Supplementary Material

A. More details of HM3D-DUET

HM3D-DUET [1] is a dual-scale graph transformer with topological maps, which contains two modules: topological mapping and global action planning. The topological mapping module builds a topological map during navigation. And the global action planning module predicts the next location on the map or a stop action to end the navigation.

A.1. Topological Mapping

To build the environment graph $\mathcal{G}$ which is unknown initially, the mapping module updates node representations by adding the newly observed location gradually to the map. Specifically, the map denote as $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$. At time step t, the current node V_t and its neighboring unvisited nodes $\mathcal{N}(V_t)$ are added to $\mathcal{V}_{t-1}$.

The mapping module outputs the current panorama encoding with image features $\{r_i\}_{i=1}^n$ and object features $\{o_i\}_{i=1}^m$ and a graph with K node features $\{v_i\}_{i=1}^K$.

A.2. Global Action Planning

Dual-scale Cross-modal Encoder The module uses dual-scale architecture transformers to capture cross-modal vision-and-language relations from different scales: a fine-scale representation of the current location and a coarse-scale representation of the map.

In the coarse-scale cross-modal encoder, the inputs are map node features $\{v_i\}_{i=1}^K$ and textual features $\mathcal{T}$. The node features are embedded and combined with word embeddings into a multi-layer graph-aware cross-modal transformer to get node embedding $\widehat{v}_i$. Then the node embedding $\widehat{v}_i$ is fed into a two-layer feed-forward network (FFN) to predict a navigation score for each node s_i^c.

In the fine-scale cross-modal encoder, the inputs are fine-grained visual representations $\{\mathcal{R}_t, \mathcal{O}_t\}$, the textual features $\mathcal{T}$, and a special stop token r_0. Then the concatenated visual tokens $[r_0; \mathcal{R}_t; \mathcal{O}_t]$ and textual features $\mathcal{T}$ are fed into a standard multi-layer cross-modal transformer to get $[\widehat{r_0}; \widehat{R}_t; \widehat{O}_t]$. The navigation score for local-level s_i^f and object are predicted via FFN, a similar way in the coarse-scale cross-modal encoder.

Finally, the coarse-scale prediction s_i^c and fine-scale prediction s_i^f are dynamically fused to obtain the final navigation prediction s_i.

Algorithm 1 March in Chat

Notation Summary:
I: high-level instruction in REVERIE
$\mathcal{R}_t$: visual observation at timestep t
$\mathcal{D}$: demonstration set
$P_\circ$: prompt for LLM to generate planning
LLM: large language model
Template: templates to generate natural language description

```
t ← 0                                              ▷ Initial timestep
W_I ← I
ô ← LLM(I, P_o)                                    ▷ Target object recognition
l̂ ← LLM(ô, P_l)                                    ▷ Target location reasoning
W_G ← Template(ô, l̂)                               ▷ GOSP
W_S ← ϕ
W ← Concat(W_I, W_G, W_S)                          ▷ Assembled instruction
P_demon ← DynamicSelect(I, D)                      ▷ Dynamic demonstration
while t <max-step and â_t ≠ "stop" do
    ĉ^t_room, ĉ^t_obj ← CLIP(R_t)                  ▷ ROASP
    if ĉ^t_room ≠ ĉ^{t-1}_room then
        P_scene ← Template(ĉ^t_room, ĉ^t_obj)
        P_step ← Template(I, W_S)
        P_SODP ← Concat(P_scene, P_demon, P_step)
        I_step ← LLM(P_SODP)                       ▷ SODP
        W_S.Append(I_step)
        W.Update(W_S)                              ▷ Instruction update
    end if
    â_t ← Agent(W, R_t)
    t ← t + 1
end while
```

B. Algorithm of MiC

The algorithm of MiC pipeline is described in Algorithm 1.

C. More Qualitative Results

We provide more examples of generated instructions in Fig. 1. And Fig. 2 shows the screenshot of the user interfaces for human study.

D. Discussion of HANNA

HANNA [2] defines a task in an interactive setup. It builds a simulator to let the agent requests assistance for

Scene-Oriented Dynamic Planning	w/o Dynamic Demonstration	w/o ROASP
REVERIE Instruction: Rearrange the sofa in the office.		
Step 1: go down the stair in front of you Step 2: once at the bottom enter the door Step 3: stop once you enter the room Step 4: rotate the sofa in the office	Step 1: select the object you want to rearrange.	Step 1: go up the stairs and into the office Step 2: when you get there, turn left Step 3: when you get back to the top of the stairs
REVERIE Instruction: Dust off the picture in the lobby		
Step 1: go into the lobby Step 2: find the target picture Step 3: dry the back of the picture	Step 1: go to the lobby and click on the picture	Step 1 go to the stairs and up them Step 2: go through the doorway on the right side Step 3: go back to the staircase Step 4: go through the doorway on the left side
REVERIE Instruction: Open the bathroom window.		
Step 1: enter the bathroom through the open door Step 2: walk straight Step 3: stand next to the toilet	Step 1: go into the bathroom and open the door	Step 1: go inside the bathroom Step 2: walk towards the mirror Step 3: look at yourself in the mirror Step 4: then close the door Step 5: walk back outside
REVERIE Instruction: Go to the family room and clean the end table.		
Step 1: walk to the family room Step 2: walk to the end table Step 3: wait right by table.	Step 1: clean the end table with a damp cloth Step 2: wipe down the floor with a damp cloth	Step 1: walk towards the fireplace Step 2: and then go straight into the room on the right Step 3: wait near the table and chair Step 4: go back to the kitchen Step 5: then go to the dining room

Figure 1: Examples of generated instructions.

navigation. However, the agent just sends a signal for help when it got lost, without environmental feedback. What's more, the assistance provided by HANNA is ground truth information, which is not applicable in real-world scenarios. While in MiC, the agent could give environmental feedback in natural language, and the off-the-shelf LLM planner would correspondingly generate step-by-step instructions.

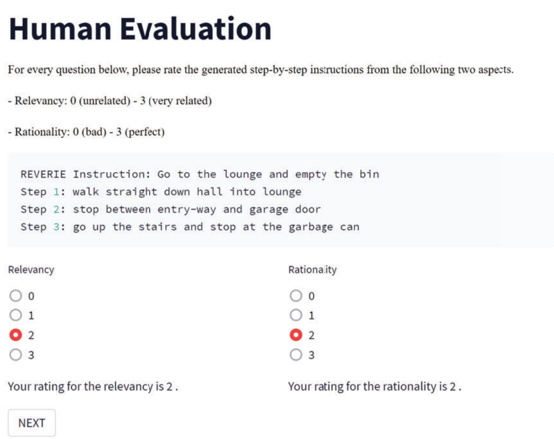

Figure 2: Screenshot of the user interfaces for our human study.

References

[1] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *ECCV*, 2022. 1

[2] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. pages 684–695. Association for Computational Linguistics, 2019. 1