

DialNav: Multi-turn Dialog Navigation with a Remote Guide

Supplementary Material

H. Data Collection Interface

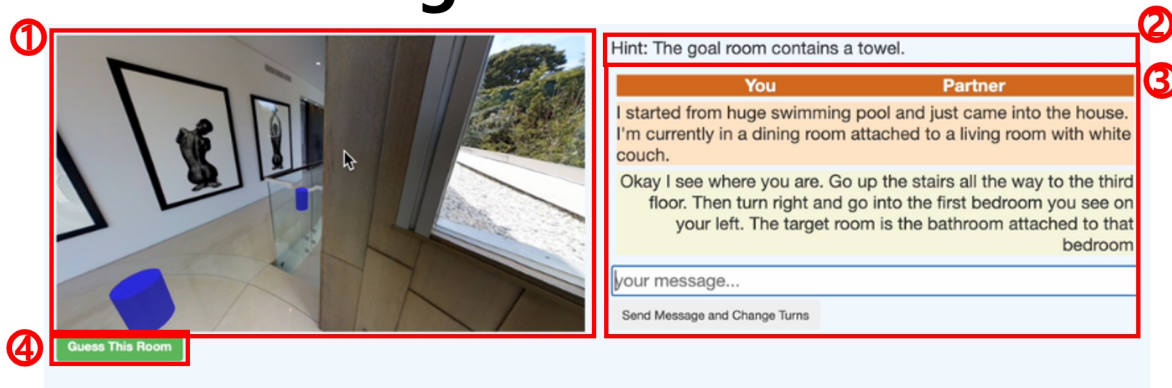
Fig. G illustrates the data collection interface for Navigator and Guide.

Navigator Interface Navigator moves toward the goal region given the initial instruction. When unsure, Navigator can ask questions to receive additional information from the

remote Guide. When Navigator asks a question, Navigator's interface is deactivated until Guide's response comes. Navigator interface consists of four main components:

- **Navigation Interface:** Allows free movement between nodes, active only during the Navigator's turn.
- **Hint:** Provides hints about objects in the goal room, formatted as "*The goal room contains {object}.*"
- **Chat Interface:** Enables the Navigator to send questions

Navigator Interface



Guide Interface

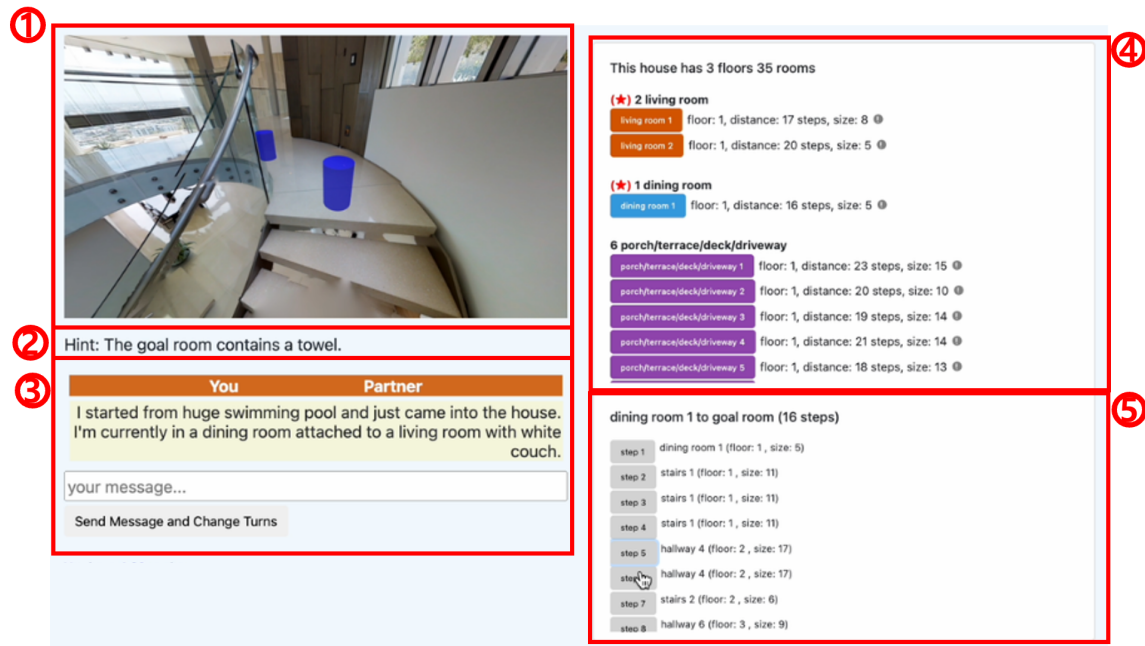


Figure G. Navigator (top) and Guide (bottom) interface

and receive responses from the Guide.

- **Guess Button:** Allows the Navigator to indicate they have reached the goal. If correct, the task ends; if incorrect, a popup indicates the mistake, and the task continues. This button can be pressed multiple times without changing turns.

Guide Interface Guide acts as a knowledgeable participant regarding the house environment. We designed a specialized Guide interface that simulates Guide to be familiar with the entire space. For the question of whether Guide interface provided sufficient support for the task, the annotators rated 4.5 out of 5 on average. Guide interface consists of five components:

- **Navigation Interface:** Allows free movement between nodes, even when it is not Guide’s turn.
- **Hint:** The same hint provided to Navigator is also shared with Guide.
- **Chat Interface:** Enables Guide to read the Navigator’s questions and send responses.
- **House Info Interface:** Offers detailed information about the house, including:
 - **Overall Information:** Presented as “*This house has N floors and N rooms.*”
 - **Room List:** Displays information about all rooms grouped by type, including size, floor location, included objects, and distance to the goal region. Clicking on a room name allows Guide to view and navigate nearby areas.
 - **Shortest Path Trajectory:** The shortest path from the current location to the goal region, detailing nodes, rooms, and floors along the path. Clicking a node navigates Guide to that node in the navigation interface.

I. Data Collection Details

We recruited annotators from volunteers within the university, though participation was not limited to students. All participants were between the ages of 20 to 35. Each participant received a compensation of 50,000 KRW for participating in the experiment for a continuous 3.5-hour session, which is equivalent to 1.45 times the minimum hourly wage.

J. Data Preprocessing

A total of 2,421 episodes were conducted. To ensure dataset quality, we first filtered out episodes that contained technical faults or exceeded the predefined time limit of 22 minutes. Additionally, we manually reviewed outlier episodes based on three criteria: trajectory detour, dialog count, and annotator scores. Specifically, we identified episodes within the top 1% of outliers for each metric—5.6 for detour, 6 for dialog count, and 1 for annotator scores. Upon review, these episodes were retained, as they were deemed sufficiently

valid for the task. Fig. H and Fig. I presents examples of qualitative DialNav samples and outliers respectively. To further refine the dataset, we manually corrected instances where dialog entries were inadvertently split due to user errors, such as accidental key presses. After these refinements, the final dataset consisted of 2,231 episodes. The dataset was then partitioned into training, validation, and test splits, with 1,559 episodes for training, 285 for testing, and 387 for validation. The validation set was further divided into seen (111 episodes) and unseen (276 episodes) environments, while the test set exclusively contained unseen environments. We adhered to the seen/unseen/test environment split established in prior work [1].

K. The RAIN-Segment Dataset

A RAIN episode consists of multiple dialog turns and corresponding actions. For training, we restructured episodes into segment instances following the methodology of CVDN [7]. Specifically, an episode with three dialog turns generates four segment instances: one representing the trajectory before any dialog and three corresponding to successive dialog turns.

Each RAIN-Segment is defined as $(G, I, \mathcal{R}, \mathcal{T}_t, \mathcal{D}_i)$, where G , I , and \mathcal{R} are common across the entire episode, while \mathcal{D}_i is dialog history until dialog turn i and \mathcal{T}_t is trajectory history up to dialog turn i . Each RAIN-Segment instance represents a specific state within an episode, capturing the interactions between the Navigator and the Guide.

For Navigation training, all instances are utilized, while for Localization and QA generation, first segment in each episode which does not contain dialog is not included. We adhered to the seen/unseen/test environment split established in prior work [1]. Refer to Tab. F for the respective counts.

Type	Train	ValSeen	ValUnseen	Test	Total
DialNav	1,559	111	276	285	2,231
Instance	4,493	337	805	768	6,403
w/ dialog	2,934	226	529	483	4,172

Table F. Dataset size per splits

L. Module-wise Training and Experiments

Navigation. We utilized two existing VLN models for navigation: HAMT [3] and DUET [5]. HAMT is first initialized with publicly available pretrained weights, and finetuned on DialNav training instances. To evaluate the impact of the large-scale pretraining on DialNav navigation, we compared the DUET with weights pretrained on ScaleVLN [37]. For each RAIN-Segment instances, the initial instruction and the dialog history is appended to form a navigation instruction for each instance. The VLN models were trained to determine the best next optimal action given the instruction. Due

Q:	Hi I'm in a grey bedroom. The bed has dark brown frame and grey patterned sheets. There's a grey sofa at the foot of the bed, facing a TV. I can see two big windows.
A:	Please exit the room using the door on the right. Then walk down the hallway and you'll find a kitchen on your left and a living room with sofas on your right. Go through them and you'll see pictures of a couple and their child on the wall. There are 6 in total with different shapes. Do you see them?
Q:	Yes. I'm standing in front of them.
A:	Ok now take a right turn and walk a few steps ahead. Then when you see two pictures of a boy and a girl on the wall, make a right turn and you'll see a bedroom. Our destination is the bathroom attached to the bedroom

Q:	I was standing in the corner, which has two ways to go. There are lots of rooms and they look all same. I entered a nearest room, and I can see a white table and a black chair. On the door, number '321' is written.
A:	Great. Go out of the room and turn left. Walk straight down the hallway up until you see 4 red sofas on your left. Tell me when you get there
Q:	I can see the red sofas. I am facing them.
A:	Okay. Turn right there. Having the red sofas on you back, walk down the hallway up until you see a white pillar and a poster of a men attached to the glass wall on you left. And a exit sign on your right
Q:	I am there! Facing the exit sign.
A:	Great! Here's the tricky part. Walk more down the hallway, until you find a messy room with a microwave on your right, and in front of the goal room, there's another small office with only one white desk and chair.

Q:	Hello, I am in a small bathroom with a toilet and a black sink. There is a door to my right which is wooden brown, and also another door to my left which is white.
A:	Alright, go through the white door, and go straight at the sofa. If you stand in front of a table with fruits, that's the goal room.

Q:	now im in a bedroom; a bed with patterned bedding, an abstract painting of two horses right above the bed, a black chair set in front of the bed.
A:	You're on the 4th floor and we're gonna have to make a way to the 2nd floor. So I want you to come out of the bedroom through the door on your right: to the TV side. Once you're out, there will be a lounge with a round wooden table and a staircase going down on your 2 o'clock. We'll take down those round stairs and let me know when you're down and find another stairs in front of you.
Q:	yep now im facing another stairs
A:	Yes head down the stairs to the entry lobby and make a U-turn to your left and there will be another short stairs for you to go down. Once you're down again from those stairs, go to the living room that you can find on your 2 o'clock: right side of the statue at your front. Once you've made your way to the living room, take a right turn and head to the pool table. There's a bathroom on the right side of that pool table and that should be our goal room.

Q:	I'm in a very fancy living room with a greyish blue fireplace with a mirror above it. There is also a chandelier, and the pattern on the floor has some red in it. Should I leave the room into a long hallway?
A:	There are several rooms that have a greyish blue fireplace, so could you please elaborate on it more? For example, it has a brown wooden round table, it has a white striped patterned sofas, like that. Thanks.
Q:	Oh, sorry. It has a long white couch, with 5 small white chairs aligned next to it. There is also a small bed in the corner.
A:	Oh, I found you. Enter the hallway, which is on the left side of the fireplace. If you walk through the hallway, you'll see the stairs on your right side. Could you go halfway down the stairs?
Q:	Yes, I'm halfway down the stairs. Keep going?
A:	You have to stop. Is the stair you've walked down on your behind?
Q:	Yes, and there are stairs on either side behind me. I'm facing the white wall.
A:	Great. Now just take one step at the stair in right side of you. That is the goal room.

Figure H. Sample RAIN dialogs

Q:	I'm in a room with a swimming pool. There are 3 sunbeds, and I see a treadmill on one side.
A:	good new. it's near. assuming you are facing the swimming pool, turn right and take a few step and turn right and you will see two spa rooms: take the one on the left and the room has this fan with red flowers and a blanket with check patterns. Text me if you aren't confused!
Q:	I'm looking at the swimming pool, and there are sunbeds behind me. If I turn right and another right, I see a concierge, not spa rooms.
A:	ok. there are two pools right? Stand in front of the larger pool. (The one on the left) and you see that there are plants next to that pool. There are two rooms behind those plants. I need you to go behind that plant, so I need you to head towards the diamond shape wall. From here, turn right. Let me know if you find the diamond wall.
Q:	Hi. I'm in the room with a huge bed which has a purple bedding, and I can also see the light green colored sofa. I guess I'm at the second floor.
A:	Hi. Is the roof inclined not flat? And do you see trees through a window?
Q:	Yes. The roof is inclined, and I can see trees through the window.
A:	Great! You're currently on the third floor. First come out from the bedroom. While you come out from that room, trees should be on your left side.
Q:	Okay. Should I go down stairs?
A:	Yes. Please go down the stairs.
Q:	I came down. Now I'm standing at the kitchen, and can see a big wood table.
A:	Terrific! You will see a hallway at the left side of that big wood table. Please go to the hallway.
Q:	I'm at the hallway now. I can see a pool through the window.
A:	Sorry! Do you see a black car through a window?
Q:	I guess I was at the different hallway. I got to the other one, and now I can see the black car through the window.
A:	Fantastic! There are brown stairs between white walls. Please go down to the first floor by following those stairs.
Q:	I came down! Now I'm looking ahead a room with work out machines.
A:	Good. Please do not go into that room. Instead enter a room next to the room with workout machines. Then you will see a wall on which a kind of graffiti is drawn.
Q:	I'm in!
A:	The place in front of that wall. That is the goal.
Q:	in front of stairs going down, on my left on the wall is an abstract painting, signiture reads what looks like 'Hunk'
A:	i got it. it's quite complicated path to our room. putting that painting on your right, please make a right turn.(kind of u turn you should make) then you might see set of wooden stairs leading to upstairs. going up, you can make a right turn to see a few black stairs. finishing that stair, there might be a kitchen. please make a right turn right at the person like-statue. then there might be a set of wooden stairs, and please go upstairs.finisnihing that, you might find a bed room and going inside, there is a spa. that's our goal room.
Q:	there's no right turn I can make if I put the painting on my right. Im in front of stairs going down, behind me is an open white door (a room with two big yellow sofa beds)
A:	oh isn't that the yellow painting?
Q:	yes, if i make a right turn, i'm in a sauna. just tell me which floor I need to be on. do i need to go downstairs?
A:	our goal room is on the 3rd floor. if that's wooden spa, i guess its our goal room
Q:	I'm at front of stairs, do I have to go down?
A:	There are 3 stairs in this house. I'll need more infos. Please give me more details.
Q:	I can see two bedrooms, one has grey blanket on the bed, and the other one has brown blanket on the bed.
A:	So, you're in hallway 1, floor2. Our goal room is bedroom 4. Please enter bedroom with the brown blanket. That's our goal room.

Figure I. **Sample dialog with outliers:** (1) Dialog with a significant detour, (2) Dialog with a high number of QA turns, (3) Dialog with a low Guide score, (4) Dialog with a low Navigation score.

Method	+SV	Val Seen	Val Unseen	Test
Shortest		22.51	23.06	25.23
Random		1.91	1.94	0.25
HAMT [3]		11.26	8.31	5.87
DUET [5]		11.13	10.22	11.53
DUET [5]	✓	12.88	12.07	12.37

Table G. **Navigation scores on RAIN-Segment.** The goal progress of navigation to the destination is measured when the initial instruction and the last answer from the previous dialogue are given. **+SV:** Pretrained on ScaleVLN [37].

Method	Val Seen		Val Unseen		Test	
	B4 ↑	RG ↑	B4 ↑	RG ↑	B4 ↑	RG ↑
LLaVA [20]	.0311	.2074	.0318	.2183	.0259	.2043
LANA [36]	.0405	.2087	.0533	.2203	.0502	.2131
LANA(pt) [36]	.0532	.2005	.0527	.2111	.0539	.2071

Table H. **Question Generation scores on RAIN-Segment.** The similarity between the questions generated by the model and those generated by humans in the same context is measured. **B4:** BLEU4, **RG:** Rouge-L

to the multi-turn nature of the dialog, the answers often do not contain the complete path to the goal. Therefore, Goal Progress (GP) [31], the distance that agent got closer towards the goal, is used as navigation metric.

Tab. G shows the navigation performance on RAIN-Segment. We report shortest path length and random agent performance serving as an upper and lower bounds respectively. The DUET [5] model, which retains past trajectory history as a graph, demonstrated superior performance in unseen environment compared to the HAMT [3] model. Additionally, pretraining on ScaleVLN [37] further improved navigation performance across all environment splits. This result indicates the advanced model and large-scale pretraining on VLN models brings high performance in DialNav navigation as well.

Through these experiments, we selected DUET [5] pretrained with ScaleVLN [37] which showed the best performance across all environments for our baseline model.

Question Generation. We employed the VLN instruction generation model LANA [36] and the multimodal model LLaVA-1.5 [20] for generating questions. We tested LANA with and without pretraining. Although Navigator has access to the past trajectory and previous dialogs, to simplifying the task, we provided only the visual input from the current viewpoint in this work. LLaVA-1.5 (7B) is simply prompted to generate question including details of current panoramic view.

The objective of Question Generation is to facilitate

Method	Language			Navigation	
	B4 ↑	RG ↑	CD ↑	GP ↑	SR ↑
<i>Val Seen</i>					
Llama-3.1	.0268	.1802	.0595	5.36	18.92
LANA	.0539	.2156	.0953	6.89	23.42
LANA(pt)	.0745	.2342	.1310	10.63	30.09
<i>Val Unseen</i>					
Llama-3.1	.0271	.1765	.0394	5.35	13.01
LANA	.0542	.2082	.0722	8.03	22.52
LANA(pt)	.0596	.2223	.1086	8.92	25.28
<i>Test</i>					
Llama-3.1	.0280	.1881	.0550	6.74	13.08
LANA	.0519	.2217	.0789	6.93	10.55
LANA(pt)	.0648	.2325	.0952	9.50	17.11

Table I. **Answer generation scores.** **LS:** Llama-S, **LF:** Llama-F, **GP (U):** GP on unseen validation set, **GP (S):** GP on seen validation set. The navigation scores (GP, SR) report the performance of the navigation task based on the answers generated by the model, using the DUET model pretrained with ScaleVLN.

human-like dialog; therefore, we evaluated the similarity of the generated questions to those asked by humans in identical situations. Tab. H, reports metrics for question generation. As discussed in the main manuscript, the improvement by pretraining LANA is limited in question generation due to the significant mismatch between pretext and target tasks.

Through these experiments, we selected LANA [36] with pretrained weight for our baseline model.

Whether to Ask. Whether to Ask (WTA) is a task to predict binary decision of to ask or not given navigation context. Previous dialog-based studies have used simple Fixed-Interval [28], Confidence Thresholding [13, 42] or adopting an additional model [29]. We tested all these 3 approaches. In Confidence Thresholding, questions are triggered when action confidence falls below a threshold. For the third approach, we added a decision head to DUET [5] model to utilize the intermediate output of the action decision and trained with RAIN-Segment. We selected the decision head for our baseline model.

Localization. Localization task predicts location of Navigator given the previous dialog. Although Guide has access to the full dialog history, we only use the last question in dialog for this work for simplification. We evaluate two ranking models from [15]: Simple Cross-Modal Network (SCN) and Graph Convolutional Network (GCN). Both compute node embeddings for the environment graph G and obtain a cross-modal feature by element-wise multiplying it with the

Method	Val Seen		Val Unseen		Test	
	LE↓	A@3↑	LE↓	A@3↑	LE↓	A@3↑
Random	19.76	7.96	18.77	4.34	21.54	3.46
SCN	10.20	44.69	14.63	23.58	16.04	19.14
SCN†	10.97	42.04	13.87	23.58	15.60	22.20
GCN	12.09	38.94	13.33	24.91	16.53	16.90
GCN†	11.47	36.73	11.63	31.19	15.28	22.77

Table J. **Localization scores on RAIN-Segment.** Scores are reported with the model with least LE for in Validation Unseen set in each model. **LE**: Localization error in meters. **A@3**: Accuracy with an allowable error margin of 3m.

query embedding, followed by a linear layer for ranking. The key difference is that SCN models each node independently, whereas GCN captures graph structure. All models are first pretrained on the WAY [15] dataset and then finetuned on RAIN-Segment. For evaluation, we adhere to the protocols in [14, 15], measuring localization error in meters and the accuracy of successful localization within 3 meters.

The SCN exhibits a significant performance disparity between seen and unseen environments, with minimal performance gains from WAY pretraining. (Tab. J ln2, 3) In contrast, the GCN model demonstrates greater gains from pretraining and better generalization to unseen environments. (Tab. J ln4, 5) This suggests that more data and complex models are required to generalize effectively to novel environments. The localization task remains underexplored, necessitating further research.

Through these experiments, we selected GCN with pre-trained weight for our baseline model.

Answer Generation. We tested VLN instruction generation model LANA with and without pretraining. We also tested Llama-3.1-8B-instruct [32] with prompting to test power of large language model for answer generation task. For Llama, we first generates detailed caption of each navigation nodes with LLaVA-1.5 [20] and prompts Llama to generate navigation instruction based on the sequence of captions through route. See Fig. M and Fig. N for LLaVA and Llama prompts. Although Guide has full access to previous dialog and entire house map, we only use the remaining trajectory to the goal as input to simplified the task.

The objective of Question Generation is to facilitate human-like dialog; therefore, we evaluated the similarity of the generated answers in identical situations. Answer generation is similar to the VLN instruction generation task, so metrics improved through pretraining. To check the impact on navigation, we also reported goal progress and SR of DUET [5] model pretrained with ScaleVLN [37] when the answers were given as instructions.

Through these experiments, we selected LANA [36] with pretrained weight for our baseline model.

M. Question Generation Examples



Ground Truth (GT): I’m at a bedroom. There is a purple sofa. The bed has red blankets, one purple cushion and two white pillows. The outside is clearly showing, with trees and multiple cactus.

LLaVA: I’m in a room with a large window, and I can see a view of a desert landscape outside.

LANA: I’m in a hallway with a large glass window. Next to the sink is a glass shelf to the outside. There is a glass door to the right next to the sink. To the sink, there is a glass shelf to the kitchen.

Figure J. Qualitative analysis of question generation from different models.

Fig. J shows example of question generation given a view-point.

N. Answer Generation Examples

Fig. K shows an example of answer generation based on a list of panoramic views along the route to the goal room.

O. LLaVA Prompt for Question Generation

The prompt for LLaVA-1.5 used in question generation is shown in Fig. L. We provide LLaVA with a panoramic image to generate a scene description as a Navigator question.

P. Llama based Answer Generation

P.1. Generating LLaVA captions

The prompt for LLaVA-1.5 caption generation is shown in Fig. M. We provide LLaVA with a list of panoramic images and extract captions for the images corresponding to the path.

P.2. Prompting Llama to generate answers

The prompt for LLaMA-3.1-8B Instruct answer generation is shown in Fig. N. The scene description extracted from LLaVA, along with the corresponding prompt, is provided as input to Llama with an one-shot example.

Q. Llama Prompt for Language Evaluation

The prompt for LLaMA evaluation is shown in Fig. O. This prompt is used to evaluate fluency scores for both questions and answers (QF and AF).



Ground Truth (GT): Almost there. Now walk towards the light grey sofa in front of you. On the left to that sofa, there is a path to a terrace where you can see the pool on the left. The terrace is our target

LLaVA: Head towards the large window with the modern design, and you'll find the dining area with a table and chairs. The kitchen is adjacent to the dining area, and you can access it through a glass door. If you'd like to explore more, the staircase leading to another floor is located in the living room, which is decorated with a vase, a TV, and several books.

LANA: great . so , facing the white couch , you will see a grey couch on your right , and a tv on your right . go straight until you see a living room with a tv on your right . that 's our goal room .

Figure K. Qualitative analysis of answer generation from different models.

Prompt:

You will be given a panoramic image of indoor scene. Create a sentence including what type of room the given image is (ex. bedroom, bathroom, empty room, stair, hallway), and details and any unique objects that would not appear in other regions of the building so that someone else can easily locate. Start with 'I'm in' or 'I can see'.

LLaVA Output:

I'm in a room with a wooden door and a blue curtain.

Figure L. Prompt for LLaVA's question generation and its results.

Prompt:

Please describe the region in this image.

LLaVA Output:

The bathroom in the image is a small space with a toilet and a sink. The sink is located near the toilet, and there is a book placed on the countertop. The bathroom appears to be part of a larger living area, as it is situated next to a wall and a door.

Figure M. Prompt for LLaVA's caption generation and its results.

Prompt:

You are an agent for creating navigation route. Given sequence of scene image description, create a navigation guide sentence for the route. You don't have to describe every single step. Try to add any unique object or landmark. Return your evaluation results in the following JSON format without any additional text:
'response': '<your response>'

[One-Shot Example]

Llama Output:

Figure N. Prompt for Llama's answer generation and its results.

Prompt:

You will be given a sentence, and your task is to evaluate its fluency. Fluency refers to how natural and grammatically correct the sentence sounds in English. Rate the sentence on a scale from 1 to 10, where:

1: Very unnatural, severe grammar or structural errors.

10: Perfectly natural, indistinguishable from a native speaker's writing.

Your output should be in JSON format, containing your evaluation results based on the criteria above.

Figure O. **Prompt for Llama evaluation of fluency scores for question and answer(QF and AF).**