Human-Aware Vision-and-Language Navigation: Bridging Simulation to Reality with Dynamic Human Interactions

Minghan Li^1∗ Heng Li^1∗ Zhi-Qi Cheng^1† Yifei Dong^2
Yuxuan Zhou^3 Jun-Yan He^4 Qi Dai^5 Teruko Mitamura^1 Alexander G. Hauptmann^1
1^Carnegie Mellon University 2^Columbia University
3^University of Mannheim 4^Alibaba Group 5^Microsoft Research

Project Page: https://lpercc.github.io/HA3D_simulator/

Introduction to Human-Aware Vision-and-Language Navigation (HA-VLN)

Point 1: *Vision-and-Language Navigation (VLN) Goals*

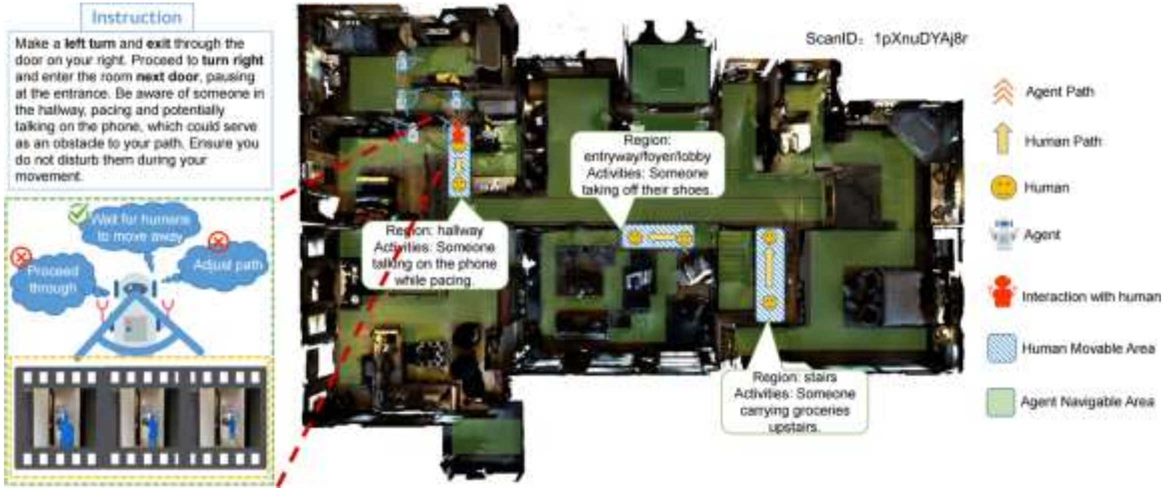
- *Description*: VLN develops embodied agents that navigate environments based on human instructions.

Point 2: *Limitations in Traditional VLN Frameworks*

- *Description*: Current VLN systems depend on static environments and optimal expert supervision, limiting real-world transferability.

Point 3: *Objective of HA-VLN*

- *Description*: HA-VLN aims to bridge the gap between simulation and reality by incorporating dynamic human activities, making navigation more applicable to real-world environments.



Instruction

Make a **left turn** and **exit** through the door on your right. Proceed to **turn right** and enter the room **next door**, pausing at the entrance. Be aware of someone in the hallway, pacing and potentially talking on the phone, which could serve as an obstacle to your path. Ensure you do not disturb them during your movement.

✓ Wait for humans to move away
☒ Proceed through
☒ Adjust path
ScanID: 1pXnuDYAj8r
Region: hallway Activities: Someone talking on the phone while pacing.
Region: entryway/foyer/lobby Activities: Someone taking off their shoes.
Region: stairs Activities: Someone carrying groceries upstairs.
Agent Path
Human Path
Human
Agent
Interaction with human
▨ Human Movable Area
🟩 Agent Navigable Area

Challenges in VLN and the Sim2Real Transfer Gap

Point 1: *Static Environments*

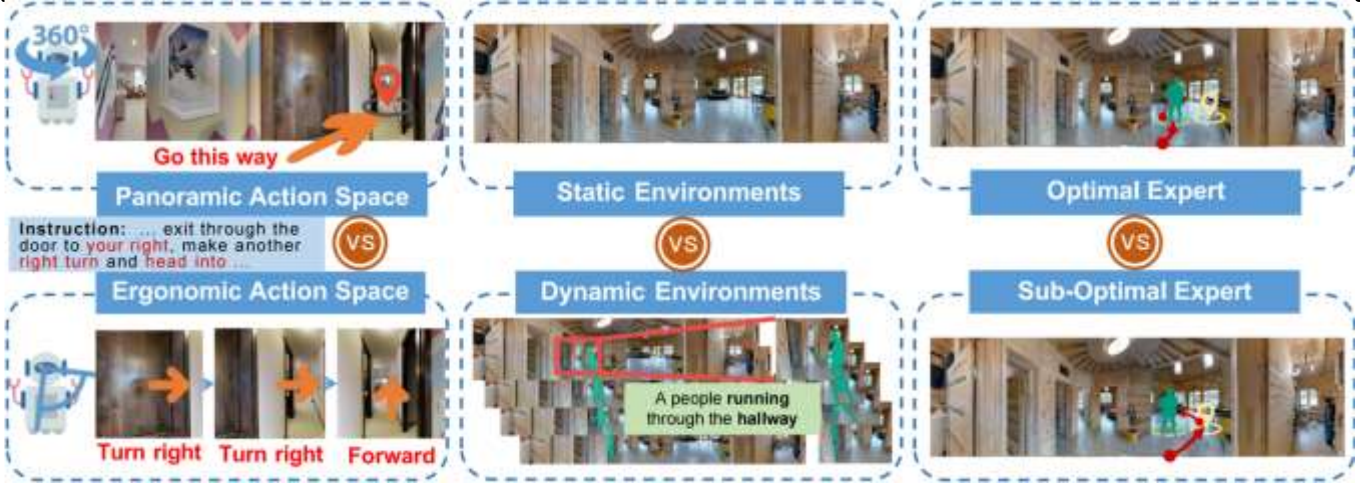
- *Description*: Traditional VLN agents operate in fixed settings, missing dynamic elements seen in real-world environments.

Point 2: *Panoramic Action Spaces*

- *Description*: Agents have an unrealistic 360° view, unlike human-limited vision, making Sim2Real transfer difficult.

Point 3: *Reliance on Optimal Expert Supervision*

- *Description*: Heavy dependence on idealized expert instructions limits agent adaptability to less predictable scenarios.



HA-VLN Scenario and the HA3D Simulator

Point 1: *Dynamic Environments in HA-VLN*

- *Description*: HA-VLN integrates 3D human motion models to simulate dynamic, realistic settings.

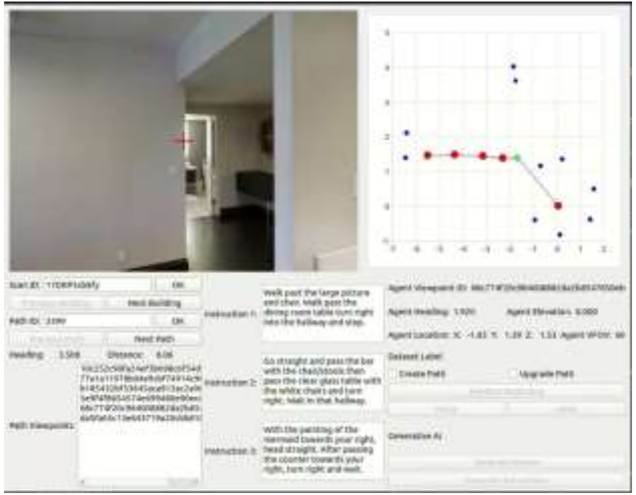
Point 2: *SMPL Model for Human Representation*

- *Description*: Uses the SMPL model to represent human motion and create interactive navigation scenarios for agents.

Point 3: *Objective of HA3D Simulator*

- *Description*: HA3D combines dynamic human activities with the Matterport3D dataset, allowing agents to interact with realistic, populated environments.

Figure: Figure 1 ("HA-VLN scenario") - illustrating agent interaction with dynamic human activities.



Scan ID: 17DRP5sb8fy OK

Next Building

Path ID: 2399 OK

Next Path

Heading: 3.5?? Distance: 0.0?

Path Viewpoints: [illegible]

Instruction 1: Walk past the large picture and chair. Walk past the dining room table turn right into the hallway and stop.

Instruction 2: Go straight and pass the bar with the chair/stools then pass the clear glass table with the white chairs and turn right. Wait in that hallway.

Instruction 3: With the painting of the mermaid towards your right, head straight. After passing the counter towards your right, turn right and wait.

Agent Viewpoint ID: [illegible]

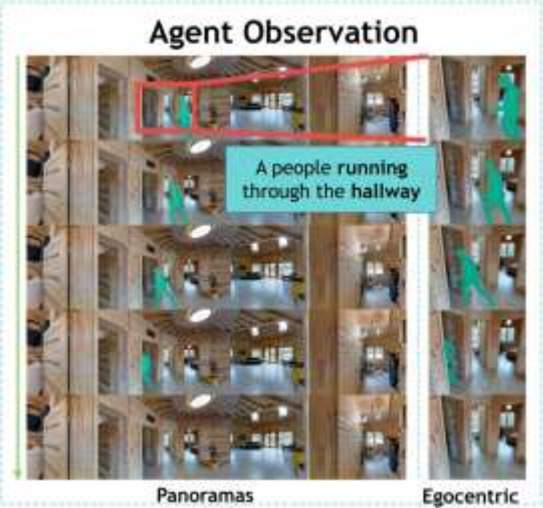
Agent Heading: 1.920 Agent Elevation: 0.000

Agent Location: X: -1.83 Y: 1.29 Z: 1.51 Agent VFOV: 60

Dataset Label:

□ Create Path □ Upgrade Path

Generation AI:



HA3D Simulator and Dataset Annotation Process

Point 1: *Integration of Human Activity and Pose Simulation (HAPS)*

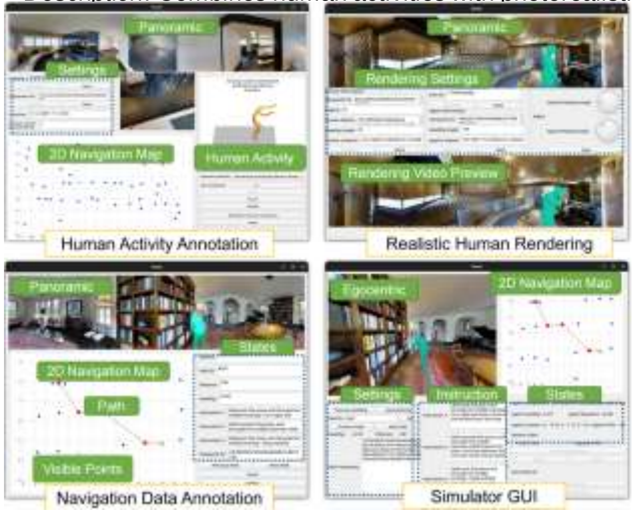
- *Description*: HA3D incorporates the HAPS dataset, offering 145 human activity descriptions and 435 3D human motion models for realistic dynamic environments.

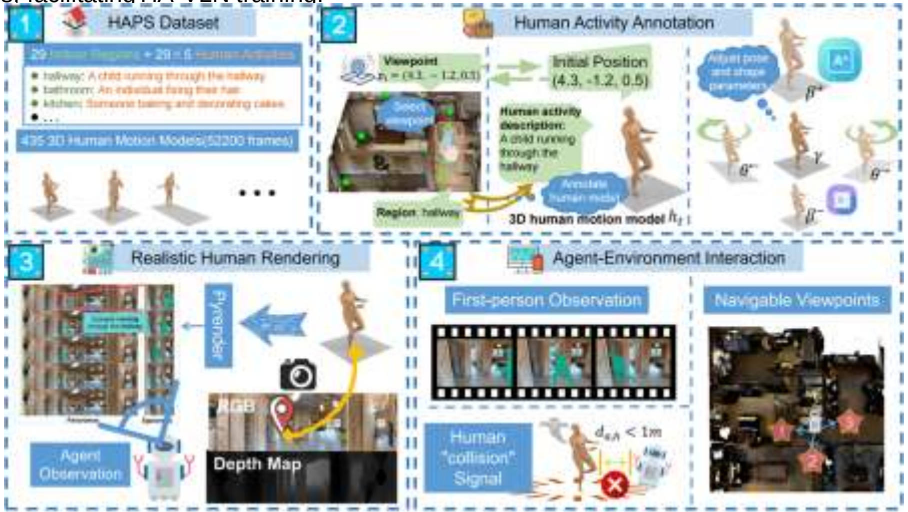
Point 2: *Annotation Tool for Human Models*

- *Description*: The simulator includes a tool for placing human models in 29 indoor areas across 90 scenes, enhancing agent interaction capabilities.

Point 3: *Realistic Environment Rendering*

- *Description*: Combines human activities with photorealistic 3D scenes, facilitating HA-VLN training.





1
HAPS Dataset
29 Indoor Regions × 29 = 6 Human Activities
hallway: A child running through the hallway
bathroom: An individual fixing their hair
kitchen: Someone baking and decorating cakes
...
435 3D Human Motion Models(52200 frames)
• • •
2
Human Activity Annotation
Viewpoint
$p_i = (4.1, -1.2, 0.1)$
Initial Position
(4.3, -1.2, 0.5)
Select viewpoint
Human activity description:
A child running through the hallway
Annotate human motion
Region: hallway
3D human motion model h_i
Adjust pose and shape parameters
β^{+}
β^{-}
γ
3
Realistic Human Rendering
Pyrender
RGB
Depth Map
Agent Observation
4
Agent-Environment Interaction
First-person Observation
Navigable Viewpoints
Human "collision" Signal
$d_{a,h} < 1m$

HA-R2R Dataset and Instruction Analysis

Point 1: *Expansion of Room-to-Room Dataset*

- *Description*: HA-R2R adds human activity descriptions to the original R2R dataset, enriching training diversity.

Point 2: *Expanded Vocabulary and Activity Coverage*

- *Description*: The dataset includes over 21,500 instructions with a broader vocabulary, facilitating more robust training scenarios.

Point 3: *Balanced Instruction Length Distribution*

- *Description*: HA-R2R has more uniform instruction lengths, aiding in balanced and adaptable training.

Figure: Figure 3 (Panels A-C) - displaying the effects of human activities on instruction length, comparing HA-R2R and R2R distributions, and illustrating viewpoints

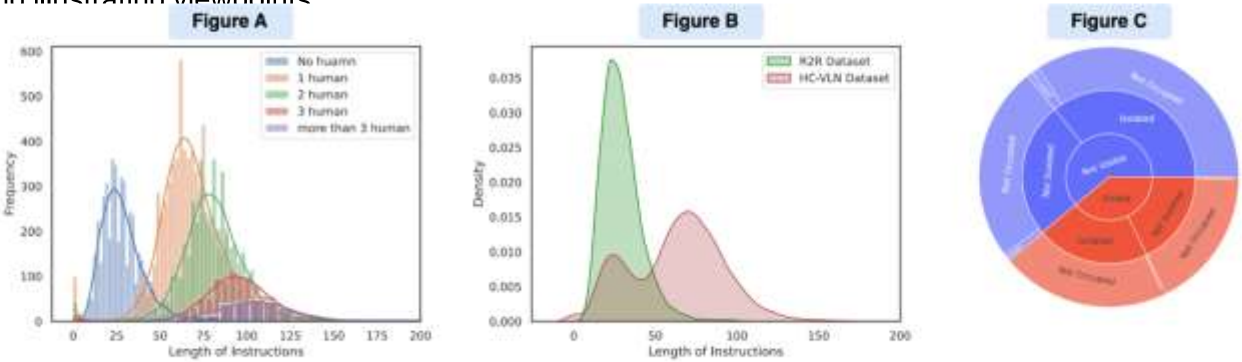


Figure A

Figure B

Figure C

VLN-CM and VLN-DT Agent Models for HA-VLN

Point 1: *Expert-Supervised Cross-Modal (VLN-CM)*

- *Description*: VLN-CM uses expert demonstrations to guide agents through navigation tasks, utilizing a cross-modality fusion module for optimal understanding.

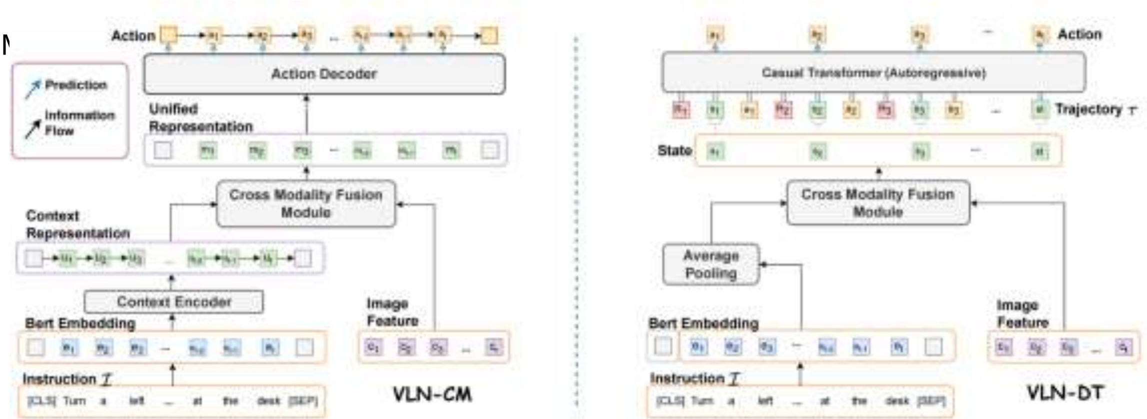
Point 2: *Non-Expert-Supervised Decision Transformer (VLN-DT)*

- *Description*: VLN-DT relies on random trajectories rather than expert supervision, demonstrating strong generalization capability.

Point 3: *Cross-Modality Fusion Module*

- *Description*: Both agents integrate visual and linguistic information dynamically, enabling nuanced navigation actions.

Figure: Figure 4 ("N lity fusion modules.



Action
Prediction
Information Flow
Action Decoder
Unified Representation
Cross Modality Fusion Module
Context Representation
Context Encoder
Bert Embedding
Image Feature
Instruction $\mathcal{I}$
[CLS] Turn a left ... at the desk [SEP]
VLN-CM
Action
Casual Transformer (Autoregressive)
Trajectory τ
State
Cross Modality Fusion Module
Average Pooling
Bert Embedding
Image Feature
Instruction $\mathcal{I}$
[CLS] Turn a left ... at the desk [SEP]
VLN-DT

Training Strategies and Expert Supervision in HA-VLN

- **Point 1**: *Expert vs. Non-Expert Supervision*
  - *Description*: Highlights the adaptability of agents trained without expert guidance, which allows for better generalization in dynamic environments.
- **Point 2**: *Adaptive Policy Development*
  - *Description*: Policies adapted to dynamic conditions in HA-VLN improve navigation flexibility and responsiveness.
- **Point 3**: *Reward Function for Safe Navigation*
  - *Description*: HA-VLN's custom reward function encourages agents to optimize paths while maintaining safe distances in human-populated environments.

Table: Table 1 ("Egocentric vs. Panoramic Action Space Comparison") and Table 2 ("Static vs. Dynamic Environment Comparison"). Table 3 ("Optimal vs. Sub-Optimal Expert Comparison") – comparing performance across various supervision levels.

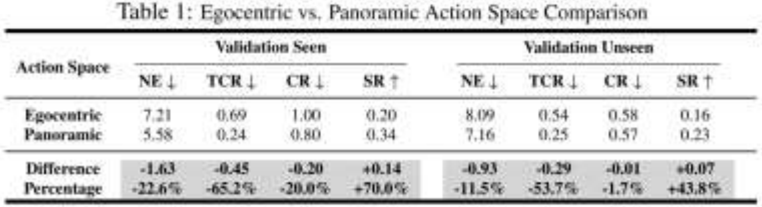


Table 1: Egocentric vs. Panoramic Action Space Comparison

Action Space	Validation Seen				Validation Unseen			
	NE ↓	TCR ↓	CR ↓	SR ↑	NE ↓	TCR ↓	CR ↓	SR ↑
Egocentric	7.21	0.69	1.00	0.20	8.09	0.54	0.58	0.16
Panoramic	5.58	0.24	0.80	0.34	7.16	0.25	0.57	0.23
Difference	-1.63	-0.45	-0.20	+0.14	-0.93	-0.29	-0.01	+0.07
Percentage	-22.6%	-65.2%	-20.0%	+70.0%	-11.5%	-53.7%	-1.7%	+43.8%

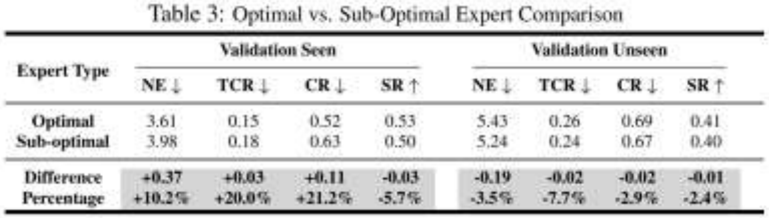


Table 3: Optimal vs. Sub-Optimal Expert Comparison

Expert Type	Validation Seen				Validation Unseen			
	NE ↓	TCR ↓	CR ↓	SR ↑	NE ↓	TCR ↓	CR ↓	SR ↑
Optimal	3.61	0.15	0.52	0.53	5.43	0.26	0.69	0.41
Sub-optimal	3.98	0.18	0.63	0.50	5.24	0.24	0.67	0.40
Difference	+0.37	+0.03	+0.11	-0.03	-0.19	-0.02	-0.02	-0.01
Percentage	+10.2%	+20.0%	+21.2%	-5.7%	-3.5%	-7.7%	-2.9%	-2.4%

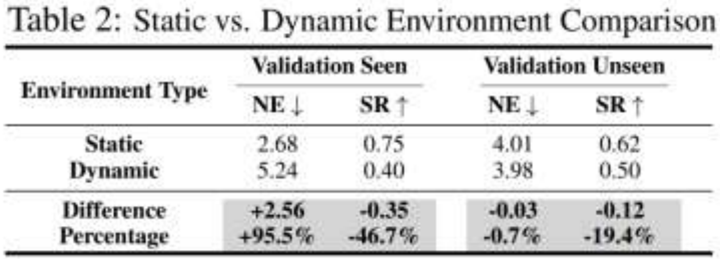


Table 2: Static vs. Dynamic Environment Comparison

Environment Type	Validation Seen		Validation Unseen	
	NE ↓	SR ↑	NE ↓	SR ↑
Static	2.68	0.75	4.01	0.62
Dynamic	5.24	0.40	3.98	0.50
Difference	+2.56	-0.35	-0.03	-0.12
Percentage	+95.5%	-46.7%	-0.7%	-19.4%

Evaluation Metrics and Results for HA-VLN

Point 1: *Human Activity-Aware Metrics*

- *Description*: New metrics focus on interactions with dynamic elements, crucial for assessing Sim2Real transfer quality.

Point 2: *Performance Gap Analysis*

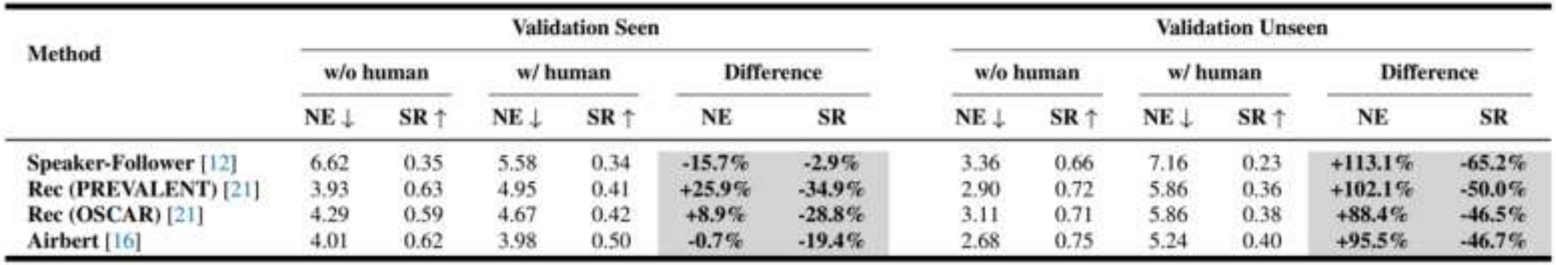
- *Description*: Evaluation reveals substantial gaps between HA-VLN agents and the Oracle, especially in dynamically populated settings.

Point 3: *Comparison with State-of-the-Art (SOTA) Agents*

- *Description*: Retrained SOTA agents show limited success in dynamic HA-VLN environments, underscoring challenges in complex real-world settings.

Table: Table 4 ("Performance of SOTA VLN Agents on HA-VLN (Retrained)") - presenting performance metrics to highlight the current challenges.

Table 4: Performance of SOTA VLN Agents on HA-VLN (Retrained)



Method	Validation Seen						Validation Unseen					
	w/o human		w/ human		Difference		w/o human		w/ human		Difference	
	NE ↓	SR ↑	NE ↓	SR ↑	NE	SR	NE ↓	SR ↑	NE ↓	SR ↑	NE	SR
Speaker-Follower [12]	6.62	0.35	5.58	0.34	**-15.7%**	**-2.9%**	3.36	0.66	7.16	0.23	**+113.1%**	**-65.2%**
Rec (PREVALENT) [21]	3.93	0.63	4.95	0.41	**+25.9%**	**-34.9%**	2.90	0.72	5.86	0.36	**+102.1%**	**-50.0%**
Rec (OSCAR) [21]	4.29	0.59	4.67	0.42	**+8.9%**	**-28.8%**	3.11	0.71	5.86	0.38	**+88.4%**	**-46.5%**
Airbert [16]	4.01	0.62	3.98	0.50	**-0.7%**	**-19.4%**	2.68	0.75	5.24	0.40	**+95.5%**	**-46.7%**