DiaLoc: An Iterative Approach to Embodied Dialog Localization

Chao Zhang   Mohan Li   Ignas Budvytis   Stephan Liwicki

Toshiba Europe Ltd

firstname.lastname@toshiba.eu

Abstract

Multimodal learning has advanced the performance for many vision-language tasks. However, most existing works in embodied dialog research focus on navigation and leave the localization task understudied. The few existing dialog-based localization approaches assume the availability of entire dialog prior to localizaiton, which is impractical for deployed dialog-based localization. In this paper, we propose DiaLoc, a new dialog-based localization framework which aligns with a real human operator behavior. Specifically, we produce an iterative refinement of location predictions which can visualize current pose believes after each dialog turn. DiaLoc effectively utilizes the multimodal data for multi-shot localization, where a fusion encoder fuses vision and dialog information iteratively. We achieve state-of-the-art results on embodied dialog-based localization task, in single-shot (+7.08% in Acc5@valUnseen) and multi-shot settings (+10.85% in Acc5@valUnseen). DiaLoc narrows the gap between simulation and real-world applications, opening doors for future research on collaborative localization and navigation.

1. Introduction

Picture yourself becoming lost within a novel building during a visit to a friend's place of residence or work place. In this scenario, you reach out to your friend and detail the elements of your environment. Through a series of exchanges, it is anticipated that your friend will eventually figure out your whereabouts. The completion of this collaborative endeavor depends on the skillful use of targeted inquiries and unequivocal linguistic responses. In this study, we formulate this task as an iterative embodied dialog localization problem involving two agents: an Observer, randomly situated within an environment, and a Locator, whose job is to localize the observer with a provided top-down map through dialog.

Iterative embodied dialog localization is of clear relevance to numerous real-world applications, including the utilization of mobile robots for search and rescue opera-
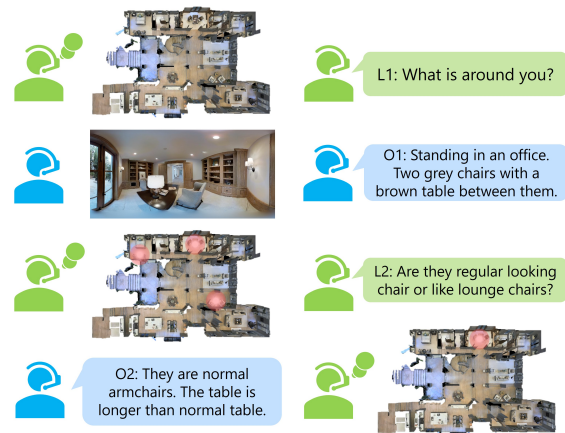

Figure 1. **Illustration of iterative embodied dialog localization.** The locator with top-down map and the observer with egocentric view engage in a cooperative dialog to assist in determining the observer's location. The locator iteratively forms predictions to enhance the estimations. The preceding dialog exchanges and cumulative predictions also influence the manner in which the locator poses new questions. As depicted, the locator predicts 3 possible locations based on the first turn. The second question is asked to disambiguate the predictions and the correct location is predicted given the answer from the observer.

tions [17, 19]. Nevertheless, it is worth noting that dialog-based localization, in comparison to navigation, is still a relatively understudied area. The dataset WAY sourced dialogs from human annotations and formulated dialog-based localization through employing a full dialogue and a top-down map to determine an agent's precise location. The sparsity and problem formulation pose however two challenges: Firstly, its utilization of multimodal information to overcome information imbalances stemming from egocentric and top-down viewpoints. Secondly, the dataset's scarcity restricts the capacity of trained models to generalize to unfamiliar environments.

In an initial attempt to address the challenge of localization through embodied dialog, Hahn *et al.* [9] introduced a novel approach by framing the task as an image-to-image

problem. They adopted LingUNet network to predict localization distribution. Language embedding generated by BiLSTM [10] is used as conditional input to guide the learning of visual features. Leveraging a dataset of human annotations, their approach assumes the availability of the complete dialog prior to making predictions. However, this approach contrasts with human intuition, as human localization predictions are characterized by incremental refinement. As illustrated in Figure 1, the locator actively makes location guesses as new dialog information arrives. This dynamic process is also reflected in questioning: the locator poses new questions based on prior dialog exchanges and their cumulative predictions. Therefore, developing a localization technique capable of continual location estimation stands as a fundamental cornerstone for realizing the practical application of cooperative search and rescue.

Inspired by cooperative search and rescue we therefore introduce a novel approach which we call *iterative embodied dialog localization*. This formulation requires the locator to possess the capability of finding out the observer's location following each turn within the dialog. When compared to existing approaches that consider the complete dialog, our iterative framework offers several merits. Firstly, it operates with greater efficiency. During the inference phase, it only processes a single turn of the dialog, as opposed to the entire conversation. In practical application, this efficiency can translate into time savings, or life savings, as the location can be estimated or narrowed down to a few points before the conversation finishes. Secondly, our approach exhibits enhanced generalization to novel locations. The risk of over-fitting to training data is mitigated by not relying on complete dialogs during the training process. Lastly, the intermediate predictions provide vital cues for assessing uncertainty and are indispensable for tasks that involve dialog generation, such as embodied visual dialog and cooperative localization. In summary, this transition towards an iterative dialog-based localization approach not only aligns with engineering perspectives but also paves the way for improved performance and broader applicability across various related tasks.

Multimodal learning, particularly through the integration of vision and language, has emerged as a prominent topic in deep learning era. The paradigm of Vision-Language Pretraining (VLP) [13, 14] has advanced the performance of numerous vision-language tasks. VLP employs Transformers to acquire either unimodal representations or fused multimodal representations. Inspired by VLP's success in multiple vision language understanding and generation tasks, such as image captioning and Visual Question Answering(VQA), we embrace the Transformer-based encoders. This forms the backbone for our novel multimodal multi-shot localizer, which becomes the state of the art.

In summary, we make the following **contributions**: (i)

We introduce an iterative approach towards practical embodied dialog localization. This formulation aligns with the intuitive behavior of a human locator who employs incremental predictions to improve localization performance. (ii) We propose a novel multimodal multi-shot localizer DiaLoc based on Transformer encoders. (iii) We demonstrate state-of-the-art performance for embodied dialog localization. Our proposed iterative solution exhibits enhanced generalization capabilities. (iv) We showcase that the proposed approach posses the ability to rectify previous prediction with new information and may help a human operator in question formulation for search and rescue applications.

2. Related Work

Vision and Language Vision and language understanding has been extensively explored across tasks such as image captioning, visual question answering (VQA) and visual grounding [1, 4, 7, 24]. The fusion of language and vision for recognition is also an active domain. Notably, CLIP [20] effectively learns visual concepts through natural language supervision. The fundamental insight is to employ natural language as a versatile space for predictions, facilitating zero-shot generalization and transfer learning.

In the context of language-driven tasks, LSeg [12] was proposed for language-driven semantic image segmentation using flexible text embeddings. It allows LSeg to generalize to unseen categories at test time. Delving into the intersection of vision and language with navigation, Visual and language navigation (VLN) combines visual and linguistic inputs to enhance robot navigation in building-scale environments [2]. Moreover, vision and language navigation in continuous environments (VLN-CE) is an instruction-guided navigation task. It relaxes the assumptions inherent in the original VLN task, striving to bridge the gap between simulated environments and real-world scenarios. Instead of utilizing a labeled dataset with trajectories annotations, LM-Nav [21] shows that a robotic navigation system can be constructed purely from pre-trained models using CLIP and GPT-3 [4] without any fine-tuning or language-annotated robot data.

Embodied Localization Estimating an agent's position within the environment forms a vital aspect of applications in embodied AI. Hahn *et al.* [9] introduced the task of localization from embodied dialog (LED), leveraging the WAY dataset containing approximately 10,000 dialogs curated by human operators. The authors also proposed a baseline for LED based on LingUnet [18], using a top-down map as input to predict location. Choosing an effective map representation is a key consideration in this task. Unlike [9], Hann *et al.* [8] proposed an Transformer-based method using navigation graph extracted from Matterport3D [5]. In a similar vein to LED, Cooperative Vision-and-Dialog Navigation (CVDN) [22] introduces a dataset

consisting of dialogues where an oracle assists a navigator in completing a navigation task. The baseline approach adopts a sequence-to-sequence approach. Both WAY and CVDN are constructed upon the Matterport3D dataset, albeit with different focuses: WAY concentrates on localization, while CVDN focus on navigation. Further, Robot-Slang [3] brings a dataset that contains 169 dialogs between a human directing a robot and another human offering guidance towards navigation goals. Beyond multimodal localization, Text2Pos [11] delves into cross-modal text to point cloud localization based on Kitti360 [15]. Within this context, the proposed approach identifies positions specified by template-based language descriptions of immediate surroundings within the environment represented as a point cloud.

3. Iterative Embodied Dialog Localization

We first describe the task of iterative embodied dialog based localization. Then, the architecture of our proposed model is introduced. These are followed by the training objectives and implementation details.

3.1. Task and Evaluation Metrics

We perform embodied localization using dialogs $D = (L_1, O_1, ..., L_T, O_T)$ and the global map e, here the human dialog is represented by T turns between the utterance of the locator L_i and the observer O_i. The goal is to predict the observer's final location p_T. Different from the previous approaches making a single prediction at the end of the whole dialog, the proposed iterative approach outputs multiple predictions for each turn $p_1, ..., p_T$. The goal is to enhance communication efficacy and spatial understanding through a series of information exchanges between the locator (top-down view) and the observer (egocentric). To evaluate the localization error (LE), we follow [8] to use geodesic distance instead of euclidean distance, as geodesic distance is more meaningful for determining error across rooms. Pixel coordinate p_T of the location is snapped to the nearest way-point node g_T before the calculation of localization error as: LE $= ||g_T - \hat{g}_T||^2$, where g_T is the ground-truth and $\hat{g}_T$ is the prediction. Binary success rate for LE at threshold k is also reported for represented values such as 0m and 5m. We draw Cumulative Matching Characteristics (CMC) curves to compare different approaches to characterize the holistic performance.

The WAY dataset [9] is used to evaluate the proposed method. It is split into train (9,955), valSeen (305), valUnseen (579) and test sets (1,200). There exists two evaluation protocols: floor-level [9] and building-level [8]. As opposed to non-iterative methods, our proposed method focus on exploiting the sequential context from a dialog and has the potential of stopping the dialogs when predictions are accurate. Achieving the early-stopping requires the turn-

based locations since the observer is allowed to move during the dialogs. To further improve the training efficiency, our method follows [9] and use the floor-based setting assuming the known floor. We found that neither the target floor information or the per-turn location is available from the test split and hence not feasible for the evaluation purpose. As such, all methods are trained using train split and validated using valSeen, and the performance on valUnSeen is used to show the generalization in novel scenes.

There are two advantages of evaluating on the dialogs with known floor and turn-based dense locations: (1) Comparing to methods that use all floors during training, evaluating using the target floor map is efficient and easier for us to investigate the iterative improvement; (2) With corresponding locations at each turn of the dialog, we could measure the localization performance at different timesteps. This is useful to quantify the potential of early termination.

Compared to non-iterative embodied dialog localization, the new iterative approach offers improved adaptability and accuracy. It enables agents to progressively refine their spatial understanding through ongoing dialogs, accommodating dynamic environments representation and nuanced contextual information. This contrasts with the previous single-shot methods, and allows for more effective training, better generalization to unseen environments, and ultimately more practical embodied applications.

3.2. Architecture

Previously, LED [9] formalized embodied dialog localization as one image-to-image translation problem. The whole dialog is used as conditional input to the encoder-decoder network. Unlike LED, we argue that multi-turn predictions happened at the end of each turn are crucial for the applications such as search and rescue. This also resembles the way how human performs dialog-based localization. To reduce this gap, we adopt Transformer [23] backbone for the new task. Inspired by recent work on multimodal learning [13, 14]. Our design features an image backbone which interacts with dialog input iteratively. Location predictions are made at multiple timesteps. We now describe the network architecture details as shown in Figure 2.

Unimodal Encoders. To obtain unimodal embeddings, we use Transformer-based encoder for both vision and language inputs. Vision transformers (ViT) pre-trained on ImageNet is used as the image encoder for top-down map e. We fine-tune ViT to generate the visual embedding $V \in R^{M,C}$, where $M = 196$ is the number of visual tokens. Input map is resized to 224×224. To encode dialog $D = (L_1, O_1, \cdots, L_T, O_T)$ of T turns, we adopt the pre-trained Bert [6] as the text encoder. The dialog D is represented as features $L \in R^{N,C}$, where $N = 100$ is the max token length. Note that the feature dimension C is 768 for both ViT and BERT.
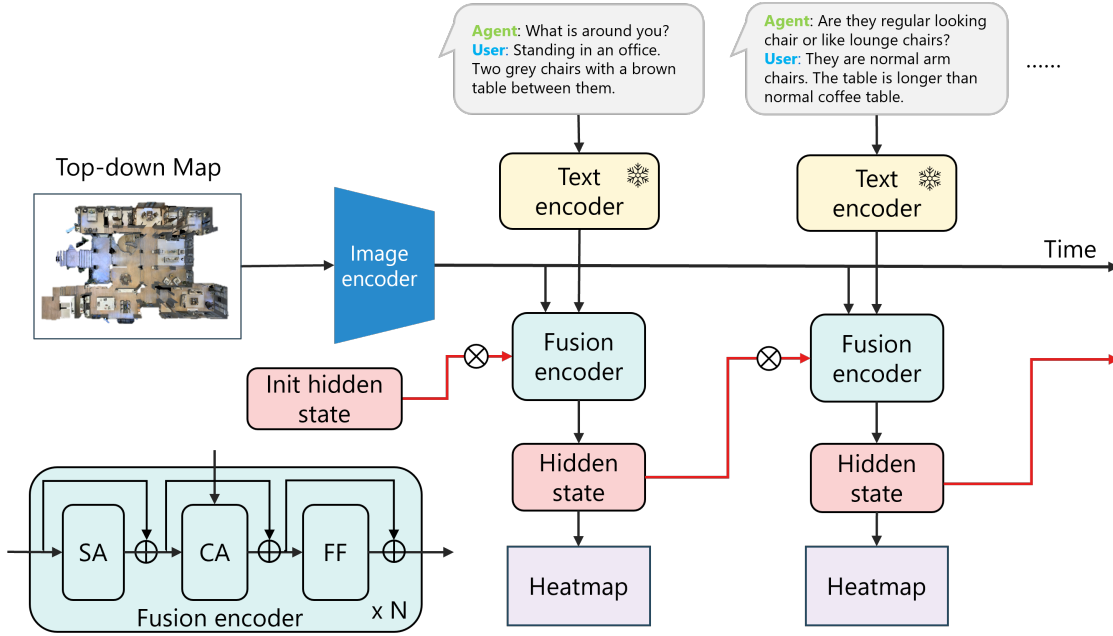
Figure 2. **DiaLoc-e: the proposed multi-shot multimodal architecture for embodied dialog localization.** The approach employs the image encoder and the frozen text encoder to derive visual and linguistic unimodal embeddings. The fusion encoder integrates the unimodal inputs to update the hidden state. Multi-shot predictions are produced using the hidden state at varying timesteps. The fusion encoder comprises N blocks of Transformer encoder with cross attention layer.

Multimodal Encoder. Given visual and text embeddings V and L, we employ a stack of Transformer blocks as the multimodal encoder Φ. Each block consists of 12 sequential modules of self-attention (SA), cross-attention (CA) and feed-forward (FF) layers. The cross-attention layer is introduced for the purpose of multimodal fusion. Specifically, text embedding of current turn L_t is used to create query Q via a MLP, while the map embedding V provides key K and value V. The output of multimodal encoder is hidden state $S \in R^{M,C}$ and has the same dimension as visual embedding V. A detailed illustration of the fusion encoder is shown in Figure 2.

Prediction Head. Multimodal encoder fuses dialog input and visual map representations to output the current state S_t at turn t. Using S as input, a prediction head including multiple convolution and deconvolution layers is trained to produce location heatmap $H \in R^{h,w}$. The heatmap H is finally up-sampled to the same size as target $\tilde{H} \in R^{h_0,w_0}$ for loss caculation. At inference stage, the predicted location in image space is obtained via $\hat{p} = \text{Argmax}(\text{Softmax}(H))$.

Multi-shot Multimodal Localizer. We now discuss two fusion variants for multi-shot localization. The distinction mainly lies in how previous hidden state S_{t-1} are intergrated with new dialog input L_t. In Figure 2, the *explicit* variant, dubbed **DiaLoc-e**, is illustrated. The map embedding V is directly processed at each timestep and is fused
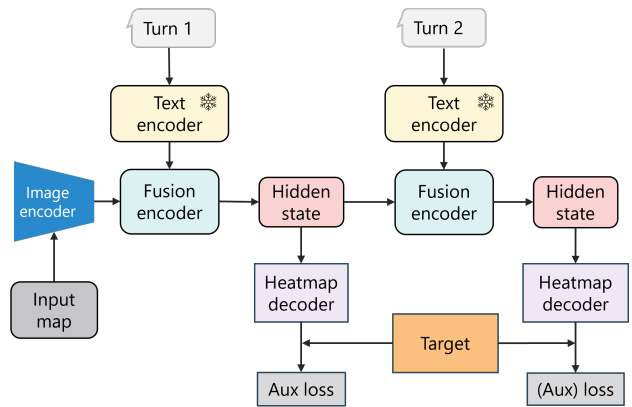

Figure 3. **DiaLoc-i: the proposed localizer variant with implicit fusion design.** The hidden state is continuously updated with dialog information at each timestep.

with the hidden state S before being fed into the fusion encoder via: $S_t = \Phi(V \odot S_{t-1}, L_t)$. The hidden state S is initialized to $\mathbf{1}$ and serves as the prior for the visual map. For the *implicit* variant, **DiaLoc-i**, depicted in Figure 3, it iteratively updates the hidden state via $S_t = \Phi(S_{t-1}, L_t)$ thus ensuring that the final state S_T are conditioned on all dialog turns $(L_1, .., L_T)$.

3.3. Loss Function

Starting with single-shot case, given the prediction H and the target $\tilde{H}$, we train the model to minimize the KL-divergence between the predicted location distribution and the ground-truth location. Similar to [9], Gaussian smoothing with standard deviation of 3m is applied to the ground-truth $\tilde{H}$. The single-shot loss function is:

$$L_{ss}(H, \tilde{H}) = \log(\tilde{H})(\log \tilde{H} - \log(\text{Softmax}(H))) \quad (1)$$

To adapt to our multi-shot approach, we simply apply single-shot loss to all predictions $\mathbf{H} = H_{t=1}^T$ made at $t \in [1, T]$. Additionally, the multiple losses are weighted summed to provide the following multi-shot loss:

$$L_{ms}(\mathbf{H}, \tilde{H}) = \frac{1}{T} \sum_{t=1}^{T} L_{ss}(H_t, \tilde{H}) \alpha^{T-t} \quad (2)$$

where $\alpha \in [0, 1]$ is the decay factor. The idea of applying decay is to penalize less for early predictions due to incomplete dialog context.

However, there is still a risk of over-fitting when intermediate predictions H_t where $t < T$ are toned down with the decay factor. Specifically we note that it is often unreasonable to find the true location after a single, ambiguous dialog turn. Unfortunately, however, The KL-divergence loss with the ground-truth target encourages peaky predictions and leads to over-confident prediction at early timesteps. To alleviate this issue, we introduce an auxiliary loss L_{aux}. The aim of this term is to promote the diversity of early predictions and avoid unreasonable early filtering of false positives. The loss function is defined as:

$$L_{aux}(\mathbf{H}, \tilde{H}) = \frac{1}{T} \sum_{t=1}^{T} L_{mse}(\text{Sigmoid}(H_t) \odot \tilde{H}_{mask}, \tilde{H}) \quad (3)$$

where $\tilde{H}_{mask}$ is the binary mask of target where $\tilde{H} > 0$. Sigmoid function is applied to H_t to produce the probability scores of location. Mean squared loss L_{mse} is used to measure the prediction error.

We combine the multi-shot loss L_{ms} and the auxiliary loss L_{aux} to get the final loss:

$$L(\mathbf{H}, \tilde{H}) = L_{ms}(\mathbf{H}, \tilde{H}) + \beta L_{aux}(\mathbf{H}, \tilde{H}) \quad (4)$$

where β as the weight factor. We set $\beta = 1.0$ for our method except ablation studies.

3.4. Training Details

In training, both top-down map and ground-truth target are resized to 224×224. Color jittering, random cropping with ratio $[0, 9, 1.0]$ and scale $[0, 75, 1.0]$, and random rotation of $180°$ are used for data augmentation. ViT-base pretrained on ImageNet-21k and Bert-base-uncased are used as unimodal encoders. We keep BERT encoder frozen in training. We train all models up to 30 epochs, with batch size 16. We use AdamW [16] as the optimizer and set the learning rate to $2e - 5$.

4. Experiments

First, we conduct several ablation studies aimed at validating the design choices, including the depth of multimodal encoders, fusion variants, and the effect of the decay factor. Additionally, we delve into the evaluation of the auxiliary loss to support its efficacy. From the perspective of data augmentation, we observe that incorporating synthetic dialogs generated by ChatGPT API [1] contributes to performance enhancement. Second, we proceed to compare our proposed methods with the baseline approaches in both single-shot and multi-shot modes. This include quantitative results as well as qualitative samples. Lastly, we provide a detailed examination of multi-shot performance, offering a fine-grained analysis of the results.

4.1. Ablations

Depth of multimodal encoder. We first study the effect of depth of the multimodal encoder in Table 1. We ablate this with the implicit fusion variant as shown in Figure 2. Results show that depth=3 gives the best performance in unseen set when vision encoder is frozen. When ViT is fine-tuned, the performance using more blocks drops. The best unseen Acc5 is 47.09 using depth=1. Overall, fine-tuning ViT improves the performance a lot because ViT needs to adapt to the top-down map visual input.

ViT	Depth	valSeen		valUnseen	
		LE↓	Acc5↑	LE↓	Acc5↑
Frozen	1	9.29	47.18	10.54	33.27
	3	9.45	43.43	9.26	38.23
	6	9.41	47.50	9.54	37.72
Fine-tune	1	**7.62**	57.37	**9.09**	**47.09**
	3	7.73	**58.01**	9.35	42.06
	6	8.44	55.12	10.02	39.61

Table 1. **Ablation on depth of multimodal encoders.** DiaLoc-i is used in this ablation with decay $\alpha = 0$.

Fusion variants and decay factor. We have proposed two variants to perform multimodal fusion in Section 3.2. DiaLoc-i updates the visual hidden state using dialog input recursively, while DiaLoc-e fuses the original map embedding with the hidden state using multiply fusion operation. To investigate how the two fusion variants perform, we report the results in Table 2. For both variants (depth=3),

we fine-tune the ViT. The results show that implicit fusion performs better on valSeen set, while explicit fusion works better on valUnseen. Implicit fusion leverages shared hidden state and has higher risk of over-fitting. Using fixed map embedding and learnable hidden state, explicit fusion shows better generalization. We also find that using non-zero decay factor leads to lower accuracy due to the early convergence.

Method	Decay	valSeen		valUnseen	
		LE↓	Acc5↑	LE↓	Acc5↑
DiaLoc-i	0.0	**7.73**	**58.01**	9.35	**42.06**
	0.5	7.95	54.68	9.74	36.48
	1.0	8.20	55.31	9.75	33.61
DiaLoc-e	0.0	8.99	50.32	**9.10**	41.60
	0.5	9.51	54.68	9.76	41.10
	1.0	8.91	51.56	9.21	37.95

Table 2. **Ablation on fusion schemes and the decay factor.** Both DiaLoc variants use depth=3 without auxiliary loss.

Impact of auxiliary loss. We ablate the impact of using auxiliary loss as defined in Eq 3.3. We train the model with depth as 1 with varying decay factor of KL-divergence term. For each decay choice, the same model is trained with (β=1) and without (β=0) using auxiliary loss. Results are shown in Table 3, combining the two losses helps to improve the performance consistently across all setups.

Decay α	Aux β	DiaLoc-e				DiaLoc-i			
		valSeen		valUnseen		valSeen		valUnseen	
		LE↓	Acc5↑	LE↓	Acc5↑	LE↓	Acc5↑	LE↓	Acc5↑
0.0	0.0	7.81	57.18	9.99	37.89	7.54	58.65	9.61	34.19
	1.0	7.76	58.01	9.56	40.23	7.39	60.57	9.19	38.58
0.5	0.0	7.95	54.68	9.74	36.48	7.16	58.97	9.98	36.41
	1.0	6.76	63.46	9.58	40.01	7.79	60.25	10.54	32.53
1.0	0.0	8.20	55.31	9.75	33.61	7.99	56.41	11.01	34.36
	1.0	7.92	58.65	8.99	42.06	7.55	59.29	10.42	36.47

Table 3. **Ablation on auxiliary loss using explicit fusion variant DiaLoc-e.** We vary the decay factor α to train multiple versions of the model, with and without the proposed auxiliary loss. DiaLoc-e generalizes (9/12 valUnseen) better than DiaLoc-i (10/12 valSeen)

Dialog augmentation using LLM. We hypothesize that increasing the dialog diversity during training could improve the performance because text encoder is frozen in our work. We leverage GPT to study the effect of dialog augmentation on the task. For each training sample, we prompt GPT API to paraphrase the ground-truth dialog. During training, either the GPT version or the ground-truth dialog is chosen randomly. Results in Table 4 show that using extra data helps improve the performance for seen and unseen maps.

Method	Dialog	valSeen		valUnseen	
		LE↓	Acc5↑	LE↓	Acc5↑
DiaLoc-e	GT	8.91	51.56	9.21	37.95
	GT+GPT	**7.07**	**60.00**	**9.09**	**40.71**

Table 4. **Ablation on using extra dialogs generated by GPT.** DiaLoc-e is configured to use depth=3, $\alpha = 1$, and $\beta = 0$

Mode	Method	valSeen		valUnseen	
		Acc0	Acc5	Acc0	Acc5
Single shot	LingUNet	19.87	59.29	6.16	33.33
	BLIP-Res18◇	13.44	50.82	6.73	34.71
	BLIP(DiaLoc)†	**25.64**	**66.02**	7.02	40.41
Multi shot	LingUNet-i	4.32	24.16	3.25	20.21
	LingUNet-e	14.47	46.15	5.31	36.30
	DiaLoc-i	18.43	57.18	6.42	37.89
	DiaLoc-e	18.36	60.00	**8.44**	**47.15**

Table 5. **Evaluations on WAY dataset under single-shot and multi-shot scenarios.** We configure DiaLoc-i and DiaLoc-e to use $d = 3, \alpha = 0, \beta = 1$. BLIP-Res18◇ and BLIP(DiaLoc)† are adapted so that the visual branch is configured as the main backbone to predict heatmap.

4.2. Comparison to the SOTA

Single-shot and multi-shot. In this section, we undertake a comparative analysis of our method with state-of-the-art approaches in the LED task. For single-shot mode, LingUNet [9] using complete dialog stands as the current SOTA method. We report the geodesic distance performance using the author-provided code. We also adapt BLIP [14] as BLIP-Res18 and BLIP(DiaLoc) to use dialogs as conditional input for image-based localization task.

In the multi-shot mode, we adapt LingUNet similarly to our DiaLoc-i and DiaLoc-e, enabling it to generate multi-shot predictions Specifically, the fusion is implemented by multiplying the output (H1 preceding the final MLP) with image features (F1). More details in Supp.Mat. We refer to this method as LingUNet-i/e. Our proposed method is denoted as DiaLoc-i and DiaLoc-e, reflecting the different fusion variants implicit and explicit respectively. In alignment with [9], we report Acc0 and Acc5 as evaluation metrics for both valSeen and valUnseen sets in Table 5.

When utilizing the entire dialog both during training and inference, our method outperforms LingUNet by a substantial margin. The improvement in Acc5 on the valUnseen is 7.08. This performance enhancement can be attributed to the learning capabilities and flexibility of Transformers for both unimodal and multimodal encoders.

In the context of the proposed multi-shot localization task, our evaluation focuses on final prediction accuracy. First, in comparison to the mode using complete dialogs, the performance of all methods exhibits a decline in the
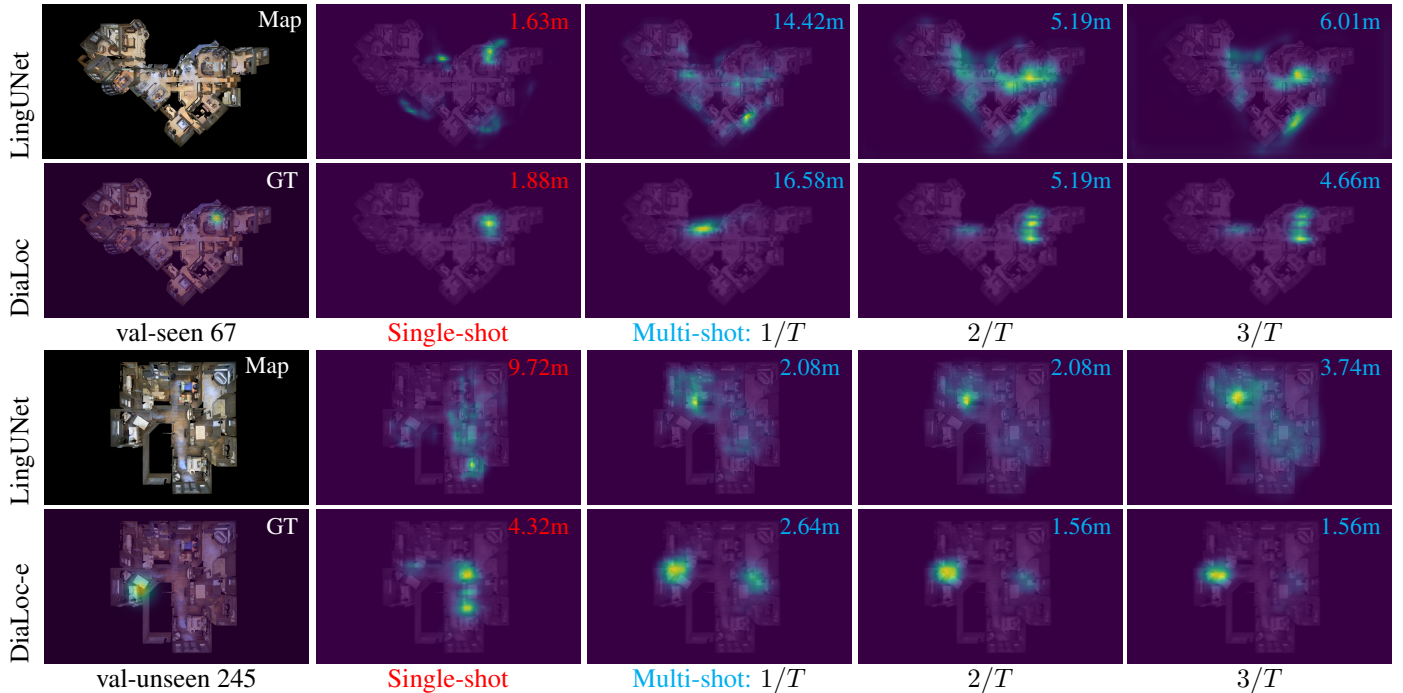
Figure 4. **Qualitative results of single-shot and multi-shot location predictions are presented.** In the first column, the visual map is displayed alongside its corresponding ground truth (GT) location. The second column displays the single-shot predictions, with LingUNet results above and DiaLoc results below. The last three columns showcase the multi-shot predictions. Regarding the results of valseen case 67, DiaLoc effectively corrects the prediction after the second turn, whereas LingUNet-ms produces noisy distributions. For valUnseen case 245, both LingUNet and DiaLoc failed in the single-shot mode. Nevertheless, in the multi-shot mode, DiaLoc succeeds in refining the prediction. In contrast, LingUNet-ms converges towards an incorrect area. Localization Error (LE) is displayed for the predictions.
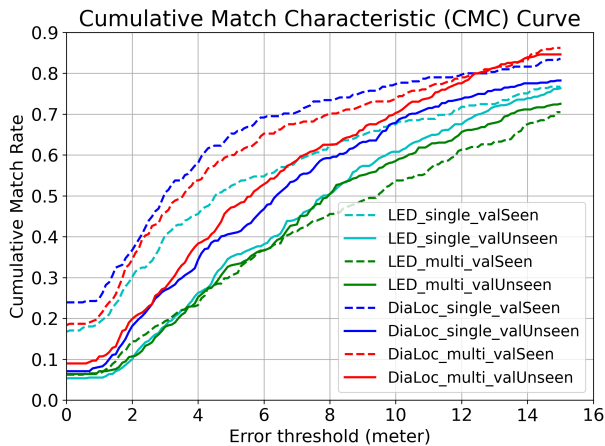

Figure 5. **CMC curves on WAY dataset.** We depict the CMC curves for both our DiaLoc and LingUNet for single-shot and multi-shot settings. DiaLoc consistently outperforms the baseline. X-axis denotes the error threshold for LE and the Y-axis denotes the success rate.

valSeen set. Our methods demonstrate competitive performance in valSeen set. Notably, DiaLoc-e surpasses all other methods in the valUnseen set, demonstrating robust gener-

alization and reasoning capabilities. It is also noteworthy that LingUNet-ms performs well in the valUnseen set, even though its performance drops in valSeen. This observation suggests that the proposed iterative multi-shot approach effectively reduces the issue of over-fitting.

Considering computational complexity, there are significant savings when switching from the one-shot to the multi-shot scenario, despite multi-shot demanding multiple forward passes. Now, we compare the computational costs during the dialog embedding stage with Bert. For a dialog with T turns, where each turn is encoded by N tokens, the single-shot mode involves $O(T^2 N^2)$ multiplications for self-attention computation. However, this complexity is reduced to $O(N^2)$ in the multi-shot mode.

CMC Curves. To conduct a more comprehensive assessment of the performance of proposed method and the state-of-the-art method, We present the Cumulative Match Characteristic (CMC) curves for both DiaLoc and LingUNet(ms) in Figure 5. DiaLoc consistently outperforms the baseline in both single-shot and multi-shot configurations. In val-Unseen set, our multi-shot performance (red solid line) also surpasses that of the single-shot baseline (cyan solid line) and single-shot DiaLoc (blue solid line), showing improved
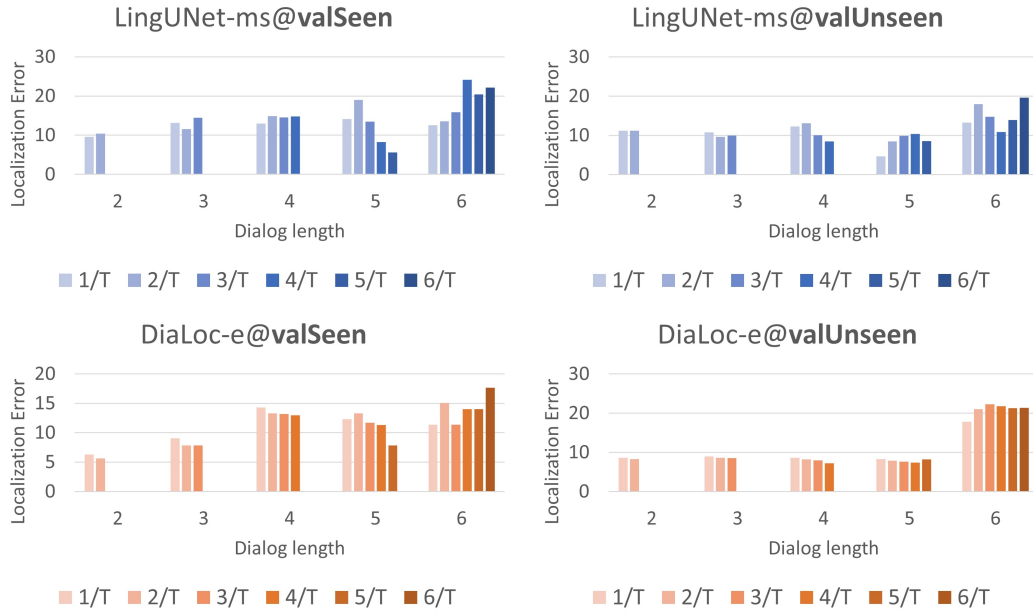
LingUNet-ms@valSeen — Localization Error vs Dialog length (2, 3, 4, 5, 6); legend: 1/T 2/T 3/T 4/T 5/T 6/T

LingUNet-ms@valUnseen — Localization Error vs Dialog length (2, 3, 4, 5, 6); legend: 1/T 2/T 3/T 4/T 5/T 6/T

DiaLoc-e@valSeen — Localization Error vs Dialog length (2, 3, 4, 5, 6); legend: 1/T 2/T 3/T 4/T 5/T 6/T

DiaLoc-e@valUnseen — Localization Error vs Dialog length (2, 3, 4, 5, 6); legend: 1/T 2/T 3/T 4/T 5/T 6/T

Figure 6. **Multi-shot localization error analysis.** To study performance across varied dialog length, we group samples based on their length T and report LE for each group. Within each sub-plot, the LE at t where $1 \leq t \leq T$ is detailed. Our method depicts a trend that more turns is helpful to localization while LingUNet shows mixed performance.

performance in novel environments.

Qualitative results. In Figure 4, we show two examples along with the corresponding predictions from the valSeen and valUnseen splits of WAY dataset comparing LingUNet with DiaLoc. Our single-shot method yields more precise and concentrated predictions. Regarding the multi-shot results, DiaLoc showcases the capability to rectify its previous predictions by leveraging the most recent dialog information as guidance.

4.3. Multi-shot Predictions Analysis

As one of the most attractive aspects, employing multi-shot localization holds the potential to early terminate the dialog in real-world searching and rescue applications. In this section, we analyze the performance of multi-shot methods using dialog up to timestep t. There exist several ways to determine the predicted locations. Examples encompass hard argmax and soft threshholding. Similar to single-shot evaluation, we use argmax to identify the top-1 location based on the predicted heatmap. Alternatively, soft threshholding identifying a set of top-K locations with probabilities surpassing the threshold. Subsequently, evaluation metrics such as Recall and Precision can be employed for measuring the performance. For simplicity, we adopt argmax as a straightforward way to evaluate the interim predictions, presenting the results in Figure 6. This will be unfair if multiple peaks show up, and one is true. It is worth looking at the prediction probability at ground-truth pixel.

To comprehend performance across various dialog length T, we group samples from valSeen and valUnseen based on their respective dialog length. Each plot showcases the mean localization error (LE) for different sample groups. In each sub-plot, the LE at t/T where $1 \leq t \leq T$ depicts the performance using incomplete dialog. Our method consistently demonstrates a trend wherein using more turns leads to decreased LE across valSeen and valUnseen split, while LingUNet exhibits varied behavior. Another notable observation is that both methods exhibit lower performance in the case of lengthier dialogs, particularly struggling with dialogs comprising $T = 6$ turns. Introducing new turns does not yield performance improvements, and is possibly due to the unbalanced training data.

5. Conclusion

Localization using dialog is an important task in the field of Embodied AI and significant progress has been achieved thanks to the advancements in vision and language learning. In this work, we propose a novel approach DiaLoc for iterative dialog-based localization. The key idea is to overcome the limitations of existing methods that requires entire dialog for localization. We introduce DiaLoc, a multimodal localizer, to continually fuse visual and dialog inputs for accurate location prediction. Our work narrows the gap between simulation and real-world applications, opening doors for future work on multimodal dialog generation and collaborative localization.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] 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. 2

[3] Shurjo Banerjee, Jesse Thomason, and Jason Corso. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, pages 1384–1393. PMLR, 2021. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[5] 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. *arXiv preprint arXiv:1709.06158*, 2017. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[7] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palme: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2

[8] Meera Hahn and James M Rehg. Transformer-based localization from embodied dialog with large-scale pre-training. *arXiv preprint arXiv:2210.04864*, 2022. 2, 3

[9] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James M Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. *arXiv preprint arXiv:2011.08277*, 2020. 1, 2, 3, 5, 6

[10] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. 2

[11] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6687–6696, 2022. 3

[12] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[13] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 3, 6

[15] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[17] Melvin P Manuel, Mariam Faied, Mohan Krishnan, and Mark Paulik. Robot platooning strategy for search and rescue operations. *Intelligent Service Robotics*, pages 1–12, 2022. 1

[18] Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping instructions to actions in 3d environments with visual goal prediction. *arXiv preprint arXiv:1809.00786*, 2018. 2

[19] Jorge Pena Queralta, Jussi Taipalmaa, Bilge Can Pullinen, Victor Kathan Sarker, Tuan Nguyen Gia, Hannu Tenhunen, Moncef Gabbouj, Jenni Raitoharju, and Tomi Westerlund. Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision. *Ieee Access*, 8:191617–191643, 2020. 1

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

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

[22] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[24] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 2