T). The episode is successful if the prediction is within k meters of the true *final* position – i.e. $||p_T - \hat{p}_T||_2 < k$ m. This does not depend on the initial position, encouraging movement to easily-discriminable locations.

Map Representation. The Locator is shown top-down views of Matterport textured meshes as environment maps. In order to increase the visibility of walls in the map (which may be mentioned by the Observer), we render views using perspective rather than orthographic projections (see left in Fig. 1). We set the camera near and far clipping planes to render single floors such that multi-story buildings contain an image for each floor.

3.1 Collecting Human Localization Dialogs

To provide a human-performance baseline and gather training data for agents, we collect human

Table 1: Comparison of the language between the WAY dataset and related embodied perception datasets.

Method	Dataset Size	Vocab Size	Avg Text Length	Noun Density	Adj Density	Preposition Density	Dialog
CVDN	2050	2165	52	0.20	0.06	0.09	Yes
TtW	10K	7846	110	0.20	0.07	0.11	Yes
VLN	21K	3459	29	0.27	0.03	0.17	No
WAY	6154	5193	61	0.30	0.12	0.18	Yes

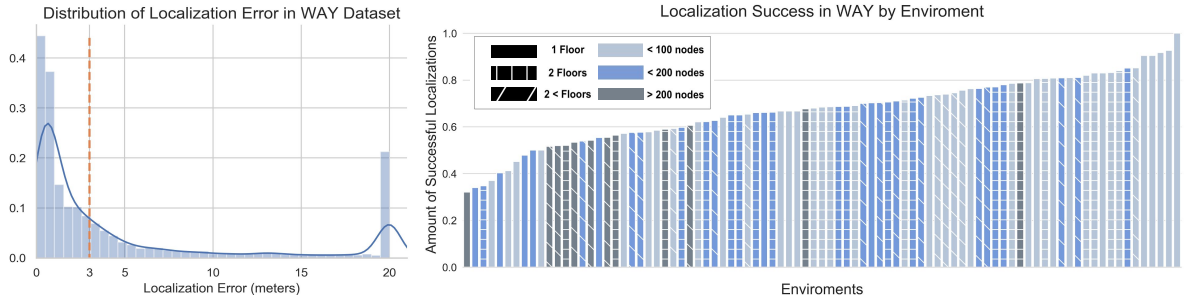


Figure 2: Left: Distribution of human localization error in WAY (20+ includes wrong floor predictions). Right: Human success rates (error <3m) by environment. Bar color indicates environment size (number of nodes) and pattern the number of floors.

localization dialogs in these environments.

Episodes. We generate 2020 episodes across 87 environments by rejection sampling to avoid spatial redundancy. For each environment, we iteratively sample start locations, rejecting ones that are within 5m of already-sampled positions. Three environments were excluded due to their size (too large or small) or poor reconstruction quality.

Data Collection. We collect dialogs on Amazon Mechanical Turk (AMT) – randomly pairing workers into Observer or Locator roles for each episode. The Observer interface includes a first-person view of the environment and workers can pan/tilt the camera in the current position or click to navigate to adjacent panoramas. The Locator interface shows the top-down map of the building, which can be zoomed and tilted to better display the walls. Views for each floor can be selected for multi-story environments. Both interfaces include a chat window where workers can send their message and end their dialog turn. The Locator interface also includes the option to make their prediction by clicking a spot on the top-down map – terminating the dialog. Note this option is only available after two rounds of dialog. Refer to the appendix for further details on the AMT interfaces.

Before starting, workers were given written instructions and a walk-through video on how to perform their role. We restricted access to US workers with at least a 98% success rate over 5,000 previous tasks. Further, we restrict workers from repeating tasks on the same building floor. In order to filter bad-actors, we monitored worker performance based on a running-average of localization error in meters and the number of times they disconnected from dialogs – removing workers who exceeded a 10m threshold and discarding their data.

Dataset Splits. We follow the standard splits for the Matterport3D dataset (Chang et al., 2017) – dividing along environments. We construct four splits: train, val-seen, val-unseen, and test comprising 3,967/299/561/1,165 dialogs from 58/55/11/18 environments respectively. Val-seen contains new start locations for environments seen in train. Both val-unseen and test contain new environments. This allows us to assess generalization to new dialogs and to new environments separately in validation. Following best practices, the final locations of the observer for the test set will not be released but we will provide an evaluation server where predicted localizations can be uploaded for scoring.

WAY includes dialogs in which the human Locator failed to accurately localize the Observer. In reviewing failed dialogs, we found human failures are often due to visual aliasing (e.g., across multiple floors), or are relatively close to the 3m threshold. We therefore expect that these dialogs still contain valid descriptions, especially when paired with the Observer's true location during training. In experiments when removing failed dialogs from

the train set, accuracy did not significantly change.

3.2 Dataset Analysis

Data Collection and Human Performance. In total, 174 unique workers participated in our tasks. On average each episode took 4 minutes and the average localization error is 3.17 meters. Overall, 72.5% of episodes where considered successful localizations at an error threshold of 3 meters. Each starting location has 3 annotations by separate randomly-paired Observer-Locator teams. In 40.9% of start locations, all 3 teams succeeded, in 36.3% 2, 18.5% 1, and 4.3% 0 teams succeeded. Fig. 2 left shows a histogram of localization errors.

Why is it Difficult? Localization through dialog is a challenging task, even for humans. The teams success depends on the uniqueness of starting position, if and where the Observer chooses to navigate, and how discriminative the Locator's questions are. Additionally, people vary greatly in their ability to interpret maps, particularly when performing mental rotations and shifting perspective (Kozhevnikov et al., 2006), which are both skills required to solve this task. We also observe that individual environments play a significant role in human error – as illustrated in Fig. 2 right, larger buildings and buildings with multiple floors tend to have larger localization errors, as do buildings with multiple similar looking rooms (e.g., multiple bedrooms with similar decorations or office spaces with multiple conference rooms). The buildings with the highest and lowest error are shown in Fig. 3.

Characterizing WAY Dialogs. Fig. 4 shows two example dialogs from WAY. These demonstrate a common trend – the Observer provides descriptions of their surroundings and then the Locator asks clarifying questions to refine the position. More difficult episodes require multiple rounds to narrow down the correct location and the Locator may ask the Observer to move or look for landmarks. On average, dialogs contain 5 messages and 61 words.

The Observer writes longer messages on average (19 words) compared to the Locator (9 words). This asymmetry follows from their respective roles. The Observer has first-person access to high-fidelity visual inputs and must describe their surroundings, *'In a kitchen with a long semicircular black countertop along one wall. There is a black kind of rectangular table and greenish tiled floor.'*. Meanwhile, the Locator sees a top-down view and uses messages to probe for discriminative details, *'Is it a round or rectangle table between the chairs?'*, or to prompt movement towards easier to discriminate spaces, *'Can you go to another main space?'*.

As the Locator has no information at the start of the episode, their first message is often a short prompt for the Observer to describe their surroundings, further lowering the average word count. Conversely, the Observer's reply is longer on average at 24 words. Both agent's have similar word counts for further messages as they refine the location. See the appendix for details on common utterances for both roles in the first two rounds of dialog.

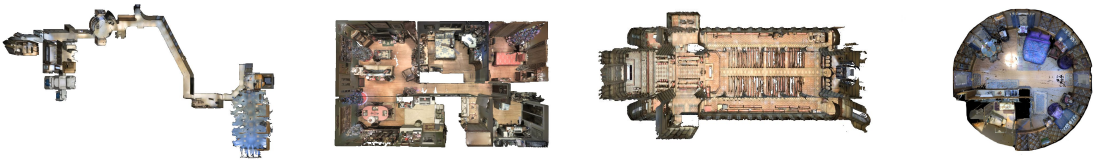
Role of Navigation. Often the localization task can be made easier by having the Observer move to reduce uncertainty (see bottom example of Fig. 4). This includes moving away from nondescript areas like hallways and moving to unambiguous locations. We observe at least one navigation step in 62.6% of episodes and an average of 2.12 steps. Episodes containing navigation have a significantly lower average localization error (2.70m) compared to those that did not (3.98m). We also observe the intuitive trend that larger environments elicit more navigation. The distributions for start and end locations for the most and least navigated environments in the appendix.

3.3 WHERE ARE YOU? Tasks

We now formalize the LED, EVD and CL tasks to provide a clear orientation for future work.

Localization from Embodied Dialog. The LED task is the following – given an episode comprised of a environment and human dialog – $(e, L_0, O_0, \ldots L_{T-1}, O_{T-1})$ – predict the Observer's final location p_T. This is a grounded natural language understanding task with pragmatic evaluations – localization error and accuracy at a variable threshold which in this paper is set to 3 meters. This task does not require navigation or text generation; instead, it mirrors AI-augmented localization applications. An example would be a system that listens to emergency services calls and provides a real time estimate of the caller's indoor location to aid the operator.

Embodied Visual Dialog. This task is to replace the Observer by an AI agent. Given a embodied first-person view of a 3D environment (see Observer view in Fig. 1), and a partial history of dialog consisting of k Locator and $k-1$ Observer message pairs (L_0: *'describe your location.'*, O_0: *'I'm in a kitchen with black counters.'*, $L_1 \ldots$): pre-



a) Longest navigation distance b) Shortest navigation distance c) Highest localization error d) Lowest localization error

Figure 3: Environments with the largest/smallest mean navigation distance (a, b) and mean localization error (c, d). Observers tend to navigate more in featureless areas, such as the long corridor in (a). Localization error is highest in buildings with many repeated indistinguishable features, such as the cathedral with rows of pews in (c).

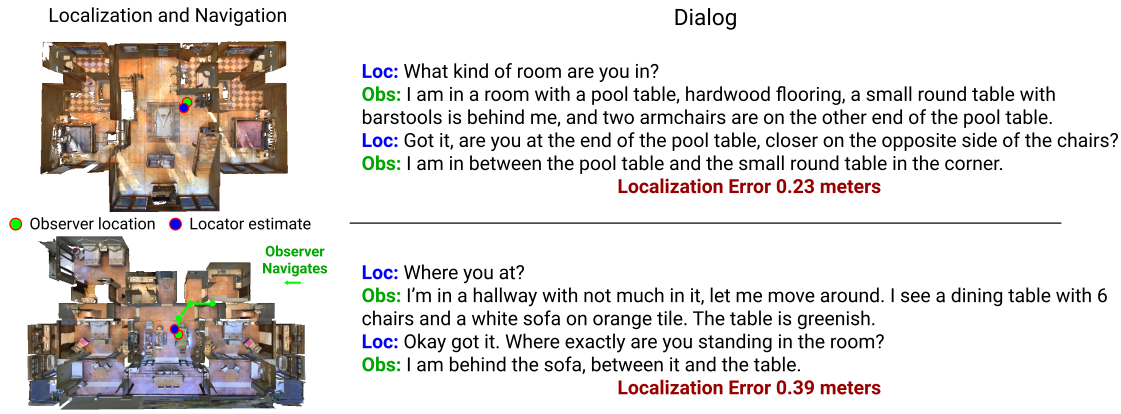


Figure 4: Examples from the dataset illustrating the Observer's location on the top-down map vs. the Locator's estimate (left) and the associated dialog (right). In the bottom example the Locator navigates to find a more discriminative location, which is a common feature of the dataset. The Observer navigates in 63% of episodes and the average navigation distance for these episodes is 3.4 steps (7.45 meters).

dict the Observer agent's next navigational action and natural language message to the Locator. To evaluate the agent's navigation path, the error in the final location can be used along with path metrics such as nDTW (Ilharco et al., 2019). Generated text can be evaluated against human responses using existing text similarity metrics.

Cooperative Localization. In this task, both the Observer and the Locator are modeled agents. Modeling the Locator agent requires goal-oriented dialog generation and confidence estimation to determine when to end the task by predicting the location of the Observer. Observer and Locator agents can be trained and evaluated independently using strategies similar to the EVD task, or evaluated as a team using localization accuracy as in LED.

4 Modeling Localization From Embodied Dialog

While the WAY dataset supports multiple tasks, we focus on Localization from Embodied Dialog as a first step. In LED, the goal is to predict the location of the Observer given a dialog exchange.

4.1 LED Model from Top-down Views

We model localization as a language-conditioned pixel-to-pixel prediction task – producing a probability distribution over positions in a top-down view of the environment. This choice mirrors the environment observations human Locators had during data collection, allowing straightforward comparison. However, future work need not be restricted to this choice and may leverage the panoramas or 3D reconstructions that Matterport3D provides.

Dialog Representation. Locator and Observer messages are tokenized using a standard toolkit (Loper and Bird, 2002). The dialog is represented as a single sequence with identical 'start' and 'stop' tokens surrounding each message, and then encoded using a single-layer bidirectional LSTM with a 300 dimension hidden state. Word embeddings are initialized using GloVe (Pennington et al., 2014) and finetuned end-to-end.

Environment Representation. The visual input to our model is the environment map which we scale to 780×455 pixels. We encode this map using a ResNet18 CNN (He et al., 2016) pretrained on ImageNet (Russakovsky et al., 2015), discarding

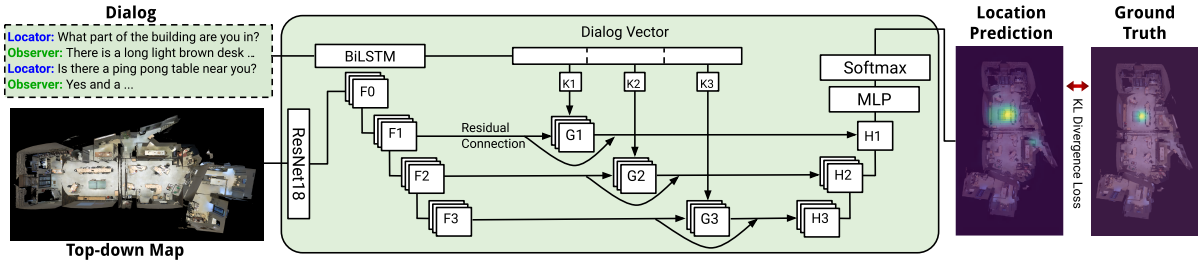


Figure 5: The 3-layer LingUNet-Skip architecture used to model the Localization from Embodied Dialog task.

Table 2: Comparison of our model with baselines and human performance on the LED task. We report average localization error (LE) and accuracy at 3 and 5 meters (all ± standard error). * denotes oracle access to Matterport3D node locations.

	val-seen			val-unseen			test		
Method	LE ↓	Acc@3m ↑	Acc@5m ↑	LE ↓	Acc@3m ↑	Acc@5m ↑	LE ↓	Acc@3m ↑	Acc@5m ↑
Human Locator	3.26±0.71	72.3±3.0	78.8±3.0	1.91±0.32	79.7±3.0	85.2±1.7	3.16±0.35	70.4±1.4	77.2±1.3
Random	12.39±0.31	5.4±0.9	15.0±1.3	10.18±0.16	7.0±0.7	21.3±1.1	13.10±0.17	6.6±0.5	15.2±0.7
Random Node*	8.27±0.44	18.1±2.2	37.8±2.7	10.44±0.31	15.8±1.1	29.0±1.4	13.19±0.32	12.8±0.7	24.9±0.9
Center	6.13±0.25	23.1±2.4	46.5±2.9	**4.90±0.12**	29.8±1.9	61.0±2.1	**6.71±0.14**	22.6±1.2	42.3±1.4
Heuristic	11.6±0.49	12.5±1.8	23.6±2.4	10.10±0.28	10.5±1.2	25.7±1.8	13.45±0.32	9.1±0.8	18.4±1.1
No Language	7.17±0.42	26.1±2.5	44.8±2.9	5.72±0.20	32.1±2.0	58.1±2.1	7.67±0.18	22.3±1.2	42.4±1.4
No Vision	11.36±0.46	9.4±1.7	18.4±2.2	8.58±0.20	7.8±1.1	22.1±1.8	11.62±0.23	7.7±0.8	18.3±1.1
LingUNet	**4.73±0.32**	**53.5±2.9**	**67.2±2.7**	5.01±0.19	**45.6±2.1**	**63.6±2.0**	7.32±0.22	**32.7±1.4**	**49.5±1.5**

the 3 final conv layers and final fully-connected layer in order to output a 98×57 spatial map with feature dimension 128. Although the environment map is a top-down view which does not closely resemble ImageNet images, in initial experiments we found that using a pretrained and fixed CNN improved over training from scratch.

Language-Conditioned Pixel-to-Pixel Model. We adapt a language-conditioned pixel-to-pixel LingUNet (Misra et al., 2018) to fuse the dialog and environment representations. We refer to the adapted architecture as LingUNet-Skip. As illustrated in Fig. 5, LingUNet is a convolutional encoder-decoder architecture. Additionally we introduce language-modulated skip-connections between corresponding convolution and deconvolution layers. Formally, the convolutional encoder produces feature maps $F_l = \text{Conv}(F_{l-1})$ beginning with the initial input F_0. Each feature map F_l is transformed by a 1×1 convolution with weights K_l predicted from the dialog encoding, i.e. $G_l = \text{Conv}_{K_l}(F_l)$. The language kernels K_l are linear transforms from components of the dialog representation split along the feature dimension. Finally, the deconvolution layers combine these transformed skip-connections and the output of the previous layer, such that $H_l = \text{Deconv}([H_{l+1}; (G_l + F_l)])$. There are three layers and the output of the final deconvolutional is processed by a MLP and a softmax to output a distribution over pixels.

Loss Function. We train the model to minimize the KL-divergence between the predicted location distribution and the ground-truth location, which we smooth by applying a Gaussian with standard deviation of 3m (matching the success criteria). During inference, the pixel with highest probability is selected as the final predicted location. For multi-story environments, we process each floor independently and the location with the highest predicted probability over all floors is selected.

4.2 Experimental Setup

Metrics. We evaluate performance using localization error (LE) defined as the Euclidean distance in meters between the predicted Observer location $\hat{p}_T$ and the Observer's actual terminal location p_T: $\text{LE} = ||p_T - \hat{p}_T||_2$. We also report a binary success metric that places a threshold k on the localization error – $\mathbb{1}(\text{LE} \leq k)$ – for 3m and 5m. The 3m threshold allows for about one viewpoint of error since viewpoints are on average 2.25m apart. We use euclidean distance for LE because localization predictions are not constrained to the navigation graph. Matterport building meshes contain holes and other errors around windows, mirrors and glass walls, which can be problematic when computing

geodesic distances for points off the navigation graph.

Training and Implementation Details. Our LingUNet-Skip model is implemented in PyTorch (Paszke et al., 2019). Training the model involves optimizing around 16M parameters for 15–30 epochs, requiring ∼8 hours on a single GPU. We use the Adam optimizer (Kingma and Ba, 2014) with a batch size of 10 and an initial learning rate of 0.001 and apply Dropout (Srivastava et al., 2014) in non-convolutional layers with $p = 0.5$. We tune hyperparameters based on val-unseen performance and report the checkpoint with the highest val-unseen Acc@3m. To reduce overfitting we apply color jitter, 180° rotation, and random cropping by 5% to the map during training.

Baselines. We consider a number of baselines and human performance to contextualize our results and analyze WAY:

- **Human Locator.** The average performance of AMT Locator workers as described in Sec. 3.
- **Random.** Uniform random pixel selection.
- **Center.** Always selects the center coordinate.
- **Random Node.** Uniformly samples from Matterport3D node locations. This uses oracle knowledge about the test environments. While not a fair comparison, we include this to show the structural prior of the navigation graph which reduces the space of candidate locations.
- **Heuristic Driven.** For each dialog D_t in the validation splits we find the most similar dialog D_g in the training dataset based on BLEU score (Papineni et al., 2002). From the top-down map associated with D_g, a 3m x 3m patch is taken around the ground truth Observer location. We predict the location for D_t by convolving this patch with the top-down maps associated with D_t and selecting the most similar patch (according to Structural Similarity). The results (below) are only slightly better than random.

4.3 Results

Tab. 2 shows the performance of our LingUNet-Skip model and relevant baselines on the val-seen, val-unseen, and test splits of the WAY dataset.

Human and No-learning Baselines. Humans succeed 70.4% of the time in test environments. Notably, val-unseen environments are easier for humans (79.7%), see appendix for details. The Random Node baseline outperforms the pixel-wise Random setting (Acc@3m and Acc@5m for all splits) and this gap quantifies the bias in nav-graph positions. We find the Center baseline to be rather strong in terms of localization error, but not accuracy – wherein it lags behind our learned model significantly (Acc@3m and Acc@5m for all splits).

Table 3: Modality, modeling, and dialog ablations for our LingUNet-Skip model on the validation splits of WAY.

	val-seen		val-unseen	
	LE ↓	Acc@3m ↑	LE ↓	Acc@3m ↑
Full LingUNet-Skip Model	**4.73**±**0.32**	**53.5**±**2.9**	5.01±0.19	45.6±2.1
w/o Data Aug.	5.98±0.35	41.1±2.0	5.44±0.18	35.7±2.1
w/o Residual	5.26±0.33	47.5±2.9	4.74±0.17	43.1±2.1
No Dialog	7.17±0.42	26.1±2.5	5.72±0.20	32.1±2.0
First-half Dialog	5.06±0.33	50.5±2.8	**4.71**±**0.18**	**46.2**±**2.1**
Second-half Dialog	5.29±0.28	41.8±2.8	5.06±0.17	38.7±2.1
Observer-only	5.73±0.36	45.2±2.9	4.77±0.17	44.9±2.1
Locator-only	6.39±0.37	30.4±2.7	5.63±0.19	33.3±2.0
Shuffled Rounds	5.32±0.32	42.8±2.8	4.67±0.18	44.9±2.1

LingUNet-Skip outperforms baselines. Our LingUNet-Skip significantly outperforms the hand-crafted baselines in terms of accuracy at 3m – improving the best baseline, Center, by an absolute 10% (test) to 30% (val-seen and val-unseen) across splits (a 45-130% relative improvement). Despite this, it achieves higher localization error than the Center model for val-unseen and test. This is a consequence of our model occasionally being quite wrong despite its overall stronger localization performance. There remains a significant gap between our model and human performance – especially on novel environments (70.4% vs 32.7% on test).

4.4 Ablations and Analysis

Tab. 3 reports detailed ablations of our LingUNet-Skip model. Following standard practice, we report performance on val-seen and val-unseen.

Navigation Nodes Prior We do not observe significant differences between val-seen (train environments) and val-unseen (new environments), which suggests the model is not memorizing the node locations. Even if the model did, learning this distribution would likely amount to free-space prediction which is a useful prior in localization.

Input Modality Ablations. No Vision explores the extent that linguistic priors can be exploited by LingUNet-Skip, while No Dialog does the same for visual priors. No Dialog beats the Center baseline (32.1% vs. 29.8% val-unseen Acc@3m) indicating that it has learned a visual centrality prior that is stronger than the center coordinate. This makes sense because some visual regions like nondescript hallways are less likely to contain terminal Ob-

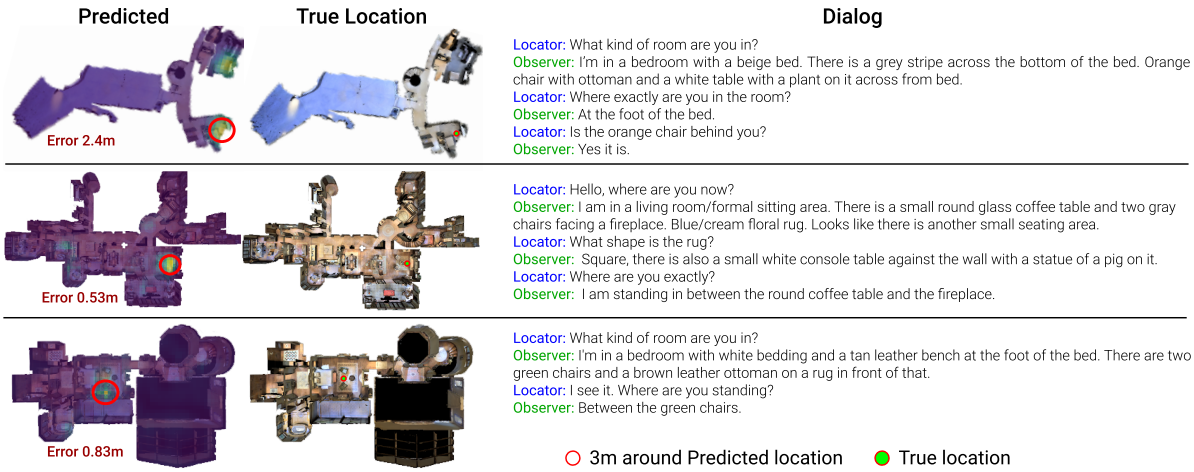


Figure 6: Examples of the predicted distribution versus the true location over top down maps of environment floors for dialogs in val-unseen. The red circle on the left represents the three meter threshold around the predicted localization. The green dot on the middle image represents the true location. The localization error in meters of the predicted location is shown in red.

server locations. Both No Vision and No Dialog perform much worse than our full model (7.8% and 32.1% val-unseen Acc@3m vs. 45.6%), suggesting that the task is strongly multimodal.

Dialog Halves. First-half Dialog uses only the first half of dialog pairs, while Second-half Dialog uses just the second half. Together, these examine whether the start or the end of a dialog is more salient to our model. We find that First-half Dialog performs marginally better than using the full dialog (46.2% vs 45.6% val-unseen Acc@3m) which we suspect is due to our model's failure to generalize second half dialog to unseen environments and problems handling long sequences. Further intuition for these results is that the first-half of the dialog contains coarser grained descriptions and discriminative statements ("I am in a kitchen"). The second-half of the dialog contains more fine grained descriptions (relative to individual referents in a room). Without the initial coarse localization, the second-half dialog is ungrounded and references to initial statements are not understood, therefore leading to poor performance.

Observer dialog is more influential. Observer-only ablates Locator dialog and Locator-only ablates Observer dialog. We find that Observer-only significantly outperforms Locator-only (44.9% vs. 33.3% val-unseen Acc@3m). This is an intuitive result as Locators in the WAY dataset commonly query the Observer for new information. We note that Observers were guided by the Locators in the collection process (e.g. *'What room are you in?'*), and that ablating the Locator dialog does not remove this causal influence.

Shuffling Dialog Rounds. Shuffle Rounds considers the importance of the order of Locator-Observer dialog pairs by shuffling the rounds. Shuffling the rounds causes our LingUNet-Skip to drop just an absolute 0.7% val-unseen Acc@3m (2% relative).

Model Ablations. Finally, we ablate two model-related choices. Without data augmentation (w/o Data Aug.), our model drops 9.9% val-unseen Acc@3m (22% relative). Without the additional residual connection (w/o Residual), our model drops 2.5% val-unseen Acc@3m (5% relative).

5 Conclusion and Future Work

In summary, we propose a new set of embodied localization tasks: Localization from Embodied Dialog - LED (localizing the Observer from dialog history), Embodied Visual Dialog - EVD (modeling the Observer), and Cooperative Localization - CL (modeling both agents). To support these tasks we introduce WHERE ARE YOU? a dataset containing ∼6k human dialogs from a cooperative localization scenario in a 3D environment. WAY is the first dataset to present extensive human dialog for an embodied localization task. On the LED task we show that a LingUNet-Skip model improves over simple baselines and model ablations but without taking full advantage of the second half of the dialog. Since WAY encapsulates multiple embodied localization tasks, there remains much to be explored.

Acknowledgments

Partial funding for this work was provided by NIH award R01MH114999.

References

Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. 2019. Chasing ghosts: Instruction following as bayesian state tracking. In *NeurIPS*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Jason Baldridge, Tania Bedrax-Weiss, Daphne Luong, Srini Narayanan, Bo Pang, Fernando Pereira, Radu Soricut, Michael Tseng, and Yuan Zhang. 2018. Points, paths, and playscapes: Large-scale spatial language understanding tasks set in the real world. In *International Workshop on Spatial Language Understanding*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. 2018. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *CoRL*.

Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, and Yoav Artzi. 2019. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *CoRL*.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *3DV*. MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP$_$TOS.pdf.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *CVPR*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *CVPR*.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

William Falcon and Henning Schulzrinne. 2018. Predicting floor-level for 911 calls with neural networks and smartphone sensor data. In *ICLR*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *CVPR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Maria Kozhevnikov, Michael A Motes, Bjoern Rasch, and Olessia Blajenkova. 2006. Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. *Applied Cognitive Psychology*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *EMNLP*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint arXiv:0205028*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*.

Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. 2020. Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. *arXiv preprint arXiv:2001.03671*.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction.

Khanh Nguyen and Hal Daumé III. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *CVPR*.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

6 Appendix

Val-Unseen has higher accuracy than other splits. Human's localization Acc@3m is 79.4% for val-unseen which is higher than all other splits such as test which as a Acc@3m of 70.4%. Following standard practice, the splits followed (Chang et al., 2017). The val-unseen split is notably smaller than the rest of the splits and through qualitative analysis, we found that the environments in the val-unseen split (Chang et al., 2017) are generally smaller and have discriminative features which we attribute to the split having a high localization performance. Our LingUNet-Skip model has lower performance on test than on val-unseen which we reason is be to the nature of the environments in the splits. Additionally the LingUNet-Skip model has lower performance on test than on val-seen which is expected because test environments are unseen environments and val-seen environments are contained in the training set.

Navigation differs between enviroments. As previously discussed, different environments in the WAY dataset have varying levels of navigation. This is likely attributed to a few factors such as size of the building and discriminative features of the building such as decorations. Additionally we see features like long hallways frequently lead to long navigational paths. The variances in navigation between environments is further illustrated in Fig. 7. While the distribution between the starting and final positions barely changes for the environment on the left, we see significant change in the environment on the right. Most noticeably we see that there are no final positions in the long corridor of the right environment despite it containing several start locations.

Data Collection Interface. Fig. 8 shows the data collection interface for the Observer and Locator human annotators. The annotator team was able to chat with each other via a message box that also displayed the chat history. The Locator had a top down map of the environment and had buttons to switch between floors. The Observer was given a first person view of the environment and could navigate the environment by clicking on the blue cylinders shown in Fig. 8

Closer Look at Dialog. Fig. 9 further breaks this down by looking at the average length of specific messages of the two agents. The Locator's first message is short in comparison to the average number of words per message of the agent. This is expected as this message is always some variation of getting the Observer to describe their location and it follows that the message has a low number of unique words. The Observer's first message is by far their longest, at 23.9 words, which is logical since in this message the Observer is trying to give the most unique description possible with no constraint on length. The distributions become more uniform in the 2nd messages from both the Locator and Observer. While the first message of the observer has a large number of unique words the distribution is not uniform over the words leading to the conclusion that the message has an common structure to it but that the underlying content is still discriminative for modeling the location. The word distribution of messages further down in the dialogue sequence are largely conditioned on the previous message from the other agent, which means that accurately encoding the dialogue history is important for accurate location estimation.

Distribution of Localization Error. In order to better understand the distribution of the LingUNet-Skip model's predictions we visualize the distributions in Fig. 10.

Success and Failure Examples. To qualitatively evaluate our model we visualize the predicted distributions, the true location over the top down map and the dialog in Fig. 11. We also show two failure cases in which the model predicts the wrong location.

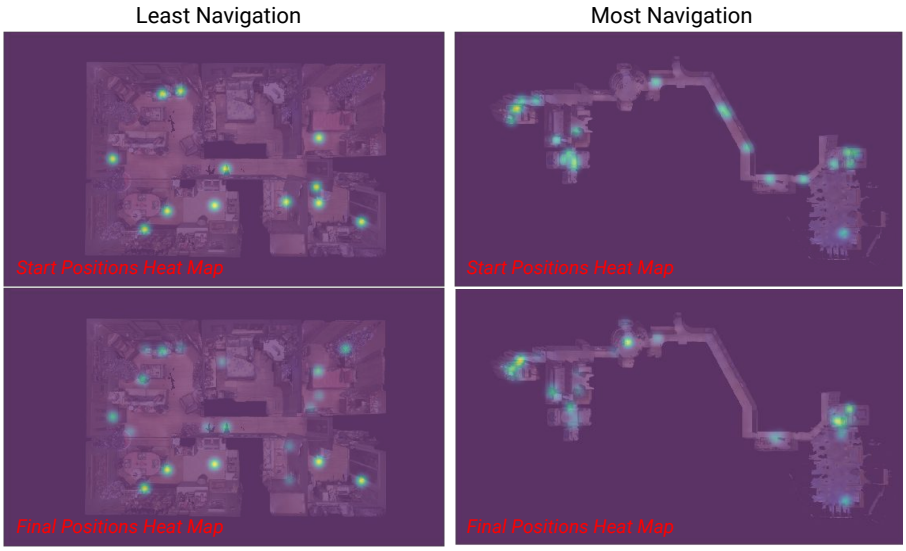


Figure 7: Shows the distribution of the starting and ending locations of the Observer for two environments in the WAY dataset. On the left is the environment that had annotations which the least amount of navigation. On the right is the environment that had annotations with the most amount of navigation.

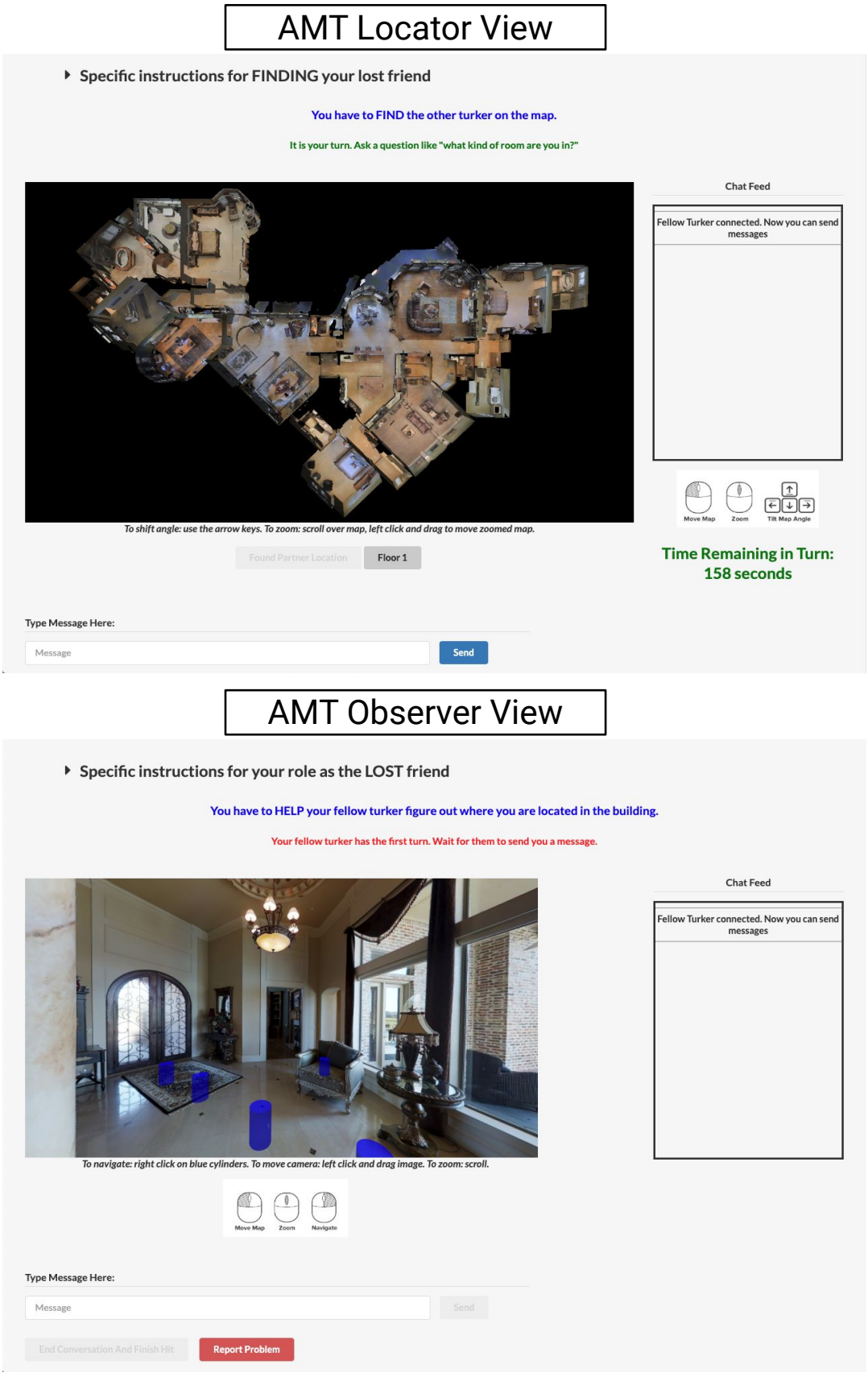
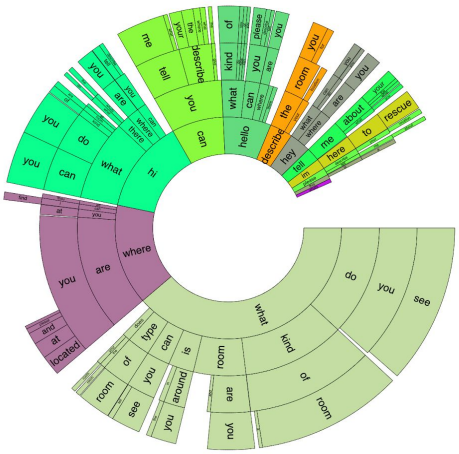
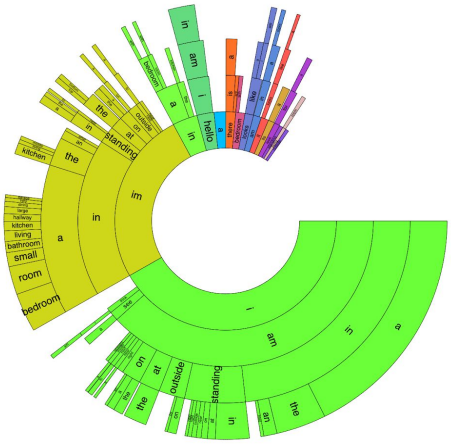


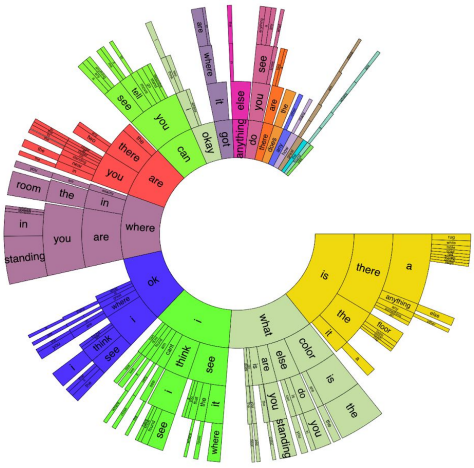
Figure 8: The dataset collection interface for WAY. These are the interfaces that the Observer and Locator workers used on Amazon Mechanical Turk.



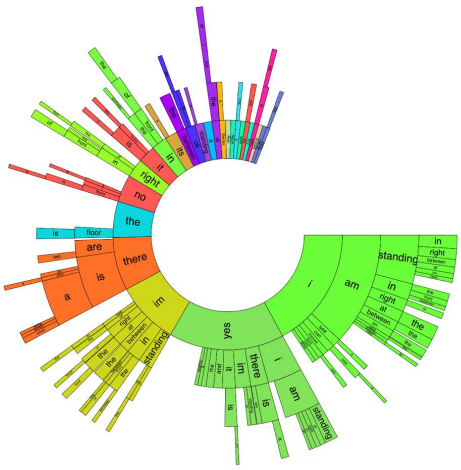
Locator 1st Message
Avg. Number of Words: *7.0*
Number Unique Words: *695*



Observer 1st Message
Avg. Number of Words: *23.9*
Number Unique Words: *3104*



Locator 2nd Message
Avg. Number of Words: *9.83*
Number Unique Words: *1663*



Observer 2nd Message
Avg. Number of Words: *16.39*
Number Unique Words: *2677*

Figure 9: Distributions of the first four words for each of the first four messages of the dialogs in the WAY dataset separated by message number and role type. The ordering of the words starts in the center and radiates outwards. The arc length is proportional to the number of messages containing the word. The white areas are words that had too low of a count to illustrate.

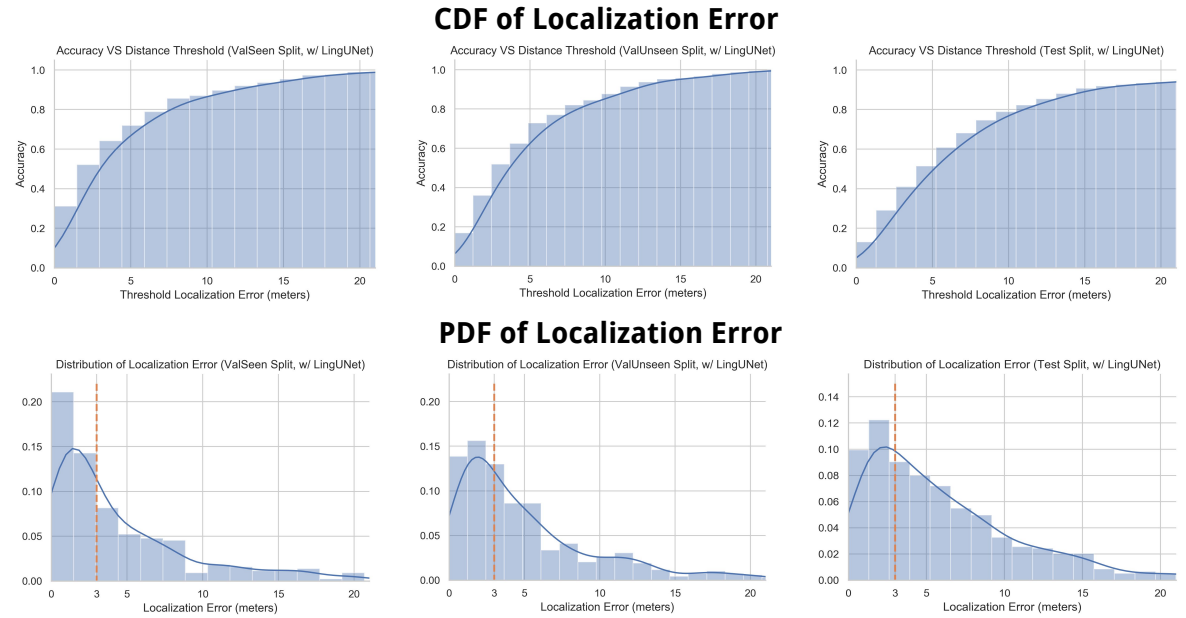


Figure 10: The top row is the cdf of localization errors on the val and test splits using the LingUNet-Skip model. These graphs can also be interpreted as the accuracy vs threshold of error which defines success. The bottom row is the probability distribution of localization errors from the LingUNet-Skip model across the val and test splits.

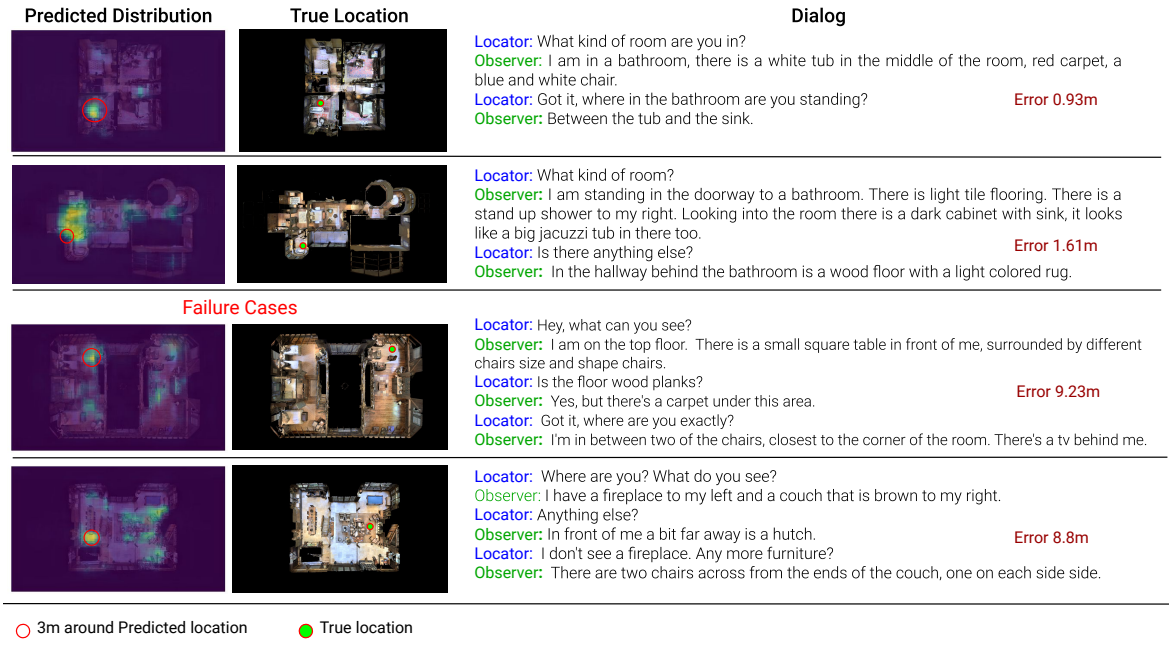


Figure 11: Examples of the predicted distribution versus the true location over top down map of a floor of an environment for a given dialog in val-unseen. On the left the red circle represents the three meter threshold around the predicted localization. On the middle image the green dot represents the true location. The localization error in meters of the predicted location shown in red in the dialog box.

Locator: Hi! What kind of room are you in?
Lost Friend: I am at the very top of the white marble spiral staircase. There is wood floor at the top and a balcony with black railing.
Locator: Are you between the two sets of stairs on the landing?
Lost Friend: I am at the top of the stairs.

Locator: Where are you located at?
Lost Friend: I'm in a kitchen that has a long curved wall. I'm near the entrance which is near some stairs.
Locator: Are the stairs going up or down?
Lost Friend: Down, it looks like I'm on the top floor.
Locator: What else do you see?
Lost Friend: There is a black table in the kitchen. One side is curved to match the curved of the kitchen wall.

Locator: What kind of room are you in?
Lost Friend: I am in a bedroom with one blue wall.
Locator: With the striped bed sheets and two tan nightstands??
Lost Friend: Yes! I am at the foot of the bed.

Locator: Hi, what kind of a room are you in?
Lost Friend: I'm in the kitchen standing in front of the stove.
Locator: How close to the stove are you?
Lost Friend: I could fry eggs if I wanted to without moving.

Locator: Hello, what type of room are you in?
Lost Friend: I am outside on the second step from the top of a windy staircase, overlooking the swimming pool.
Locator: Are you indoors or outdoors?
Lost Friend: am outside.
Locator: What material are you standing on?
Lost Friend: I am on the second step from the top of the staircase.
Locator: The only staircase I see is inside. Unless the rocks near the pool are the stairs you are talking about.
Lost Friend: I am outside, the staircase is spiral and it is black. There is a larger pool and smaller and I am on the side of the smaller one.

Locator: Hey. Where are you?
Lost Friend: I am in a study with a foosball table! There is a cream rug on the floor surrounded by white tile. A brown desk is in the corner between two brown bookcases on either side.
Locator: This is a very, very big house. Any colors that stand out will help. I would say there has got to be 30 rooms in this house. I am looking now.
Lost Friend: I am on the same floor as the swimming pool. There is a double door across from the study that looks out onto a patio and the pool is in the distance.
Locator: Found it. Where are you standing?
Lost Friend: Standing between the foosball table and the corner desk.

Figure 12: Examples of dialog in the WAY dataset.