



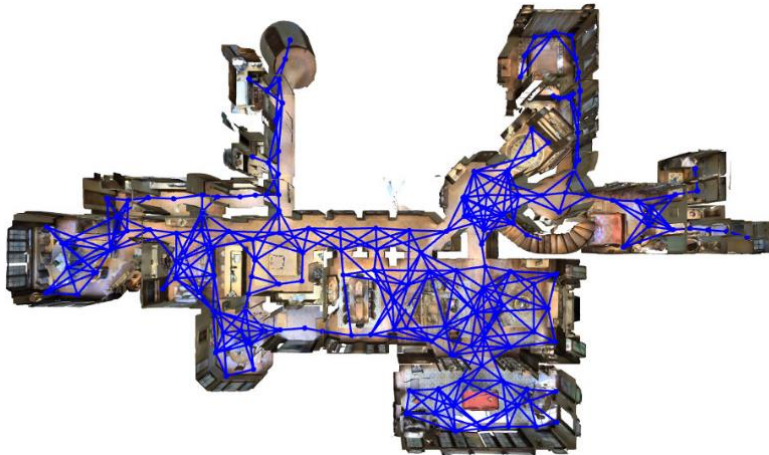
# RoomTour3D: Geometry-Aware Video-Instruction Tuning for Embodied Navigation

Mingfei Han

Mohamed Bin Zayed University of  
Artificial Intelligence

# Vision-and-Language Navigation

- VLN tasks aim to enable an embodied agent to navigate through environments following natural language instructions



**Instruction:** Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

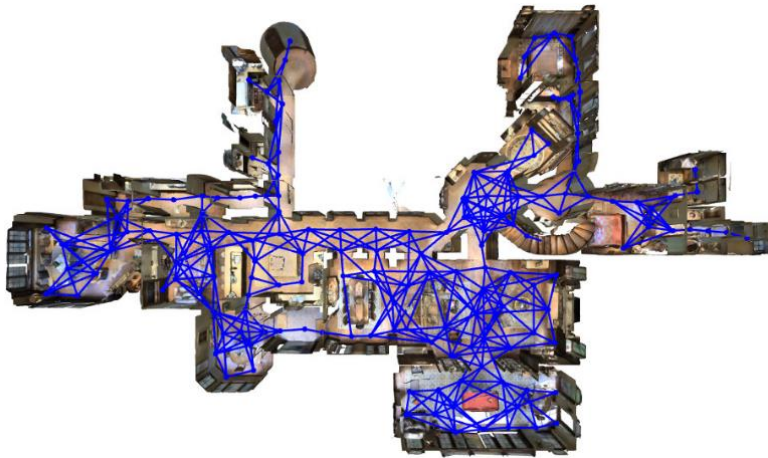
VLN (CVPR'18)

- Given an instruction, the agent navigates to a target location in **novel environments (without pre-built maps)**.
- VLN: The agent navigates on a graph, action space – selecting a node to teleport.



# Vision-and-Language Navigation

- VLN tasks aim to enable an embodied agent to navigate through environments following natural language instructions



VLN (CVPR'18)



**Instruction:** Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.



VLN-CE (ECCV'20)

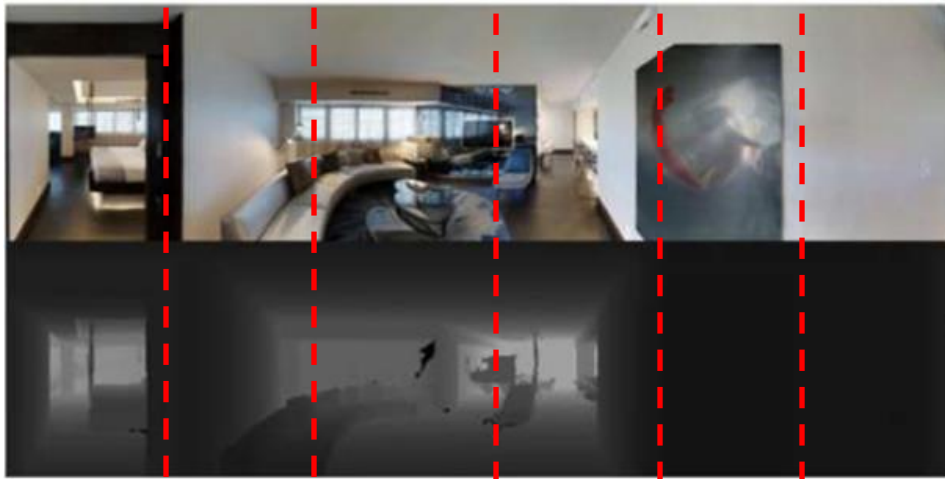


— smooth VLN-CE path  
••••• VLN nav-graph hops

- VLN-CE: The agent navigates on a 3D mesh, action space – low-level control (Rotate, Forward 0.25m).

# Vision-and-Language Navigation

- VLN tasks aim to enable an embodied agent to navigate through environments following natural language instructions



Discrete panoramas (w.r.t. different orientations)

**Inputs:** instructions, discrete panorama at each step, agent pose...

Walk straight through the room and exit out the door on the left. Keep going past the large table and turn left. Walk down the hallway and stop when you reach the 2 entry ways. One in front of you and one to your right. The bar area is to your left.

**Instruction**

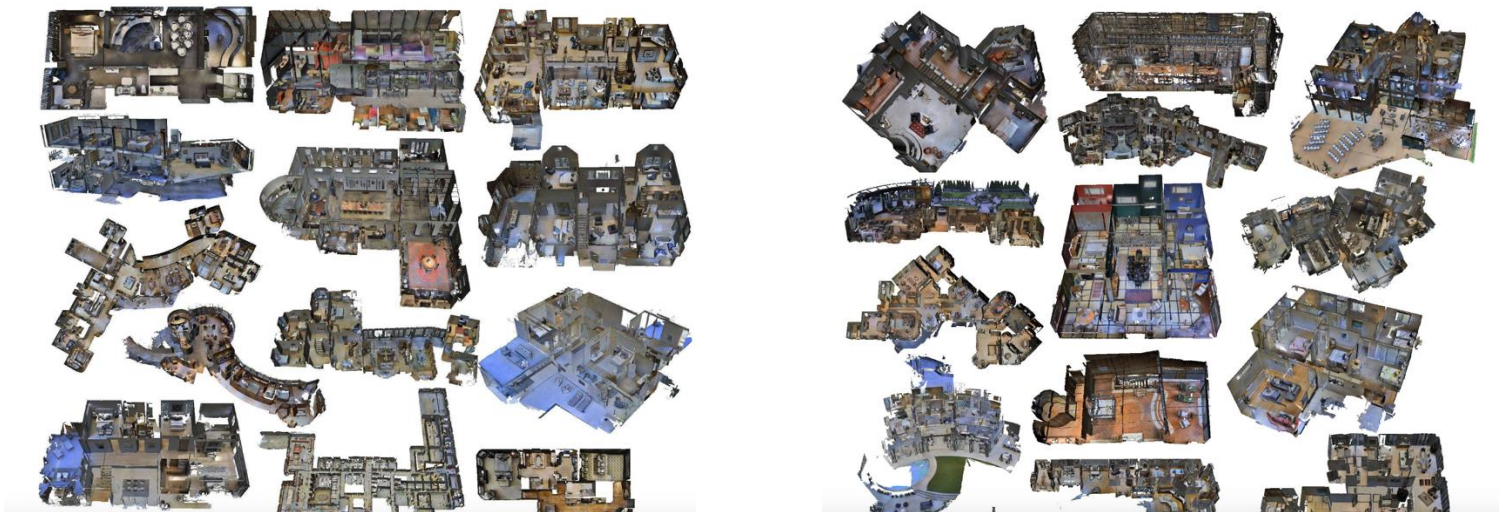
## Evaluation Metrics:

- **SR:** success rate, stop position  $< 3\text{m}$  of the target location
- **SPL:** SR penalized by path length

# Challenge

## Limited Diversity and Scale:

- Most existing datasets (e.g., R2R, CVDN, REVERIE, SOON) rely heavily on manually curated data or limited-scope simulations.
- These datasets lack variability in layouts, object variety, and realistic navigation complexity.



**Matterport3D**  
consists of 90 fully  
annotated scenes.



# Why RoomTour3D?

## Real-World Videos?

- Human behaviors – inherent decision-making and spatial understanding
- Scalability and variety.

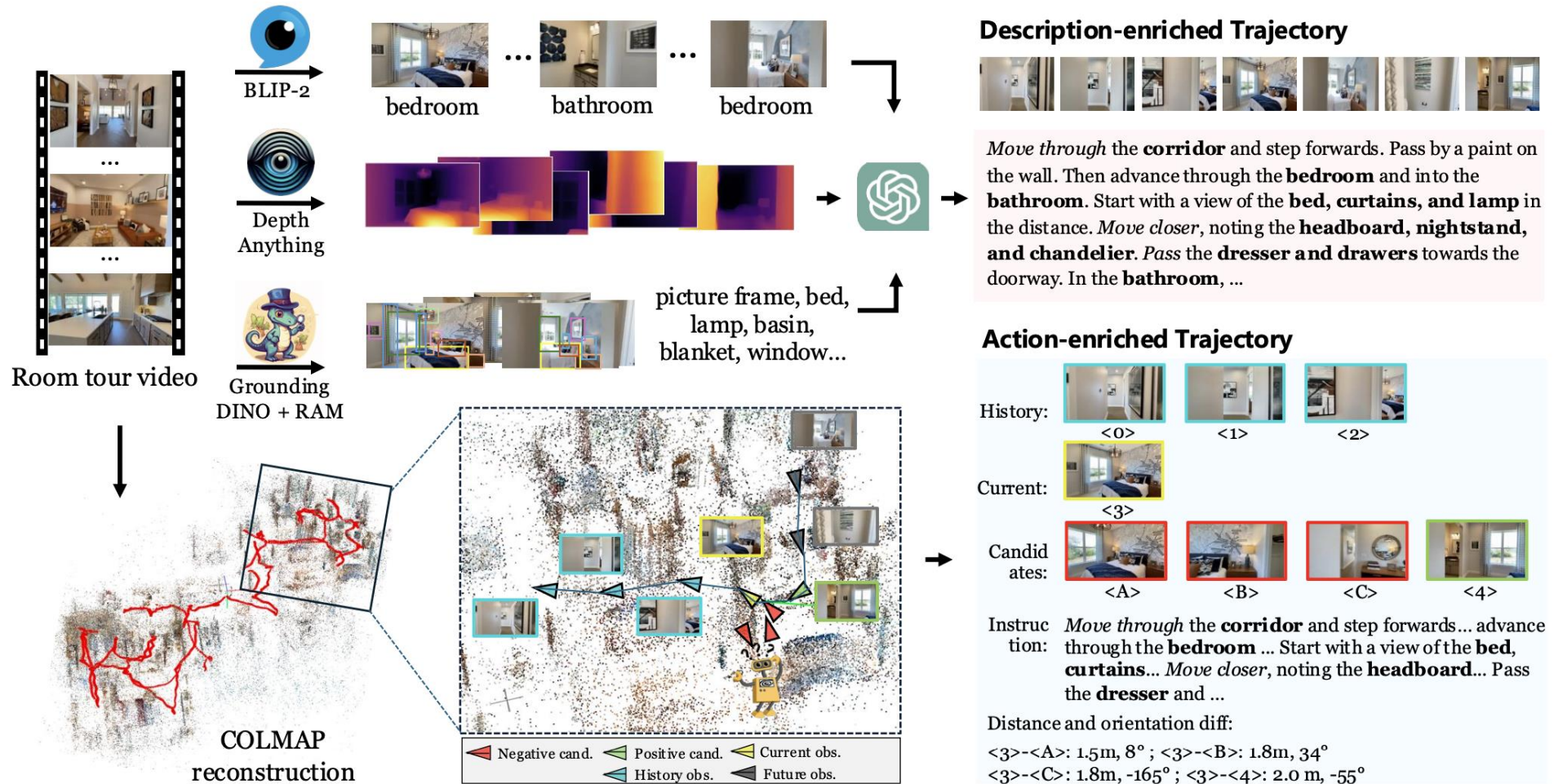


Web videos



3D scenes with human walking trajectories

# Pipeline Overview



RoomTour3D – Our automatic annotation pipeline

# Data collection – YouTube Videos

- **Video Collection:** Selecting high-quality, continuous real-world room tour videos.



## **Length & Continuity:**

Videos longer than 3 minutes to ensure rich, detailed captures.

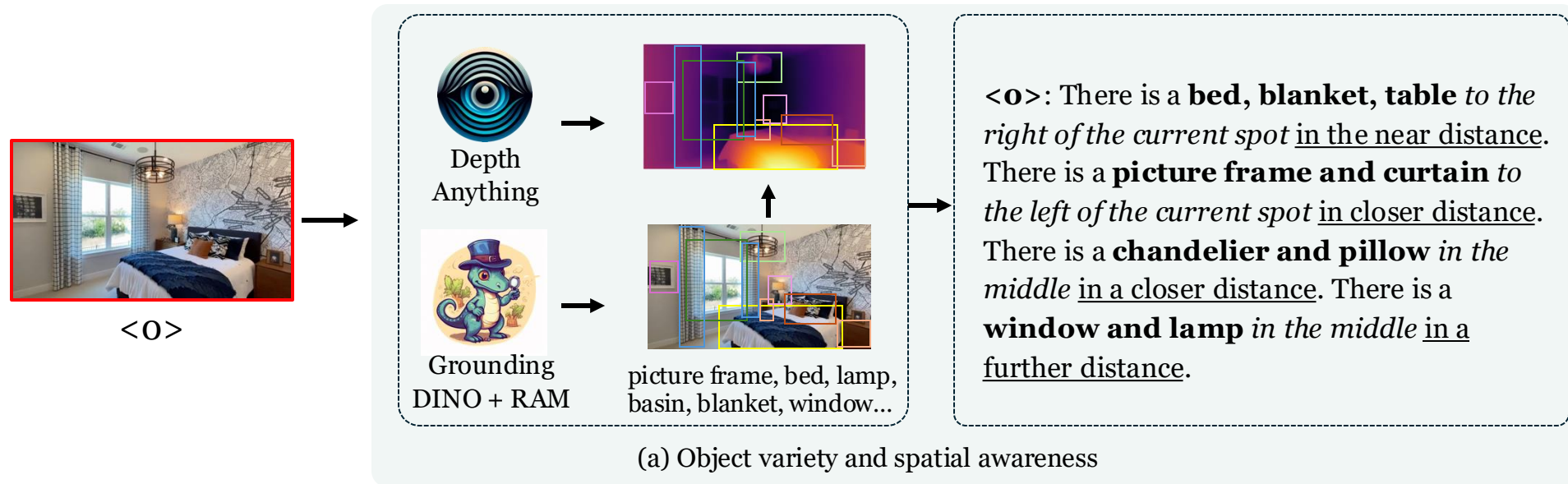
## **Minimal Shot Transitions:**

Prioritize smooth, continuous videos without abrupt cuts or frequent transitions for better reconstruction.

**1847 videos**, totaling approximately **243 hours** of diverse indoor footage.



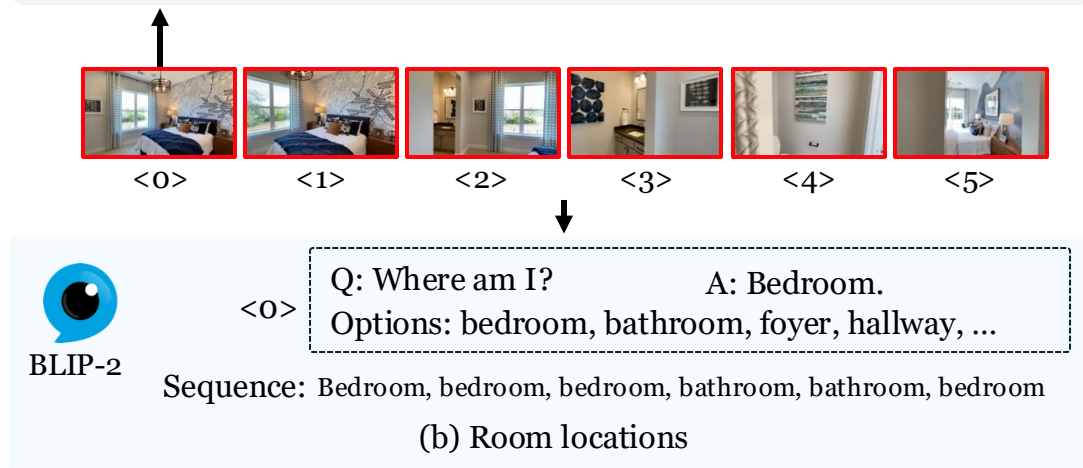
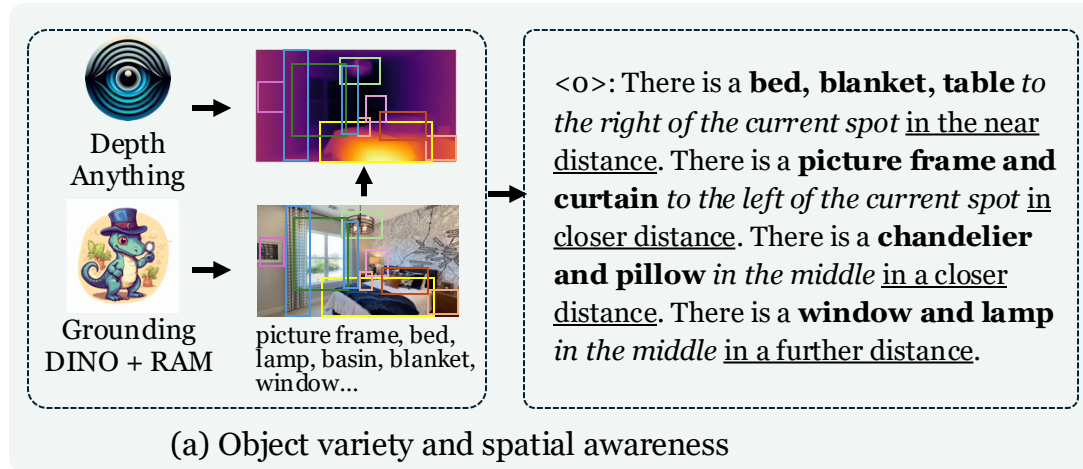
# Description-Enriched Trajectories



Single frame spatial-aware description

- We first get **open-vocabulary object** tags.
- Then, objects are detected using Grounding-DINO and enhanced with **spatial arrangement description**.
- Spatial context is enriched with **depth estimations** (Depth-Anything).

# Description-Enriched Trajectories



**Task Instruction:**  
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.

**In-context Examples:**

**Example 1:**  
<0>: In the study. there is a plant...to the left of the current spot...  
<1>: In the study. there is a bookshelf to the left of the current spot...  
<2>: In the hallway. there is a door to the left of the current spot...  
Your moving trajectory description: Exit the study. Move from left to right, start near plant, laptop, and table, pass a bookshelf..

**Example 2:**  
...

**Your turn:**  
<0>: In the **bedroom**, there is a **bed, blanket, table** to the right of the current spot...  
...  
<5>: In the **bedroom**, here is a **wall** to the left of the current spot in near distance...  
Your moving trajectory description:

(c) Controllable Instruction Generation



# D.E.T. Quality Validation

- Instruction validity check



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.



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.

Score ranks from 1 (totally irrelevant) to 4 (perfect match). The visual sequence of frames corresponds to the GPT-4 generated instruction, evaluated manually to ensure alignment, coherence, and correctness. Our annotation achieves an average relevance score of **3.08** out of 4.

# Action-Enriched Trajectories

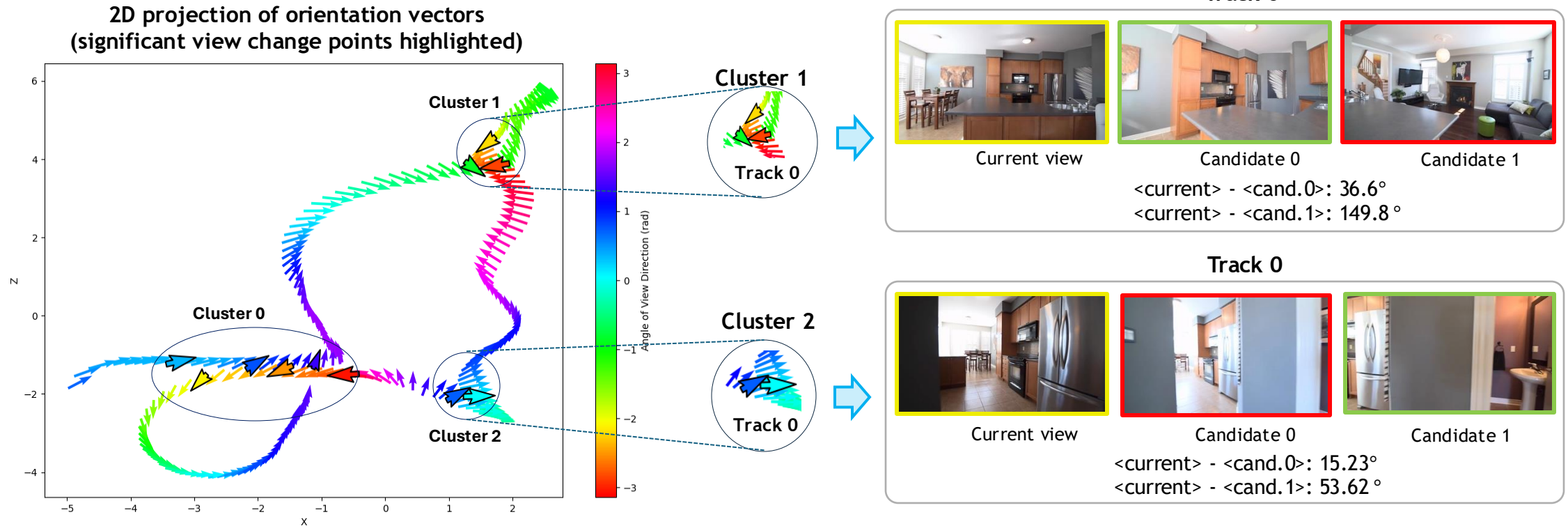


Illustration of action-enriched trajectories generation in RoomTour3D. Significant viewpoint-change points are identified, clustered, and annotated with navigable candidate frames. **Positive candidates**, **negative candidates**.



# A.E.T. Quality Validation

- Action reasonability check



Action A



Action B



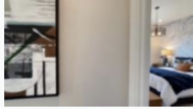



Action C

We checked the spatial proximity and viewpoint diversity of 100 random selected action points – 87% satisfied our criteria

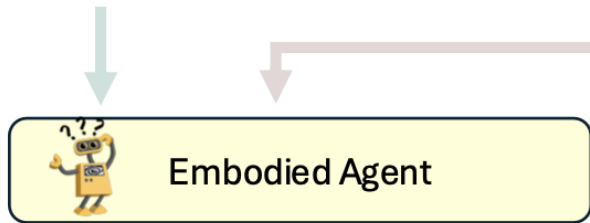
# Training Tasks

(a) Pretraining with Summarization Task

**Cand:**  ...  ...  ... 

**Task:** Describe the camera movement by listing the objects that disappear from view as it pans in one direction.

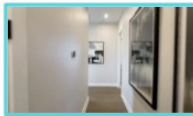

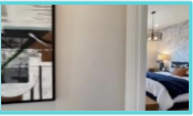
**Output:**  
Move through the corridor and step forwards. Pass by a paint on the wall. Then advance through the **bedroom** and into the **bathroom**. Start with a view of the **bed, curtain..**



**Output:**  
<View 4>





← Summarization  
← Navigation

(b) Finetuning with Navigation Task

**<hist>:**   

**Inst.:** Move through the corridor and step forwards... advance through the bedroom ... Start with a view of the bed, curtains... Move closer, noting the headboard... Pass the dresser and ...

**Task:** Compare the History and Instruction to infer your current progress, and then select the correct direction from the candidates to go to the target location.

**<cand>:**  1.5m, 8° **View A**  1.8m, 34° **View B**  1.8m, -165° **View C**  2.0m, -55° **View 4**

Model training diagram with RoomTour3D. We design two tasks for our RoomTour3D to boost NaviLLM.





# Experiments

Table 1. Overall comparison with the baseline methods. Our RoomTour3D data can boost NaviLLM by a margin on SOON, R2R and REVERIE on SPL metric and on CVDN GP metric. \*denotes reproduced results. RT3D<sub>Desc</sub> and RT3D<sub>Action</sub> stand for description-enriched trajectories only and action-enriched trajectories.

| Methods                                    | CVDN        |             | SOON        |             | R2R         |             | REVERIE     |             |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|  | Val-U       | Test        | Val-U       | Test        | Val-U       | Test        | Val-U       | Test        |
| <i>Models Focusing on Single Task</i>      |             |             |             |             |             |             |             |             |
| PREVALENT [16]                             | 3.15        | 2.44        | -           | -           | 53          | 51          | -           | -           |
| HOP [43]                                   | 4.41        | 3.24        | -           | -           | 57          | 59          | 26.1        | 24.3        |
| HAMT [6]                                   | 5.13        | 5.58        | -           | -           | 61          | 60          | 30.2        | 26.7        |
| DUET [7]                                   | -           | -           | 22.6        | 21.4        | 60          | 58          | 33.7        | 36.0        |
| VLN-SIG [25]                               | 5.52        | 5.83        | -           | -           | 62          | 60          | -           | -           |
| VLN-PETL [44]                              | 5.69        | 6.13        | -           | -           | 60          | 58          | 27.7        | 26.7        |
| NavGPT2 [65]                               | -           | -           | -           | -           | 61          | 60          | -           | -           |
| BEV-BERT [1]                               | -           | -           | -           | -           | <b>64</b>   | 60          | 36.4        | <b>36.4</b> |
| <i>Unified Model For All Tasks</i>         |             |             |             |             |             |             |             |             |
| NaviLLM(w. Pretrain) [63]                  | 6.16        | <b>7.90</b> | 29.2        | 26.3        | 59          | 60          | 35.7        | 32.3        |
| NaviLLM(w. Pretrain)*                      | 6.09        | -           | 28.0        | -           | 56.7        | -           | 31.4        | -           |
| NaviLLM+RT3D <sub>Desc</sub> (Ours)        | <b>6.96</b> | <u>7.55</u> | <u>30.2</u> | <u>26.5</u> | 62.3        | <u>61.8</u> | 37.1        | 35.1        |
| <b>NaviLLM+RT3D<sub>Action</sub>(Ours)</b> | <u>6.33</u> | <u>7.22</u> | <b>31.7</b> | <b>27.8</b> | <u>62.4</u> | <b>62.2</b> | <b>37.4</b> | <b>36.4</b> |



# Ablations

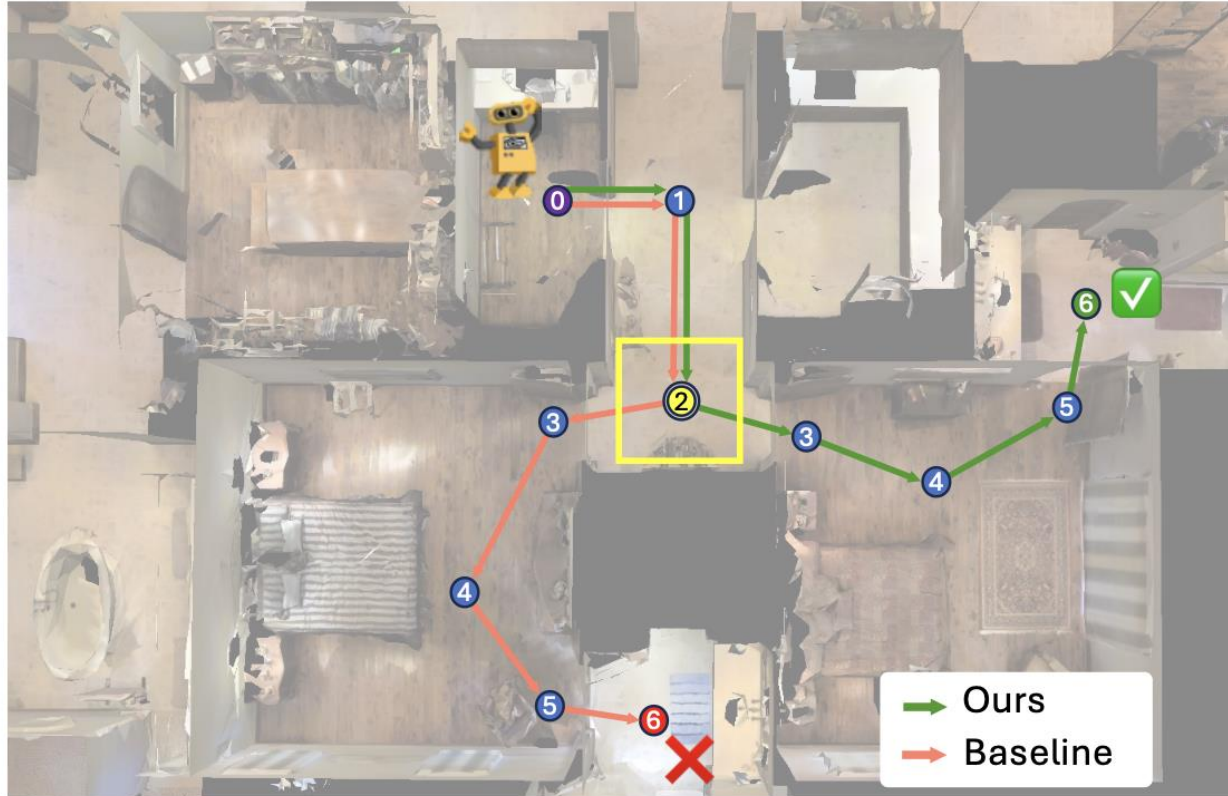
Table 2. Ablation study on the input modalities for trajectory summarization task.

| Object tags | Depth & Bounding Box | Room type | CVDN        | SOON         |              | R2R          |              | REVERIE      |              |
|-------------|----------------------|-----------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|
|             |                      |           | GP↑         | SR↑          | SPL↑         | SR↑          | SPL↑         | SR↑          | SPL↑         |
| ×           | ×                    | ×         | 6.09        | 33.64        | 28.01        | 65.52        | 56.67        | 38.32        | 31.35        |
| ✓           | ×                    | ×         | 5.41        | 32.52        | 26.51        | 63.61        | 55.76        | 42.52        | 34.37        |
| ✓           | ✓                    | ×         | 6.49        | 37.62        | <b>30.40</b> | 68.37        | 61.70        | 41.72        | 36.04        |
| ✓           | ✓                    | ✓         | <b>6.96</b> | <b>38.80</b> | 30.21        | <b>69.37</b> | <b>62.28</b> | <b>43.25</b> | <b>37.10</b> |

Ablation study showing the impact of different input modalities (Object tags, Depth & Bounding Box, and Room type annotations) on navigation performance.

# Visualization

**Instruction:** Exit the sewing room. Turn right. Go toward the glass cabinet with the dolls in it. Turn into the doorway on the left. Pass the bed and go through the next doorway on the left into the bathroom. Wait by the sink. (Instruction\_id : 4676\_0)



Step 0



⋮

Step 2



⋮

Step 4



⋮

Step 6







جامعة محمد بن زايد  
للذكاء الاصطناعي  
MOHAMED BIN ZAYED UNIVERSITY  
OF ARTIFICIAL INTELLIGENCE



# Thank you

## Q & A

<https://roomtour3d.github.io/>