Supplementary Materials:
DiaLoc: An Iterative Approach to Embodied Dialog Localization

Chao Zhang Mohan Li Ignas Budvytis Stephan Liwicki
Toshiba Europe Ltd

1. Supplementary Materials

Multi-shot Adaptation of LingUNet. LingUNet was designed for single-shot dialog localization. To enable it works for multi-shot scenario so that fair evaluation is possible under our iterative formulation, we make optimal modifications according to our proposed DiaLoc-e. Basically, the idea is to leverage hidden states of previous iteration for future predictions. The adapted LingUNet-ms is illustrated in Figure 1. Similar to our DiaLoc-e, the map embedding F1 generated by ResNet18 is fed to LingUNet alongside the dialog embedding to generate hidden states H1. The hidden states is then used as input to predict location heatmaps. In particular, at each timestep t, the hidden states of previous timestep H1_{t-1} is fused with F1 to integrate the dynamic prior information.

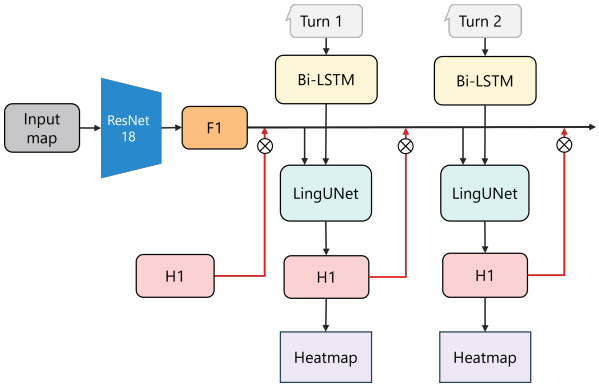

Figure 1. **LingUNet-ms: the adapted multi-shot multimodal LingUNet for iterative embodied dialog localization.**

Dialog Augmentation using LLM. In this section, we provide additional details for leveraging LLM to augment localization dialogs in WAY dataset. We employ gpt-3.5-turbo-16k as the LLM instance to rewrite the ground-truth dialogs of training set. We use the prompt as " Paraphrase the dialog". We set the temperature to 0.6 and the top-p to 0.5 in the API call.

In Figure 2, we show two examples from the train split of WAY. For each example, we display the top-down map and the corresponding target on the left. On the right size, the GT dialog is shown at the top within the blue box, and the para-phased version is shown inside the orange box. In both cases, we can see that GPT generates semantically consistent dialogs as the original version. In the second case, GPT reduced the length of the original dialog without changing the meaning. Note that, the GPT API does not use map information at all and is purely text-based.

Method	Image Size (height x width)	Runtime (second)	GPU Usage (MiB)
LingUNet-ms	455 x 780	0.768	1,273
DiaLoc-e	224 x 224	1.374	3,875

Table 1. **Average runtime and memory usage comparison at inference time.** Batch size is set to 1 for both methods. A Nvidia Titan RTX 24gb is used for the benchmarking.

Multi-shot analysis via prediction confidence. 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. In addition to the localization error based on top-1 prediction, we employ prediction confidence as an alternative in this analysis. Localization error as a metric will be unfair in case of multiple peaks show up, and one is true positive. To overcome this issue, given heatmap prediction, we report the pixel-wise probability at ground-truth location in Figure 3. In summary, our method depicts a trend that more turns is helpful to increase the confidence level at the desired location while LingUNet shows a negative trend.

Inference runtime and memory usage. One of the limitations of the proposed approach is the memory usage and decreased inference efficiency. We evaluate DiaLoc-e (depth=1) in the multi-shot mode on the valSeen split of WAY. The evaluation results are shown in Table 1.

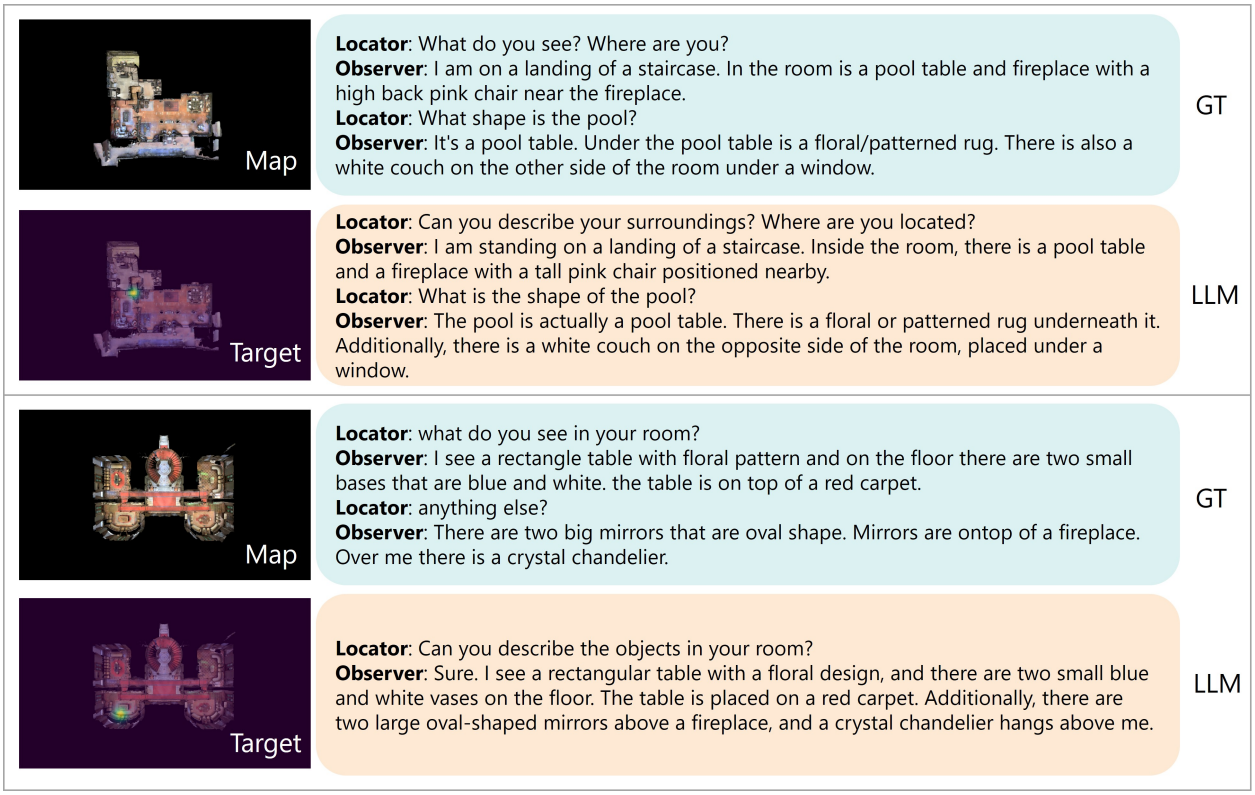

Figure 2. **Examples of ground-truth dialogs and augmented version using LLM**.

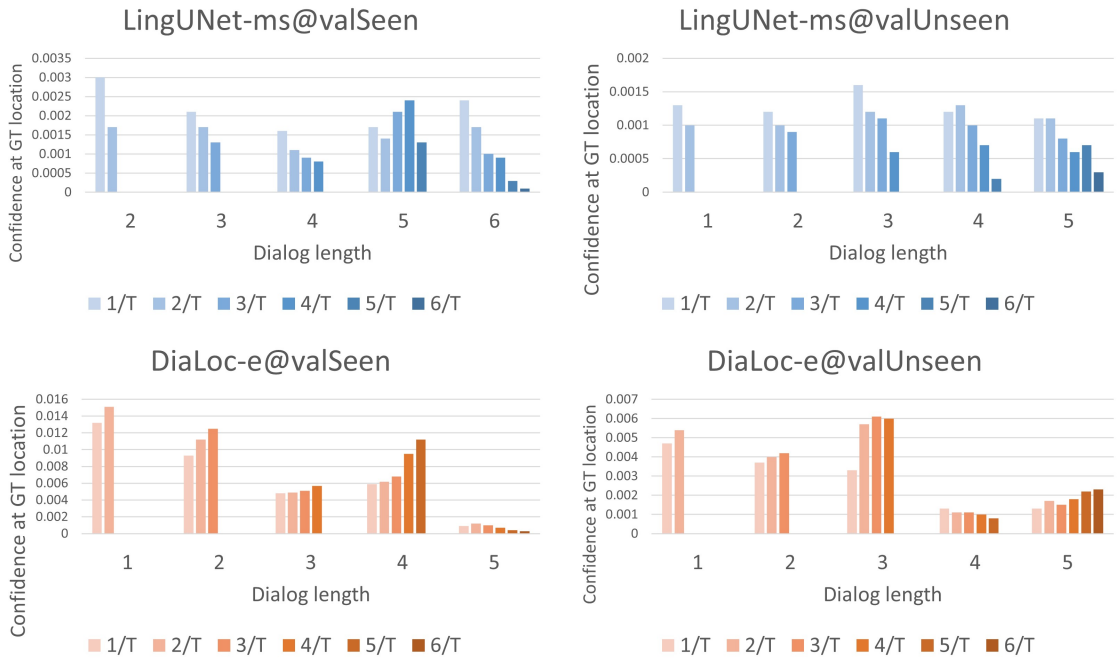

Figure 3. **Multi-shot prediction confidence analysis.** To study performance across varied dialog length, we group samples based on their length T and report average prediction confidence for each group. Within each sub-plot, the PC at t where $1 \leq t \leq T$ is detailed. Our method depicts a trend that more turns is helpful to increase the confidence at the desired location while LingUNet shows negative trend.

Additional visualizations. In Fig.4, we visualize additional localization predictions comparing LingUNet and the proposed method. For both methods, single-shot and multi-shot variants are evaluated. We now share our insights from these representative examples.

- **Val-seen 87:** DiaLoc predicts the correct location while LingUNet failed in the single-shot mode. LingUNet-ms produces noisy but correct predictions. In contrast, DiaLoc is capable of generating concentrated multi-modal (not to be confused with *multimodal learning*) predictions.
- **Val-seen 176:** Both approaches give acceptable predictions in the single-shot mode. In multi-shot mode, DiaLoc recovers from its initial incorrect prediction, out of two possible guesses.
- **Val-unseen 224:** DiaLoc succeeds while LingUNet fails in the single-shot. In multi-shot mode, LingUNet-ms generates noisy predictions, while DiaLoc continually refining its prediction and converging towards the correct location in the end.
- **Val-unseen 327:** In the single-shot mode, both methods failed. In the multi-shot mode, LingUNet-ms converged to a few locations, but none matches the GT. For our DiaLoc, the predictions are incrementally refined to the right area.

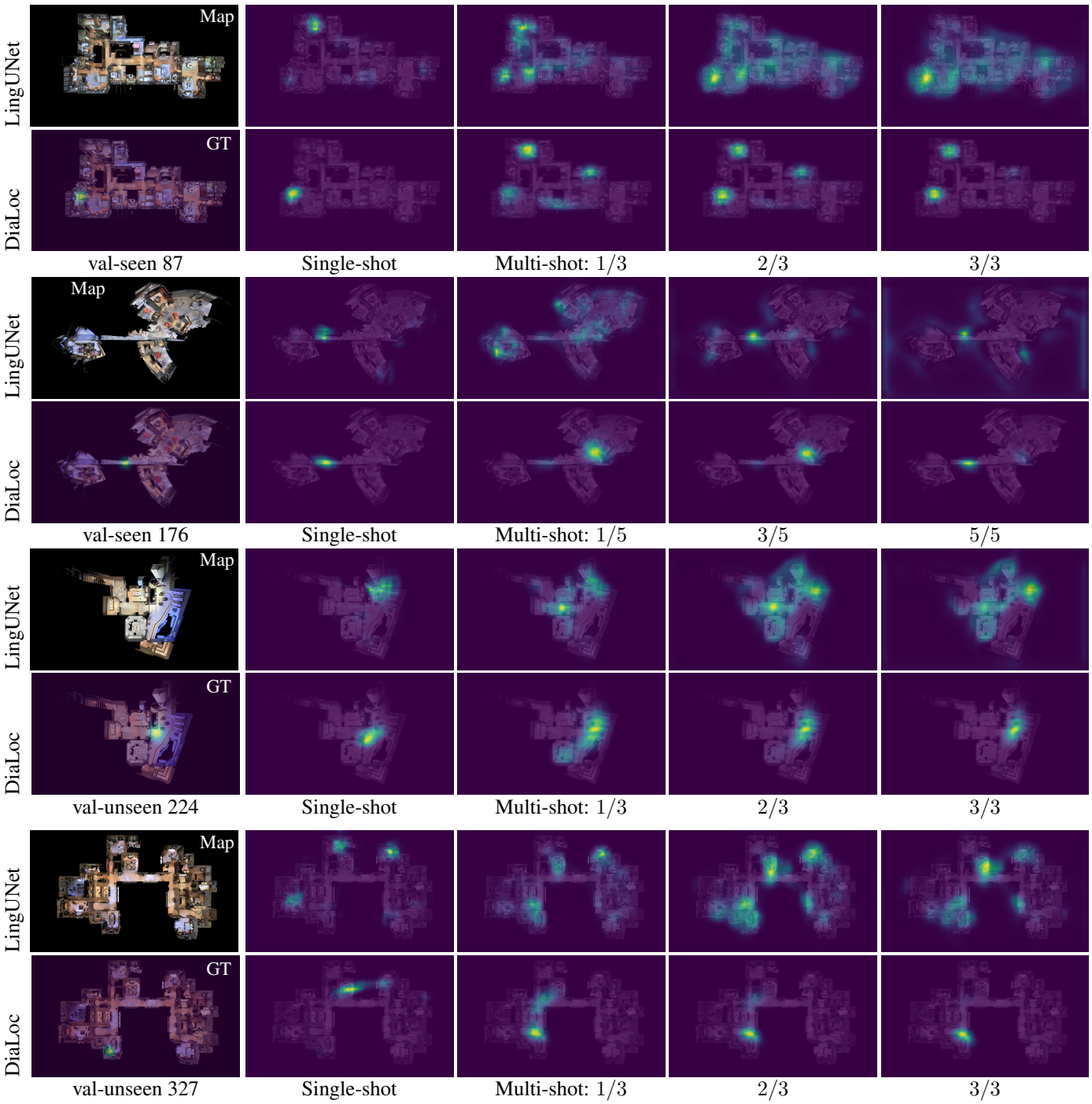

Figure 4. **Qualitative results of single-shot and multi-shot location predictions are presented.** In the first column, the top-down map is displayed alongside its corresponding ground truth (GT) location. The second column displays the single-shot predictions, with LingUNet results above and DiaLoc results below. The last three columns showcase the multi-shot predictions.