

RoomTour3D: Geometry-Aware Video-Instruction Tuning for Embodied Navigation

Supplementary Material

The indexes of figures and tables in the appendix are continuous to the main sections for easy reference.

Dataset release. Our annotations and intermediate products are released at <https://huggingface.co/datasets/roomtour3d/roomtour3d> under CC-BY-SA-4.0 license. The downsampled and sampled video frames are released at https://huggingface.co/datasets/roomtour3d/room_tour_video_3fps under CC-BY-SA-4.0 license. The codes and project updates are hosted at <https://roomtour3d.github.io/>.

Overview. In the supplementary material, we provide

- **Section 7:** Room tour video collection process.
- **Section 8:** Navigable point extraction used for action-enriched trajectory generation.
- **Section 9:** Object variety and spatial awareness for trajectory descriptions.
- **Section 10:** Room tour 3D scene reconstruction.
- **Section 11:** Further model implementation details.
- **Section 12:** Qualitative results showcasing the instruction following capabilities of our trained model.
- **Section 13:** Data samples and excerpts from our data verification report to illustrate data curation correctness.
- **Section 14:** Broader impact of our work, including limitations and future extendable works.

7. Room Tour Video Collection

To enable more diversity for indoor scenes, we leveraged the rich variety and volume of room tour videos available on YouTube. These videos, recorded with hand-held cameras from a first-person perspective, offer a realistic and dynamic view of indoor environments. We curated a dataset from 1847 YouTube room tour videos, in total 243 hours. Our data collection approach builds on the video list from YTB-VLN [34], which we further filtered and expanded to enhance diversity and quality.

To ensure high-quality data, we prioritize continuous videos with least transitions, such as human interviews or abrupt cutting into close-ups, for better 3D reconstruction. We applied a title-description-based filtering process by using GPT-4 [40] and excluded videos shorter than three minutes. Additionally, we detected abrupt video transitions, retaining videos with at least nine continuous shots occupying over 80% of the video duration. We further extended our dataset by continuously updating high-quality channels (*e.g.*, NavaRealtyGroup, Open House 24, Sona Visual) with new videos, resulting in our current 1847 room tour scenes.

To process this data, we spatially downscale the resolution to shorter side 360 and temporally downsample the frame rate to 3 frames per second. All the following processing are performed on this downsampled data.

8. Navigable points generation

To inject open-world knowledge from room tour videos into navigation agents, we propose navigating agents using video frames. Each frame in a human walking demonstration can be treated as having two next actions: move forward or stop. However, at significant view-change points — instances of distinct view shifts within a close radius — we sample frames with varied orientations as candidate actions to enhance the agent’s training. Unlike YTB-VLN [34], which composes panoramic images at room nodes, our approach involves taking every significant view-change point and its neighboring frames that meet specific criteria as candidate actions.

First, we detect significant view-change points along person’s trajectory. By reconstructing the 3D scene, we can determine the camera orientation difference and distance between frames. There are instances where the person may revisit a nearly identical location, resulting in varied views within almost the same spatial region. Additionally, turning points with notable view changes in close proximity are essential to capture. Identifying these view-change points is useful for producing diversified navigable action data, especially when panorama images are not available.

To find these points, for each point along the trajectory we calculated pairwise cosine similarity. We then applied a threshold of 45 degrees to retain only frames that demonstrate a substantial change in view. Afterwards, non-maximum suppression is performed along the trajectory to isolate local maxima in angular change to highlight the most significant view changes.

To account for the points that are close in proximity, but have different views due to an intersection in the walking trajectory, we performed DBSCAN clustering [11] of the points that were retained after Non-Maximum Suppression. This clustering step ensures a diverse set of navigable actions is maintained, even without the availability of panoramic images.

Finally, as shown in Figure 5, to extract varied navigable action candidates, we post-processed the clusters by identifying the distinct walking paths of the person within each cluster. In cases where paths intersect, the cluster may encompass two separate routes. For each walking path, we

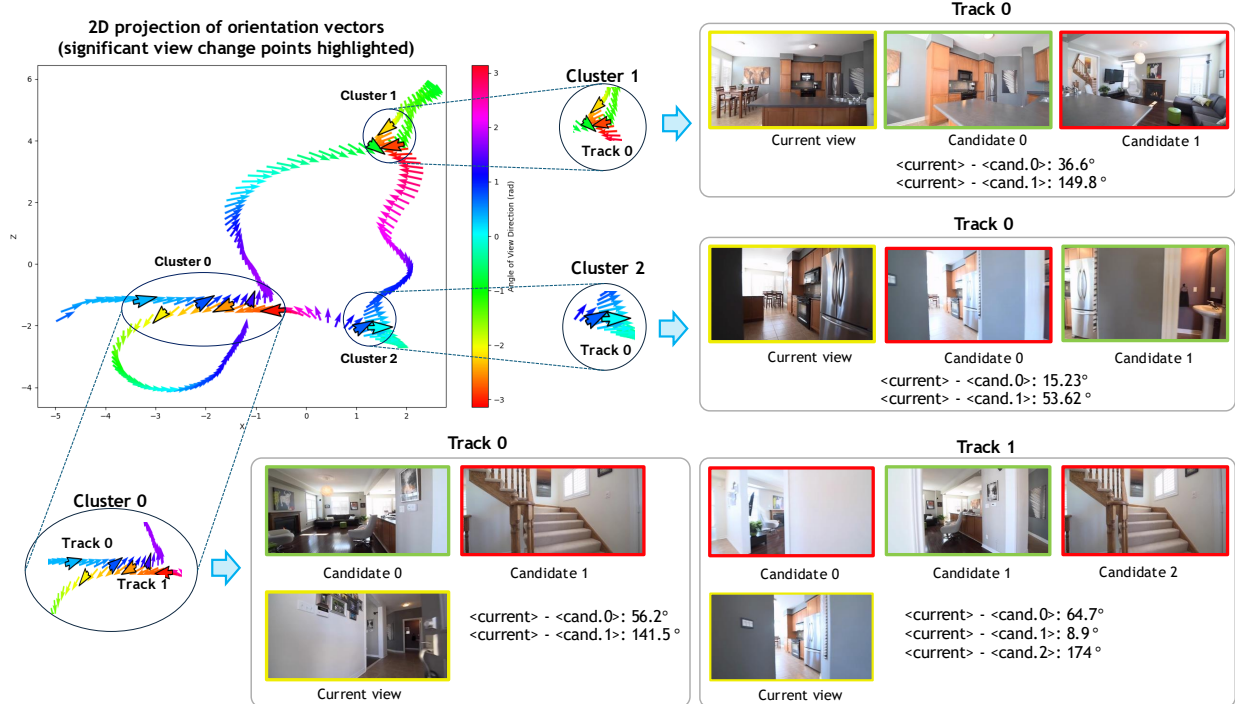


Figure 5. Visualization of significant view change point selection. For each cluster we identify the walking tracks and find the candidate views for the next action selection. This process ensures we have a diversified set of views in the setting without panorama images.

select the most recent frame on the walking path as a positive candidate, while a negative candidate is chosen as the frame within the cluster that exhibits the highest angular difference in view with the positive candidate.

9. Instruction Generation

In this section, we detail the process of transforming spatial awareness and object variety information into textual captions for use with GPT. This involves extracting multi-source data using models such as RAM (Swin-L) [61], Grounding DINO [38], and Depth-Anything [59], and then organizing this information into structured text inputs.

Object variety into texts. Web videos offer a rich, open-world setup, capturing diverse items, arrangements, room functionalities, and layouts, which are critical for training open-world navigation agents. To fully utilize this diversity and ensure a controllable generation of instructions, we use RAM [61] (Swin-L) to extract object tags in each frame. For each frame, we filter out the resulting entries indicating room types in order to be consistent with the identified room locations from BLIP-2. Then these object tags are used for grounding objects within the frames, for further integration of spatial awareness information.

Spatial awareness into texts. Navigation agents are frequently tasked with approaching or obtaining objects, making it crucial for them to sense object locations and dy-

namics during movement. To achieve this, we jointly use Grounding DINO [38] and Depth-Anything [59] models to gather detailed spatial information. The reason why we used Depth-Anything over the depth derived from COLMAP [46, 47] reconstructions is its ability to directly extract reliable depth without relying on long-range frames or structure-from-motion processes, which are prone to errors in complex video reconstructions. This spatial awareness information is then transformed into text inputs suitable for GPT, enabling effective training.

We start by using Grounding DINO to spatially localize objects within the video frames. We define spatial locations relative to the capturing spot: *to the left of the current spot*, *in the middle*, and *to the right of the current spot*. Specifically, the center 40% of the frame is considered the middle, the leftmost 30% as the left, and the rightmost 30% as the right. For depth perception, we categorize distances into three ranges: *in the near distance* (closest 30%), *in closer distance* (next 40%), and *in a further distance* (remaining 30%).

Followingly, we integrate spatial location and depth estimation by measuring the overlap ratio between objects and the defined distance ranges. For example, if a carpet overlaps with the near-distance area by more than 30%, we consider the carpet to be in the near distance to the capturing spot. Large objects that span multiple distance categories, such as a carpet visible in both near, closer, and further dis-

You will be given a set of continuous frames. The frames are captured during the camera movement. During movement, the objects in the frames change gradually, like objects passing by, objects moving towards somewhere. You should return a single and concrete sentence describing the camera moving trajectory by the object's progression in the frames. You don't need to mention all the objects. It is good to describe the moving trajectory without all of the objects.

Frames:

\t0: in the study. there is a clock to the right of the current spot in the near distance, a door on left in further distance, a window and curtains in the middle in far distance.

\t1: in the study. there is chair, laptop, table in the middle in the near distance, a door on left in further distance, a window and curtains in the middle in far distance.

\t2: in the study. there is a plant to the left of the current spot in the near distance, wall to the right of the current spot in the near distance, a bench in further distance in the middle, a window and curtains in the middle-right in further distance

Your moving trajectory description: Walk in the study. Move from right to left, pass by a clock to the right of the current spot, approach a table with a chair and laptop, and continue towards a window and curtains in a close distance, approach a plant to the left of the current spot.

Example 2:

Frames:

\t0: In the study. there is a plant, laptop, and table to the left of the current spot in the near distance, a bookshelf to the left of the current spot in the far distance, a door in the middle in further distance, and two art frames to the right of the current spot closely.

\t1: In the study. there is a bookshelf to the left of the current spot in further distance, a door in the middle in further distance, and an art frame in the middle in far distance.

\t2: In the hallway. there is a door to the left of the current spot in the near distance, art frames to the right of the current spot in closer distance

\t3: In the hallway. there is a art frame to the left of the current spot in the near distance, a switch to the right of the current spot in the near distance, a lamp and future stool in the middle in far distance

\t4: In the hallway. there is a wall to the left of the current spot in the near distance, a bed to the left of the current spot in far distance, a wall and lamp and furniture stool in the middle in closer distance.

\t5: In the hallway. there is a wall to the left of the current spot in the near distance, a bed in the middle in closer distance.

\t6: In the bedroom. there is a art frame, plant and furniture stool to the left of the current spot in the near distance, a bed in the middle in the near distance, a window and curtain in a far distance.

Your moving trajectory description: Exit the study. Move from left to right, start near a plant, laptop, and table, pass a bookshelf and approach a door, then shift towards art frames enter the hallway, before move past a switch and approach a lamp and stool, and finally arrive at the bedroom at a bed with a window and curtain in the distance.

Your turn:

Frames:

{clip_desc}

Your moving trajectory description:

Figure 6. Prompt used for GPT-based instruction generation. We provide instruction for this generation task, in-context examples.

tances, are annotated accordingly to reflect their extended presence within the scene.

This structured approach ensures that our instructions capture the relative positioning and depth of objects, providing comprehensive context for navigation tasks. These texts are then further organized into GPT to generate contextually rich instructions for navigation agents training.

GPT generation. We utilize GPT-4 to summarize the object progression during the walking trajectory, leveraging the detailed object variety and spatial awareness texts. The template used for organizing the components is depicted in

Figure 6. For each clip, we organize the object tags, spatial locations, relative distance to the camera and room locations per frame. This arranged content is then fed into GPT-4 to generate the trajectory summary and instructions. For data sample visualization, please refer to Sec. 13.

10. Room Reconstruction

To obtain complete geometric information, we adopt COLMAP [46, 47] for indoor reconstruction. In this subsection, we detail the procedure of reconstructing room tour scenes, which further facilitates sampling navigable frames.

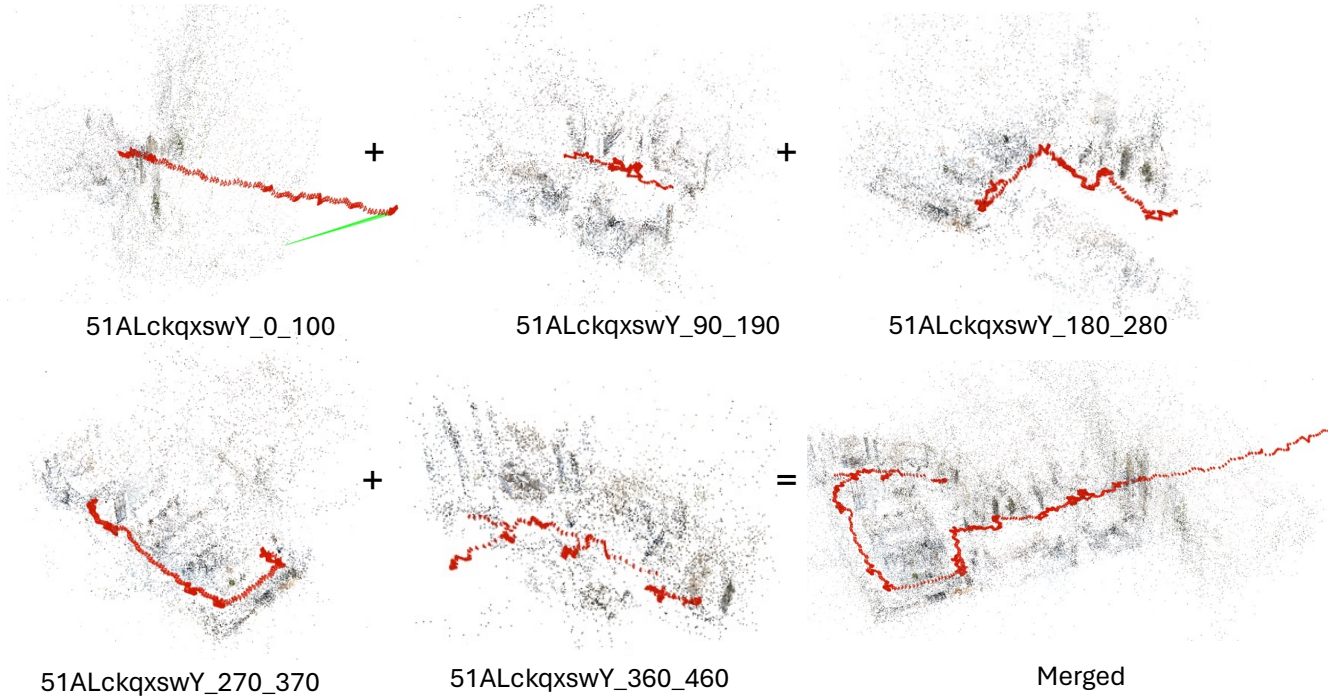


Figure 7. Illustration of the COLMAP model merging process. Reconstructed models from 5 adjacent video clips are successfully merged into one holistic model.

Reconstruction of video clip. To reconstruct video clips, we start by sampling videos at 3 frames per second to balance accuracy and execution time. This frame rate provides sufficient detail for accurate 3D reconstruction while maintaining manageable processing times. Each video is divided into 100-second clips with a 10-second overlap between adjacent clips. Using COLMAP, we perform structure-from-motion and multi-view stereo processing on each clip. It estimates camera poses and generates a sparse 3D point cloud by identifying and matching feature points across frames. The command used for reconstruction is shown as follows, in which ‘\$DATASET_PATH’ denotes the folder containing sub-clip frames and reconstructed models will be located.

```
colmap automatic_reconstructor \
  --image_path $DATASET_PATH/$IMG_FOLDER\
  --workspace_path $DATASET_PATH \
  --data_type individual \
  --quality high \
  --single_camera 1 \
  --sparse 1 \
  --dense 0 \
  --num_threads 10 --use_gpu 0
```

COLMAP model merging After performing individual reconstructions on video clips, we proceed to merge the resulting COLMAP models to create a unified 3D representation of the room tour scenes, as shown in Figure 7.

We begin by identifying overlapping frames between adjacent clips. These overlapping frames serve as common

reference points for aligning and merging the separate models. If two reconstructed models share more than 3 common frames, we will try to merge these two models using the command as the following, where model merging and bundle adjustment are conducted in sequence.

```
colmap model_merger \
  --input_path1 $MODEL_1 \
  --input_path2 $MODEL_2 \
  --output_path $RESULTED_MODEL_BEf_BA

colmap bundle_adjuster \
  --input_path $RESULTED_MODEL_BEf_BA \
  --output_path $RESULTED_MODEL_AfT_BA
```

However, due to potential variances in reconstruction quality, a single video clip may produce multiple sub-models. To handle this, we adopt a graph-based approach for merging, *i.e.*, Depth-First Search. In this approach, each sub-model is represented as a node in the graph. Edges are created between nodes that share more than three overlapping frames, indicating that these sub-models can be merged.

We iteratively merge the model nodes with edge connection existing by traversing from the first video clip (*e.g.*, clip “0_100”). The successfully merged model will be a new graph node to replace the original separated two nodes. In order to monitor the quality of this model merging operation, we use reprojection error to determine whether rolling back the merging operation. Specifically, if the error of the

merged model is even larger than the sum of the two separate models, the model merging operation will be rolled back. This iterative merging process continues until no further connections exist, resulting in a comprehensive and continuous 3D model of the room tour scenes. The final merged model provides detailed geometric information that is crucial for accurately sampling navigable frames and enhancing the training data for navigation agents.

11. Implementation details

Following the practice from NaviLLM [63], we fine-tune the multi-view fusion module and the LLM. The multi-view fusion module consists of a 2-layer transformer encoder with a hidden size of 1024, and the LLM is built upon Vicuna-7B-v1.1 [9]. The ViT in the scene encoder is EVA-CLIP-02-Large, which remains frozen during training. Our training follows a two-stage strategy using the Adam optimizer with a learning rate $3e-5$. The model is trained for 2500 steps in the pre-training stage and 1250 steps in the multi-task fine-tuning stage, with a batch size 256. The training process utilizes 4×8 Nvidia A100 GPUs. During testing, we employ a sampling strategy with a temperature of 0.01 for the SOON and REVERIE tasks to encourage exploration, while a greedy strategy is used for other tasks. This approach ensures robust performance across various evaluation scenarios.

12. Qualitative Results

This section presents qualitative results to demonstrate the effectiveness of our model trained with the RoomTour3D dataset. The model was evaluated on unseen scenes using the R2R dataset, focusing on its performance in following navigation instructions. As shown in Figure 8, we tested the model on an unseen scene, *8194nk5LbLH*, with trajectory ID 4332. Experimented with two different instructions, the agent trained our data shows its flexibility in following the instructions. For example, in (a), the agent moves straight to the bar, then reaches the three tables with chairs, and finally stops near the couch. In (b), the agent directly moves towards the window, following the instructions, then moves towards the far coach and stops. These results demonstrate the instruction-following navigation ability of the agent, which further highlights the effectiveness of our video-instruction data.

13. Data Sample Visualization

In this section, we present visualizations of data samples from the RoomTour3D dataset, as shown in Figure 9. These visualizations highlight the rich variety of indoor scenes, the spatial awareness embedded in the data, and the detailed annotations used for training navigation agents.

Data correctness verification. We provide part (14 out of 100) of manual check trajectories in Figure 10 and Figure 11. For each trajectory, sampled frames and gener-

ated descriptions are provided, along with the manual check scores. The score ranges from 1 to 4, representing “totally irrelevant”, “partially relevant”, “mostly relevant” and “perfect match” respectively. Most of the sampled trajectories gain scores 3 and 4, which shows the convincing quality of our automatically generated descriptions.

14. Broader Impact

Data Limitations and Ethical Considerations. We provide downsampled video frames instead of the original videos. Users can also download these from the original sources. Additionally, our meticulous filtering process ensures that the video frames and annotations contain only indoor rooms and houses, containing no personally identifiable information or offensive content. The authors will take responsibility for long-term maintenance.

Scope of Conclusions. It is important to recognize that experiments and data, including ours, might only represent a subset of universal realities. Nevertheless, given the wide range of room tour scenes covered in our videos, we believe our conclusions offer a robust understanding applicable to indoor embodied navigation. While specific to our dataset and results, these findings provide significant insight into the broader field of embodied navigation.

Usage of Language Models and Simulators. Our use of the LLaMA model¹ from Meta, use of MatterPort3D data [4] is authorized for research purposes. Those intending to use our model post-release should ensure they have the necessary permissions and adhere to usage restrictions. We express deep respect for the work of developers and contributors, recognizing their integral role in advancing language modeling and data collection.

Future Research and Development. Aligned with our commitment to the research community, we released our code and dataset. This is intended to encourage further research and enable others to build upon our work. Although our current experiments require up to 8×4 A100-80G GPUs for pretraining and 8 A100-80G for multi-task tuning, we are aware this may be a limitation. Consequently, we plan to focus future efforts on adapting these experiments to be compatible with parameter-efficient tuned LLMs. It’s important to note that fitting the experiments within an 8 GPU or fewer framework is not the primary focus of this paper. Still, we consider it a crucial step towards making our research more accessible and inclusive for various research groups.

Also, it would be interesting to investigate the usefulness of our data for grounded question-answering for 3D environments, particularly on the ScanQA dataset [3].

¹<https://llama.meta.com/>

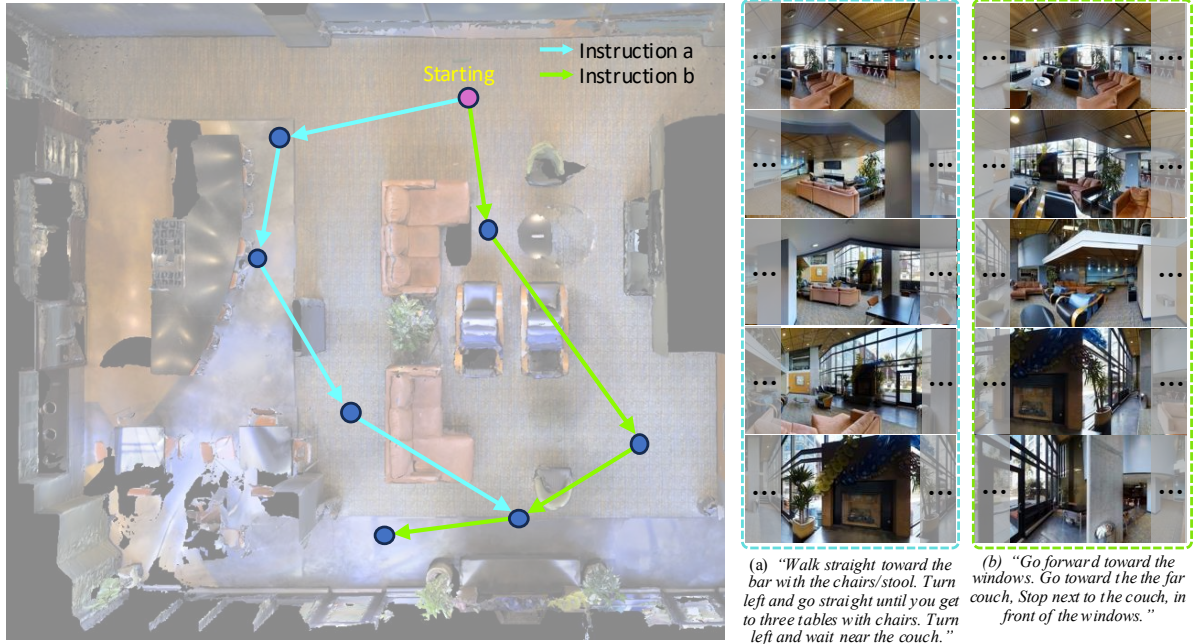


Figure 8. Visualization of the method trained with RoomTour3D on unseen scene 8194nk5LbLH with trajectory ID 4332. The agent successfully follows navigation instructions in R2R dataset. In (a), the agent first moves towards the bar and then approaches the couch. In (b), the agent moves forward towards the windows, then proceeds to the far sofa, and finally stops in front of the window.

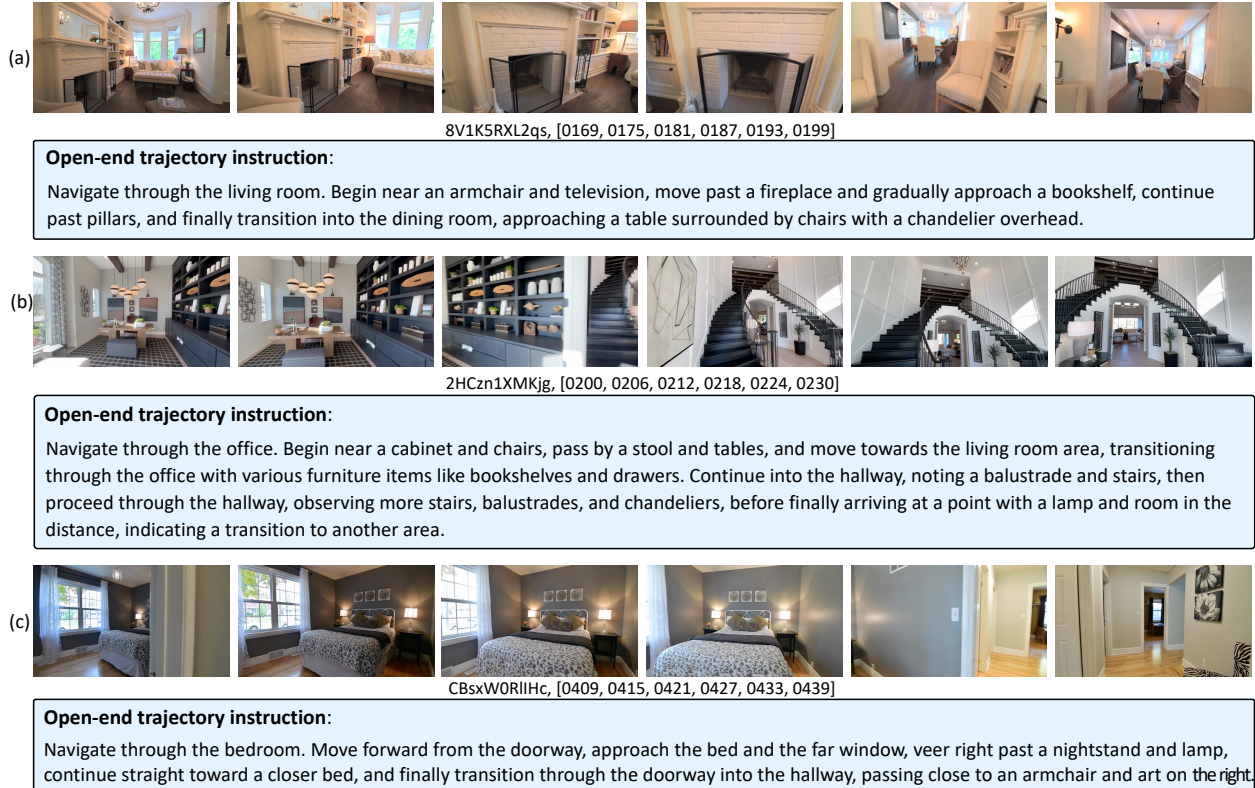


Figure 9. Example open-ended trajectories and instructions. The instruction captures the surrounding environments and the object dynamics (“move past a fireplace” in (a), “move towards the living room area” in (b)), and more importantly, the moving directions and destination (“approaching a table surrounded by chairs” in (a), “into the hallway, passing close to” in (c)). All these data are automatically generated without manual correction.

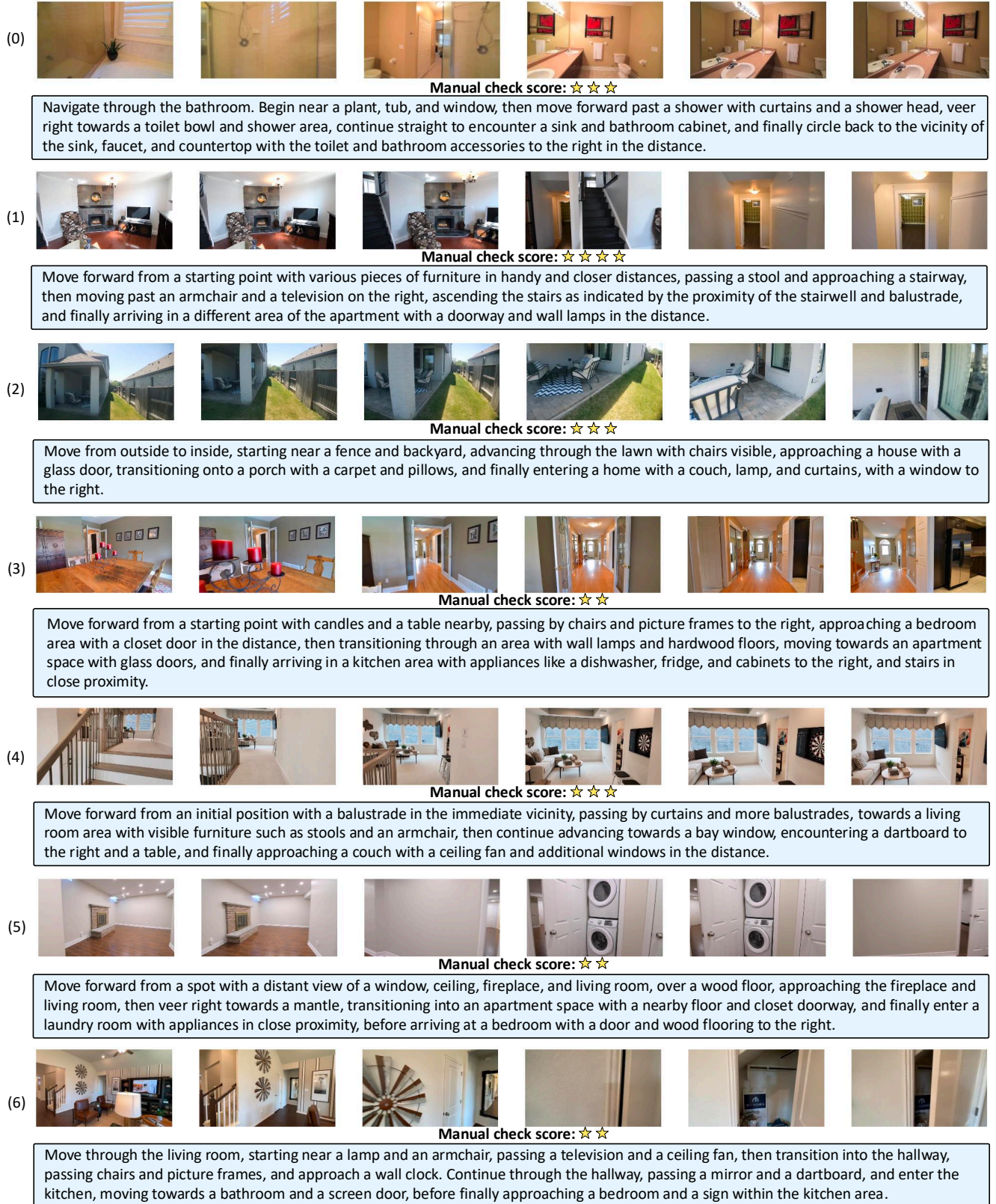


Figure 10. Trajectory samples for manual check. For each trajectory, we provide frames and descriptions for check. The rating ranges from 1 to 4, representing “totally irrelevant”, “partially relevant”, “mostly relevant” and “perfect match” respectively. 7 out of 100 samples are shown here.

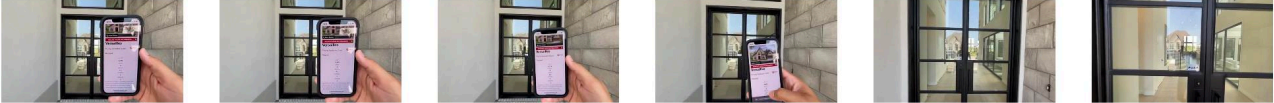
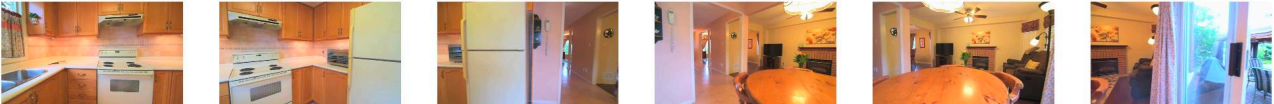
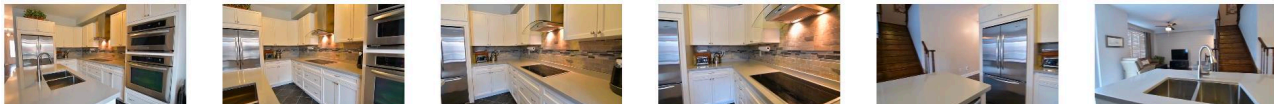

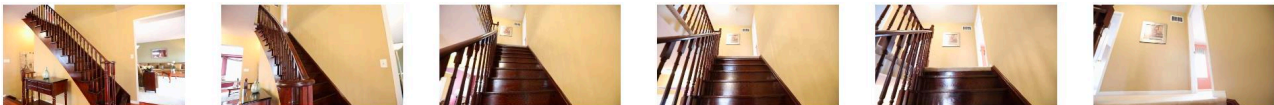
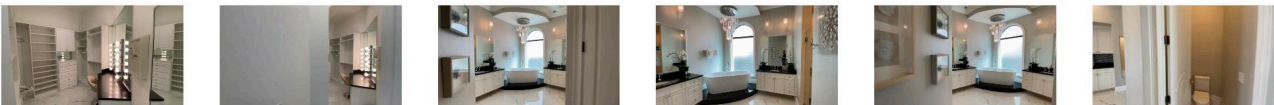
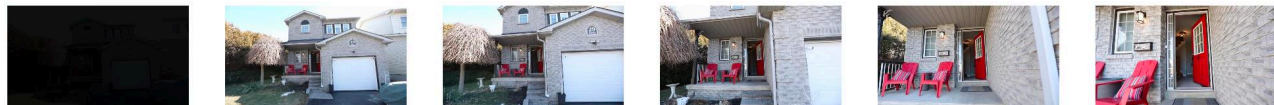
- (7) 
Manual check score: ★★
- Navigate through the bathroom. Begin near a plant, tub, and window, then move forward past a shower with curtains and a shower head, veer right towards a toilet bowl and shower area, continue straight to encounter a sink and bathroom cabinet, and finally circle back to the vicinity of the sink, faucet, and countertop with the toilet and bathroom accessories to the right in the distance.
- (8) 
Manual check score: ★★★★★
- Navigate through the kitchen into the living room. Begin near the kitchen sink and cabinets, move past appliances and countertops, approach the fridge, and continue past more cabinetry. Transition from the kitchen to the living room, passing a table and wall lamps, and finally arrive in the living room, moving towards a glass door with curtains, a fireplace, and armchairs, with a screen door to the right.
- (9) 
Manual check score: ★★★★★
- Navigate through the kitchen. Begin near the counter top and microwave, move past various appliances like a dishwasher and exhaust hood, veer right passing closer to the microwave and oven, continue towards the coffee machine and tile wall, then shift towards the sink and exhaust hood on the right, and finally approach the kitchen island with a stool, ending near the kitchen sink with a ceiling fan and stairwell in the distance.
- (10) 
Manual check score: ★★
- Move forward from a position near a girl and a phone, approaching a bathroom with a mirror and multiple doorways, then pass by a man and various bathroom fixtures such as a faucet, sink, vanity, and bathroom cabinet, before moving through the bathroom door and past curtains and a lamp, and finally turning right towards stairs and stools, indicating a transition from the bathroom area to another room or a stairway.
- (11) 
Manual check score: ★★★★★
- Navigate through the hallway. Progress forward, initially close to a balustrade, then approach a stairwell, continue past picture frames and rails, and finally head towards a room with a window and doorway in the distance, with the stairwell nearby.
- (12) 
Manual check score: ★★★★★
- Move forward from a bedroom setting, passing by a chair and dresser, towards a bathroom area, gradually approaching a stool and vanity on the right, and finally arriving at a bathroom with a tub, sink, and toilet bowl, with a closet doorway in close proximity.
- (13) 
Manual check score: ★★★★★
- Move from the outside towards a house, starting near basketball hoops, then passing by chairs and a garage door, approaching a driveway and yard, and continuing towards the house exterior and porch. Progress closer to the house, passing more chairs and approaching the doorway and stairs, before finally nearing the entrance with wall lamps, a carpet, and a pillow, indicating arrival at the home's threshold.

Figure 11. Trajectory samples for manual check - Continued. For each trajectory, we provide frames and descriptions for check. The rating ranges from 1 to 4, representing “totally irrelevant”, “partially relevant”, “mostly relevant” and “perfect match” respectively. 7 out of 100 samples are shown here.