JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Behavioral Analysis of Vision-and-Language Navigation Agents

Zijiao Yang
Oregon State University

Arjun Majumdar
Georgia Institute of Technology

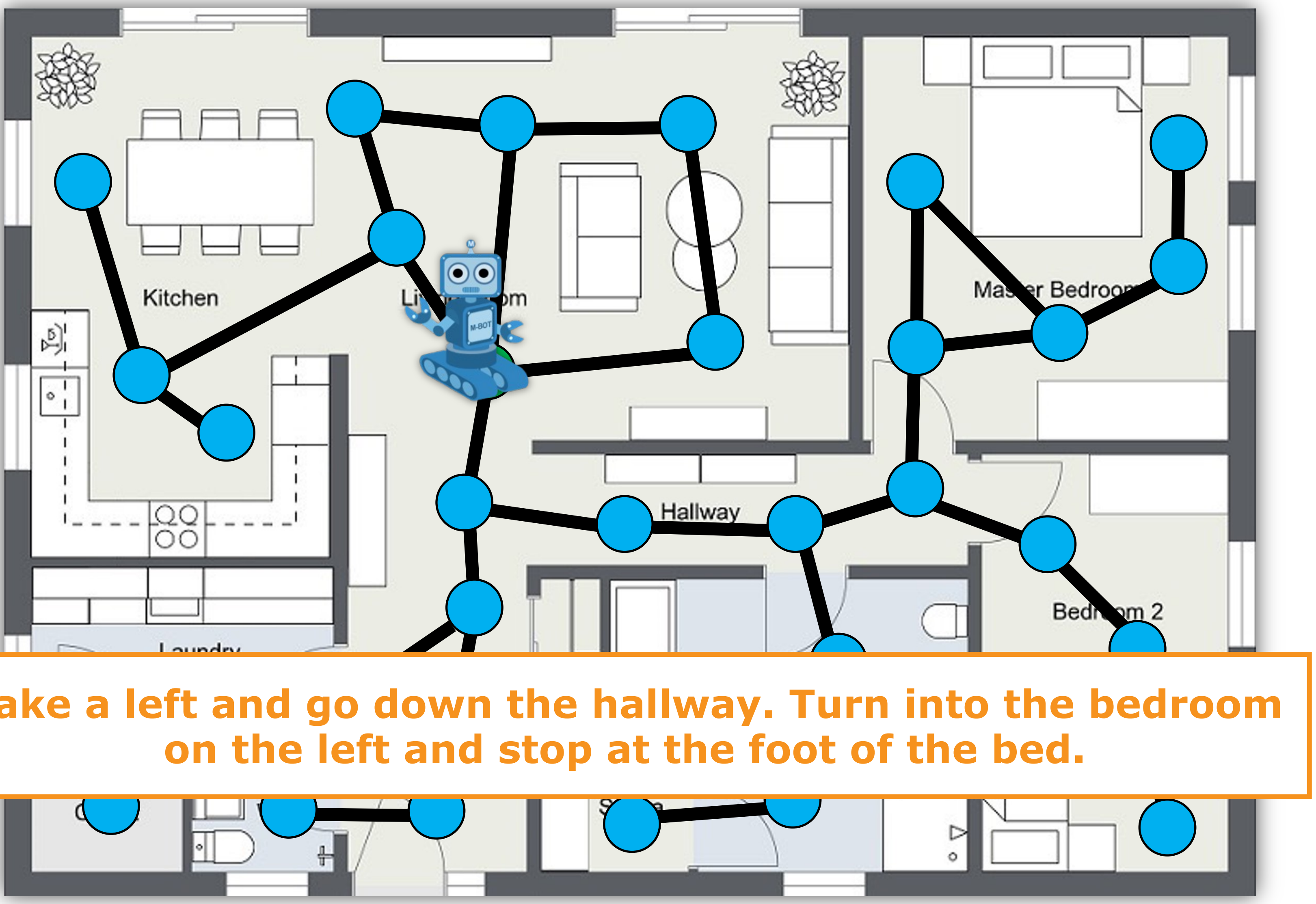
Stefan Lee
Oregon State University

Session id: TUE-AM-245

Oregon State University

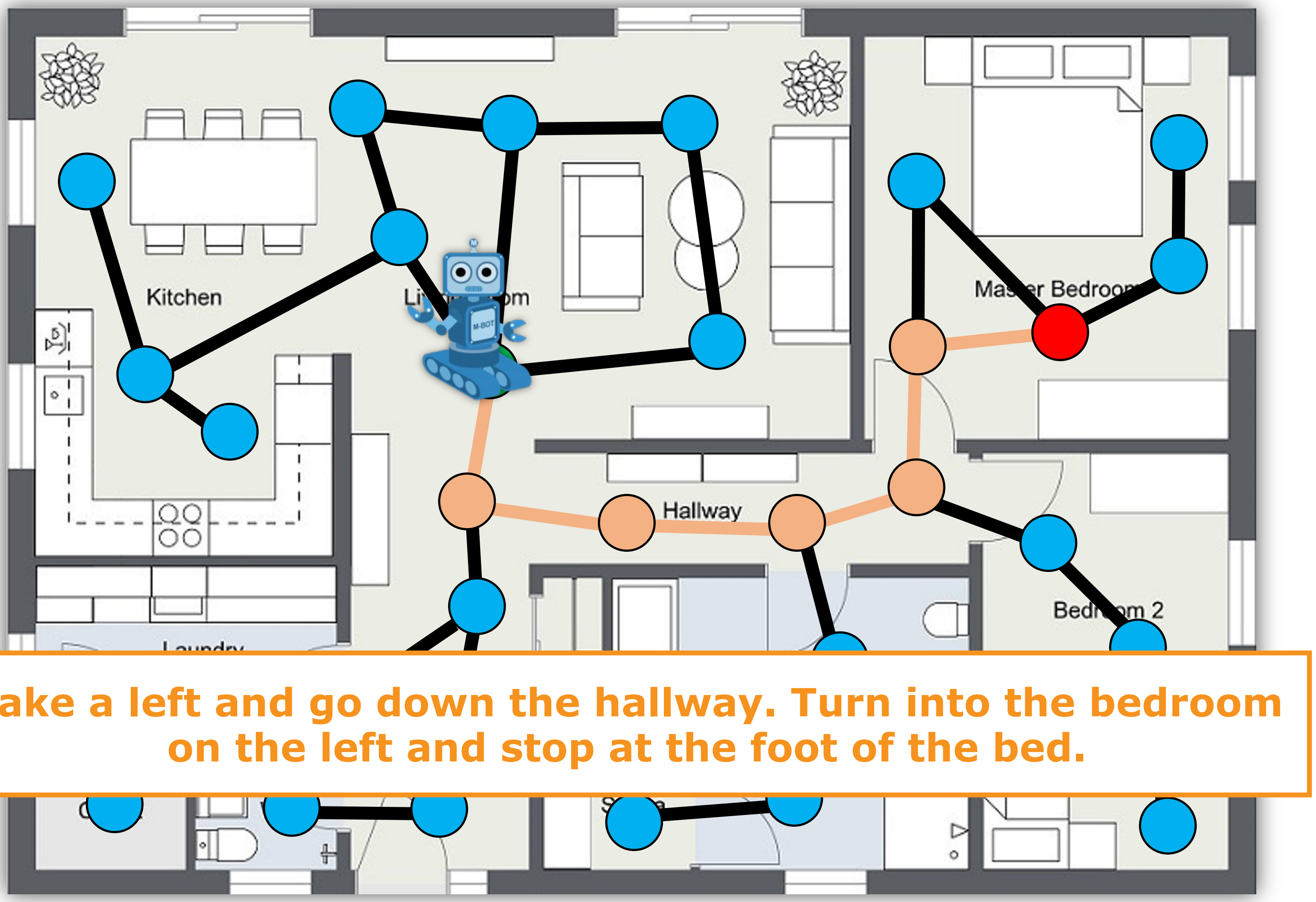
Georgia Tech

Vision-and-Language Navigation (VLN)



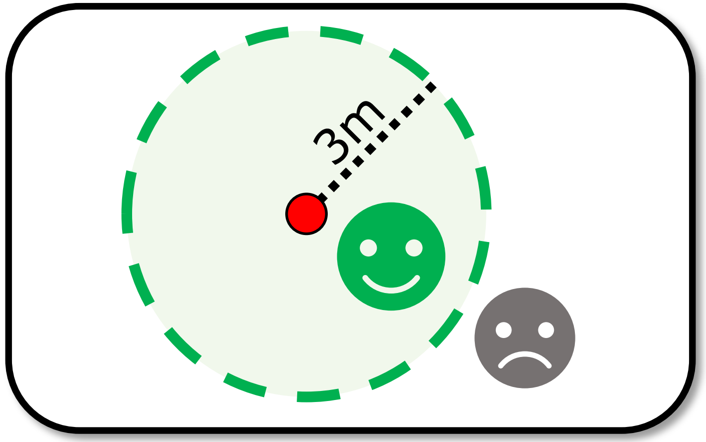
Take a left and go down the hallway. Turn into the bedroom on the left and stop at the foot of the bed.

Vision-and-Language Navigation (VLN)

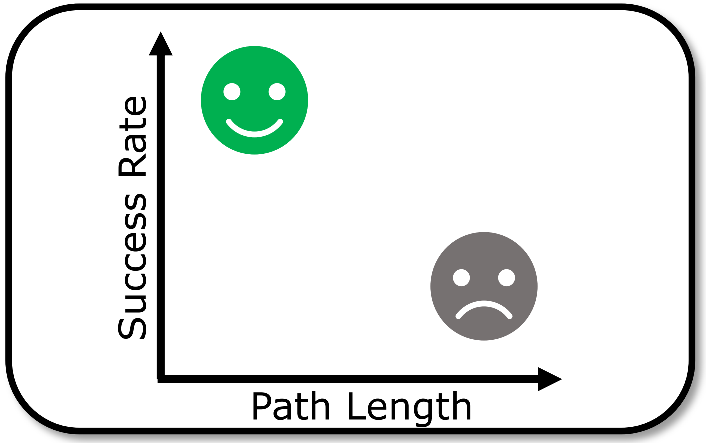


Take a left and go down the hallway. Turn into the bedroom on the left and stop at the foot of the bed.

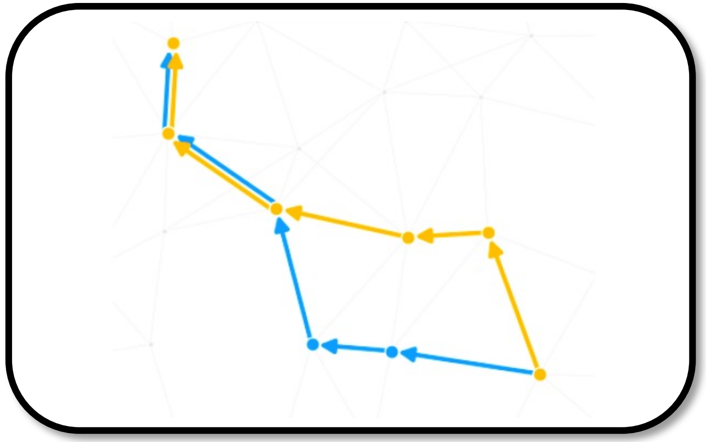
Metrics measure aggregate performance



Success Rate (SR)



Success weighted by Path Length (SPL)



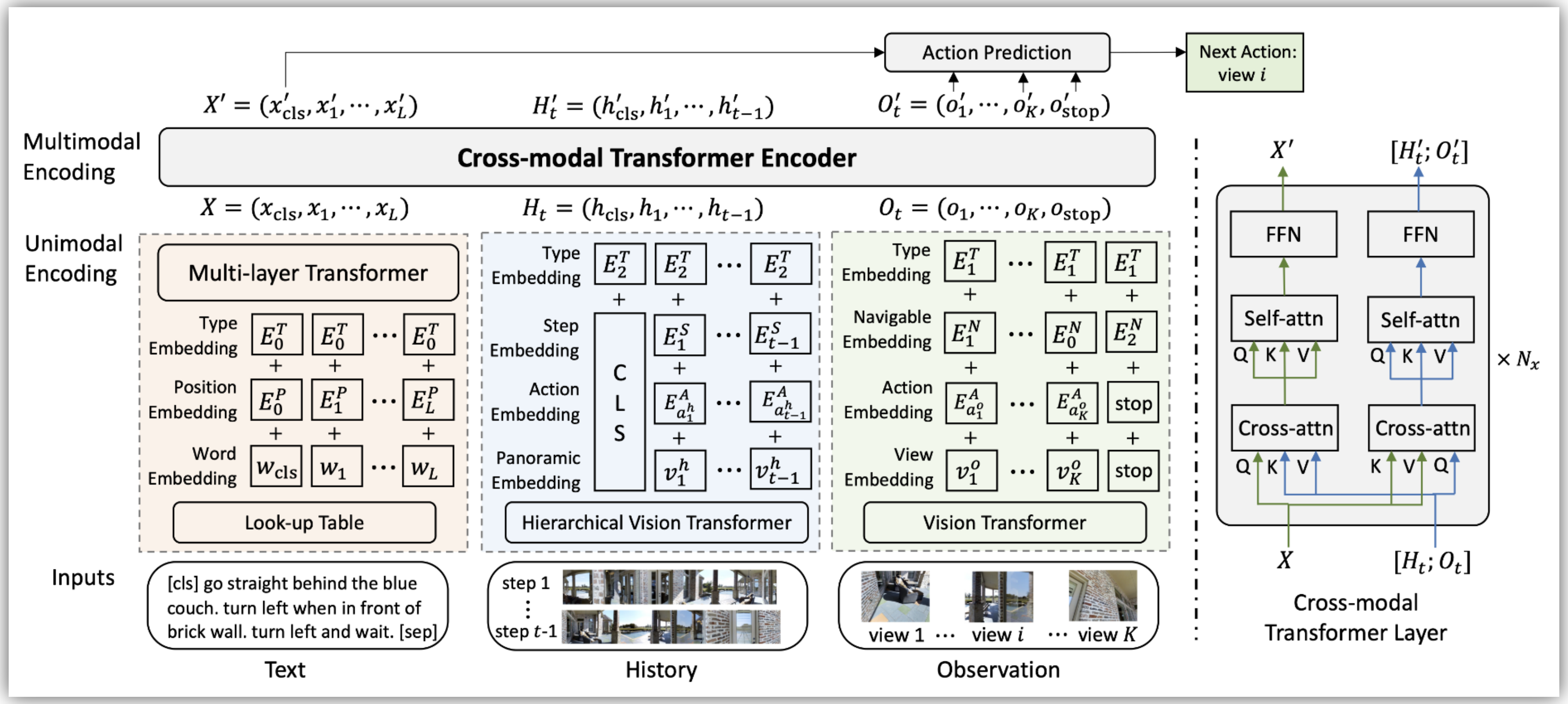
Normalized Dynamic Time Warping (nDTW)

Useful for comparing methods in aggregate but don’t speak to fine-grained agent capabilities.

Examine Four Fine-grained Skill Competencies

Explicit Stop	*“This is your destination”*
Unconditional Directional	*“Make a left”*
Object Seeking	*“Go towards the lamp”*
Room Seeking	*“Go to the kitchen”*

Case Study

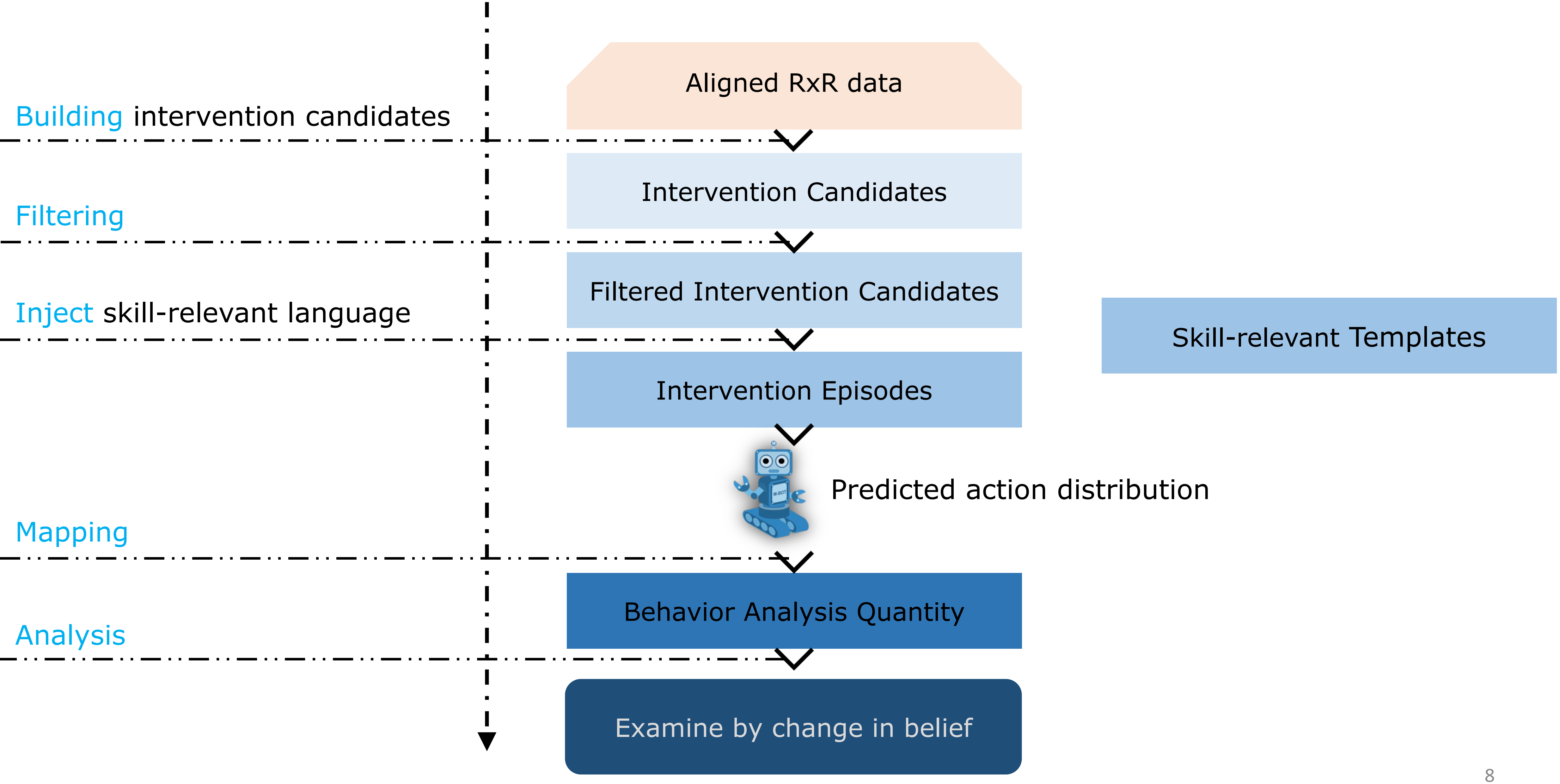

Action Prediction
Next Action: view i
$X' = (x'_{\text{cls}}, x'_1, \cdots, x'_L)$
$H'_t = (h'_{\text{cls}}, h'_1, \cdots, h'_{t-1})$
$O'_t = (o'_1, \cdots, o'_K, o'_{\text{stop}})$
Multimodal Encoding
Cross-modal Transformer Encoder
$X = (x_{\text{cls}}, x_1, \cdots, x_L)$
$H_t = (h_{\text{cls}}, h_1, \cdots, h_{t-1})$
$O_t = (o_1, \cdots, o_K, o_{\text{stop}})$
Unimodal Encoding
Multi-layer Transformer
Type Embedding
E^T_0 E^T_0 $\cdots$ E^T_0
Position Embedding
E^P_0 E^P_1 $\cdots$ E^P_L
Word Embedding
w_{cls} w_1 $\cdots$ w_L
Look-up Table
Type Embedding
E^T_2 E^T_2 $\cdots$ E^T_2
Step Embedding
E^S_1 $\cdots$ E^S_{t-1}
Action Embedding
$E^A_{a^h_1}$ $\cdots$ $E^A_{a^h_{t-1}}$
Panoramic Embedding
v^h_1 $\cdots$ v^h_{t-1}
C L S
Hierarchical Vision Transformer
Type Embedding
E^T_1 $\cdots$ E^T_1 E^T_1
Navigable Embedding
E^N_1 $\cdots$ E^N_0 E^N_2
Action Embedding
$E^A_{a^o_1}$ $\cdots$ $E^A_{a^o_K}$ stop
View Embedding
v^o_1 $\cdots$ v^o_K stop
Vision Transformer
Inputs
[cls] go straight behind the blue couch. turn left when in front of brick wall. turn left and wait. [sep]
Text
step 1
step t-1
History
view 1 ⋯ view i ⋯ view K
Observation
X'
$[H'_t; O'_t]$
FFN
FFN
Self-attn
Self-attn
Q K V
Q K V
Cross-attn
Cross-attn
Q K V
K V Q
$\times N_x$
X
$[H_t; O_t]$
Cross-modal Transformer Layer

History Aware Multimodal Transformer for Vision-and-Language Navigation (Chen et al.)

Summary of Findings

Explicit Stop	*"This is your destination"*	✓
Unconditional Directional	*"Make a left"*	✓
Object Seeking	*"Go towards the lamp"*	?
Room Seeking	*"Go to the kitchen"*	?

Behavioral Analysis Framework



Aligned RxR data

Building intervention candidates

Intervention Candidates

Filtering

Filtered Intervention Candidates

Inject skill-relevant language

Skill-relevant Templates

Intervention Episodes

Predicted action distribution

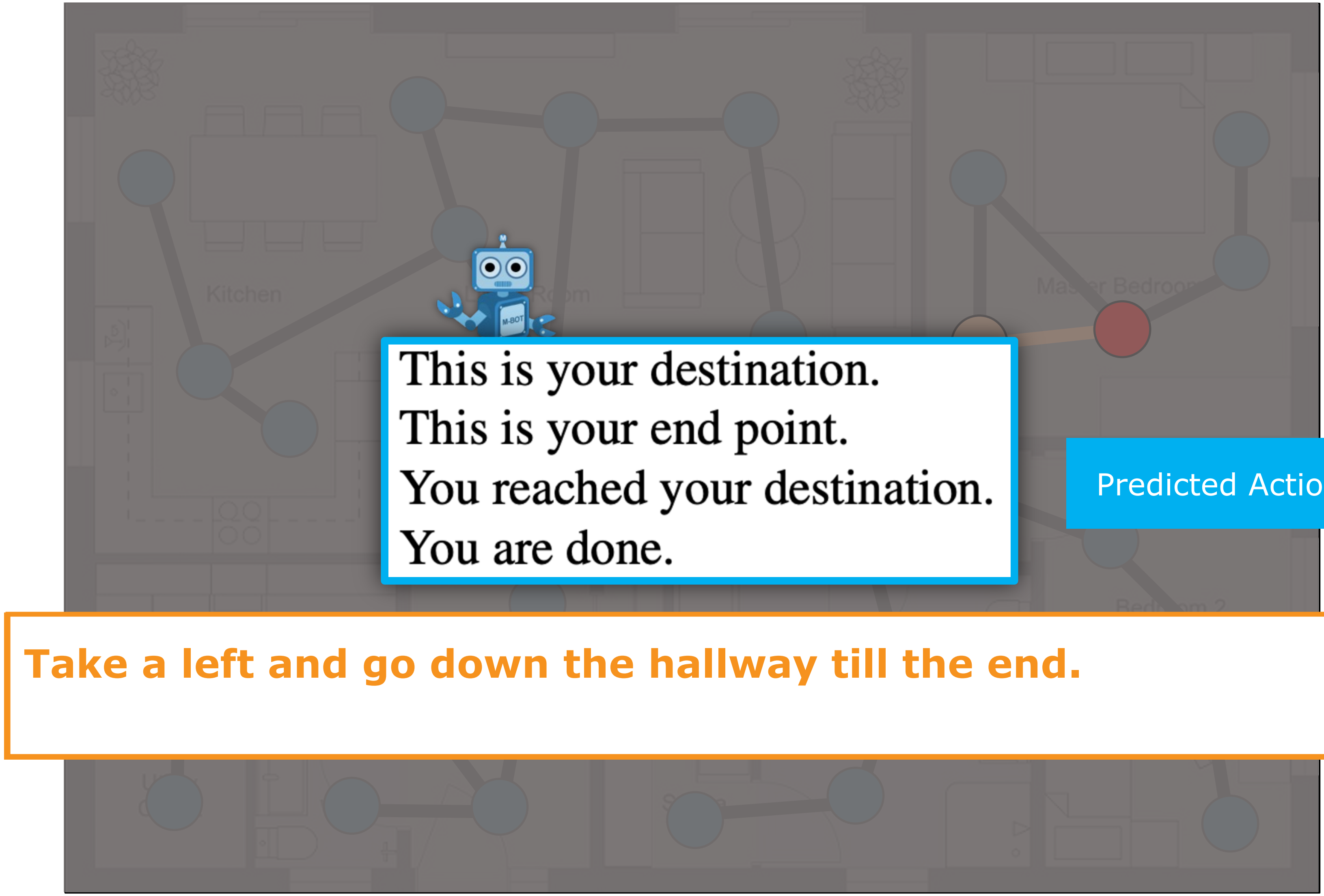
Mapping

Behavior Analysis Quantity

Analysis

Examine by change in belief

HAMT: Stop (Explicit Stop)



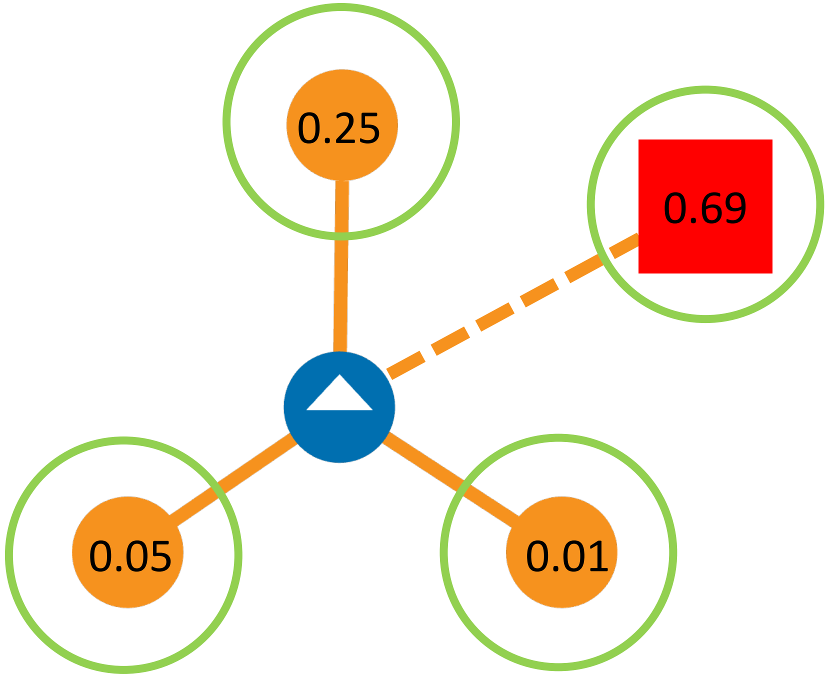
This is your destination.
This is your end point.
You reached your destination.
You are done.

Predicted Action Distribution

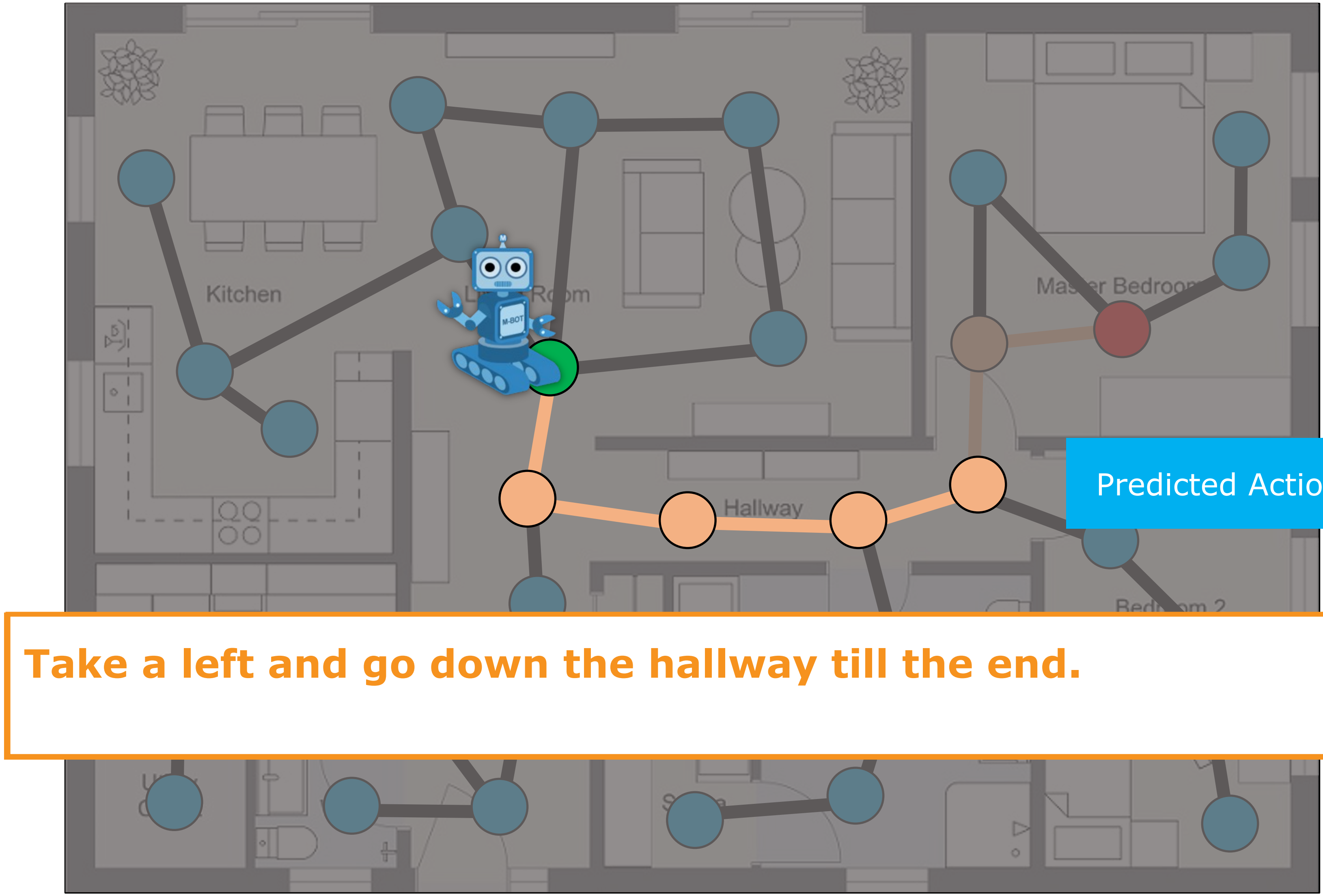
Take a left and go down the hallway till the end.

HAMT: Explicit Stop

Predicted Action Distribution



HAMT: Stop (Implicit Stop)

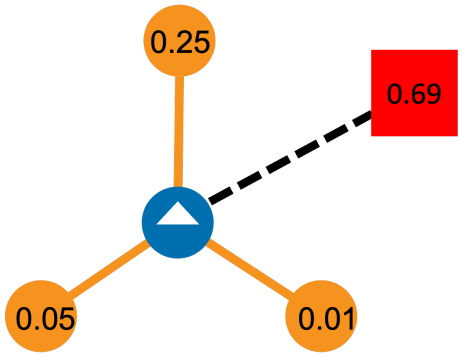


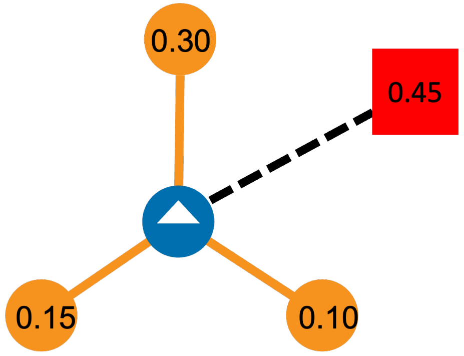
Predicted Action Distribution

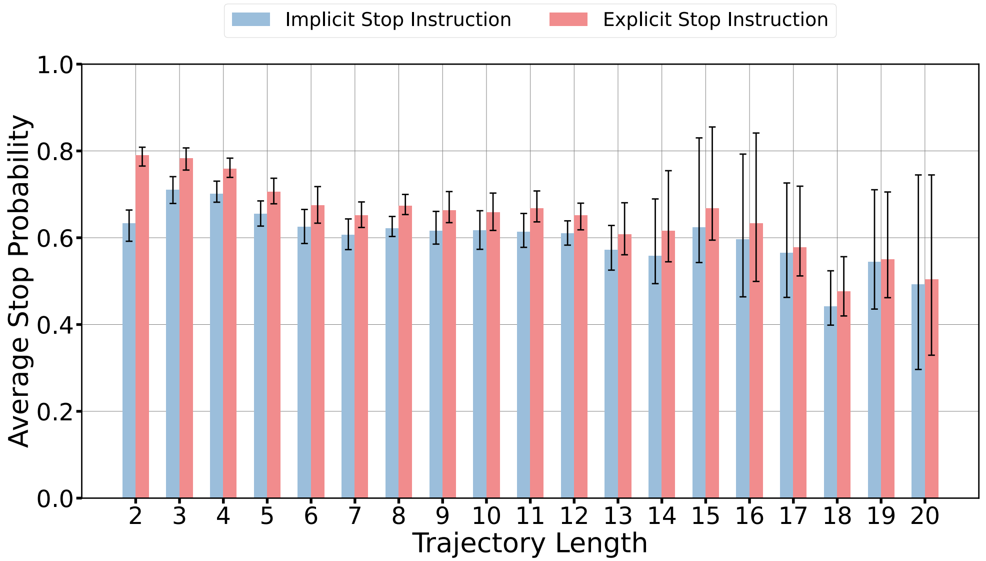
Take a left and go down the hallway till the end.

HAMT: Stop (Analysis)

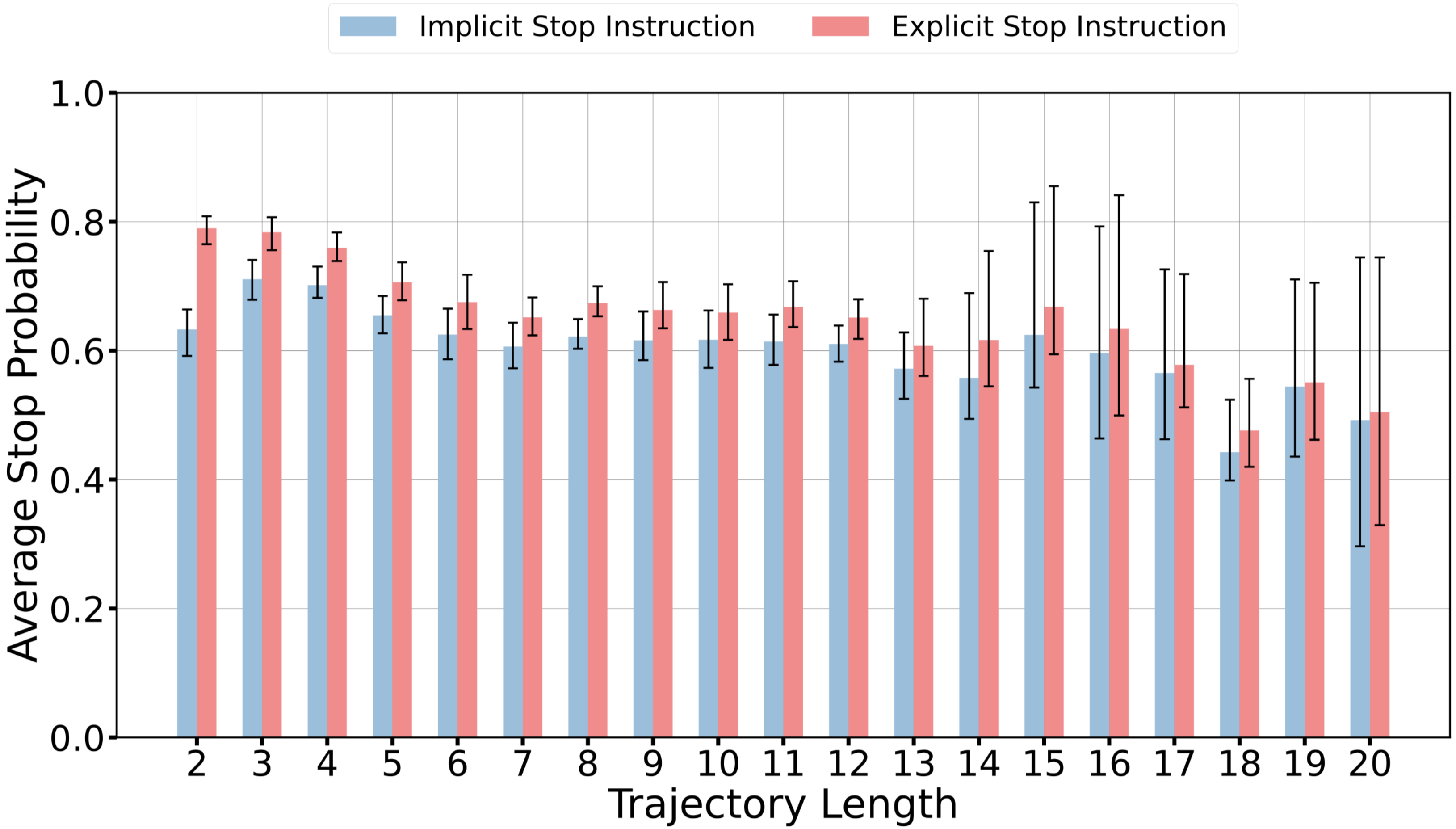
Behavior Analysis Quantity → Stop : 0.69



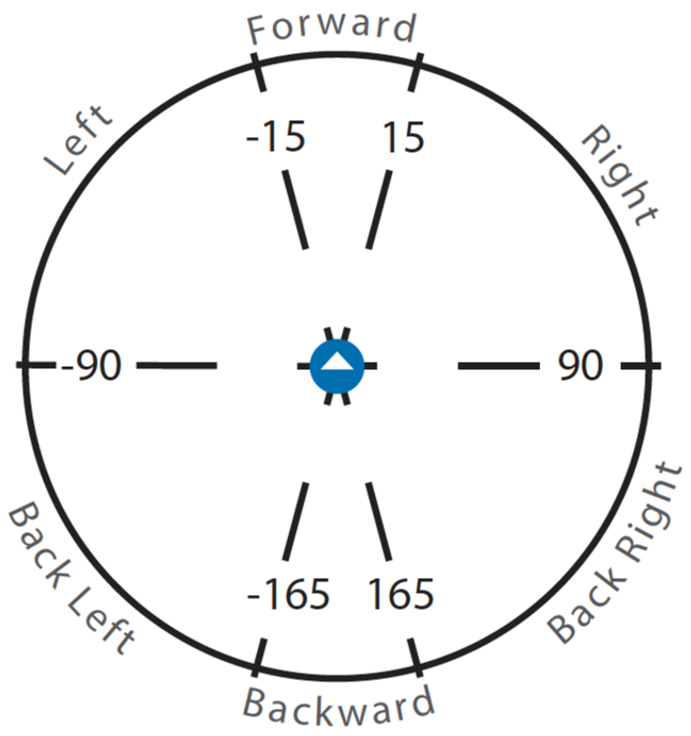




HAMT: Stop (Analysis)



HAMT: Unconditional Directional Instructions (Filtering)



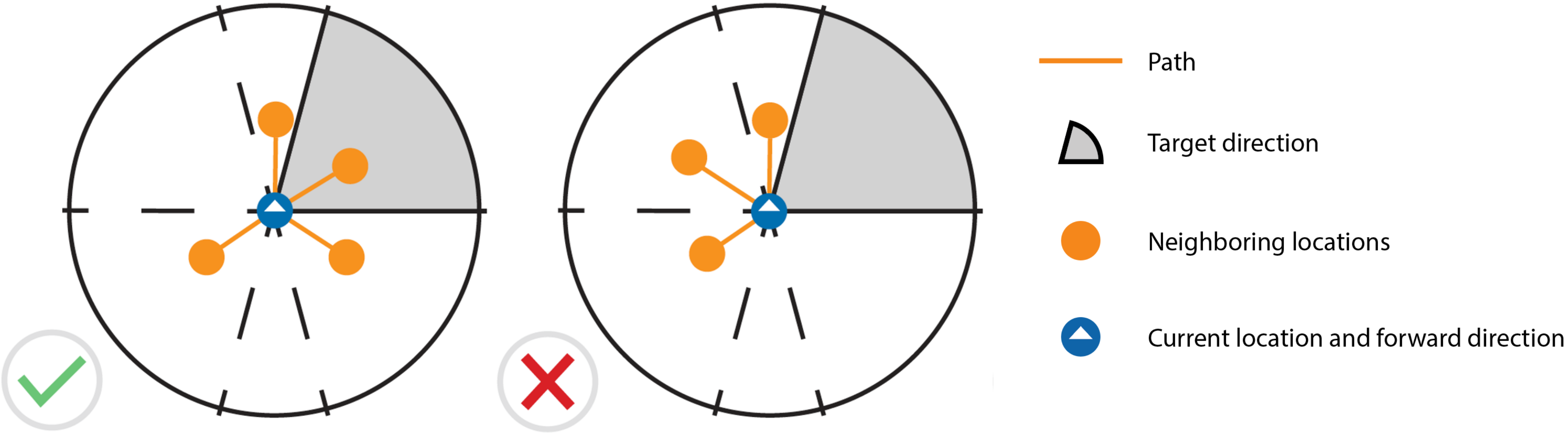
— Path

◭ Target direction

● Neighboring locations

● Current location and forward direction

HAMT: Unconditional Directional Instructions (Filtering)



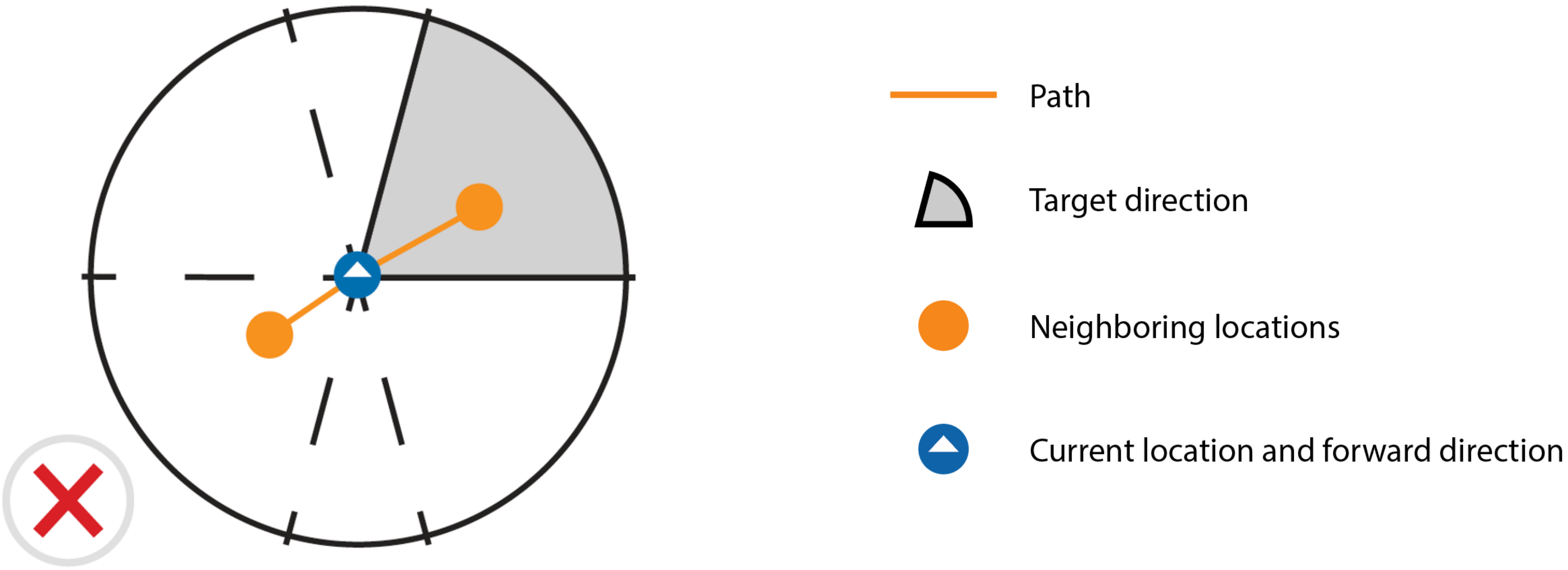
Path

Target direction

● Neighboring locations

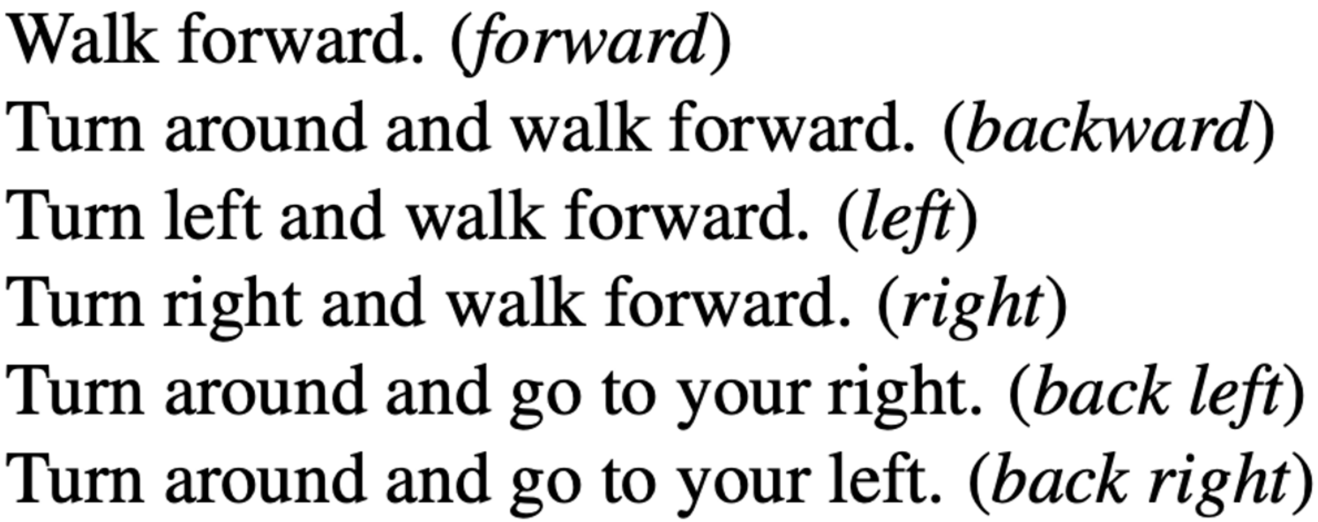
Current location and forward direction

HAMT: Unconditional Directional Instructions (Filtering)



HAMT: Unconditional Directional Instructions (Templates)

Template:



Walk forward. (*forward*)

Turn around and walk forward. (*backward*)

Turn left and walk forward. (*left*)

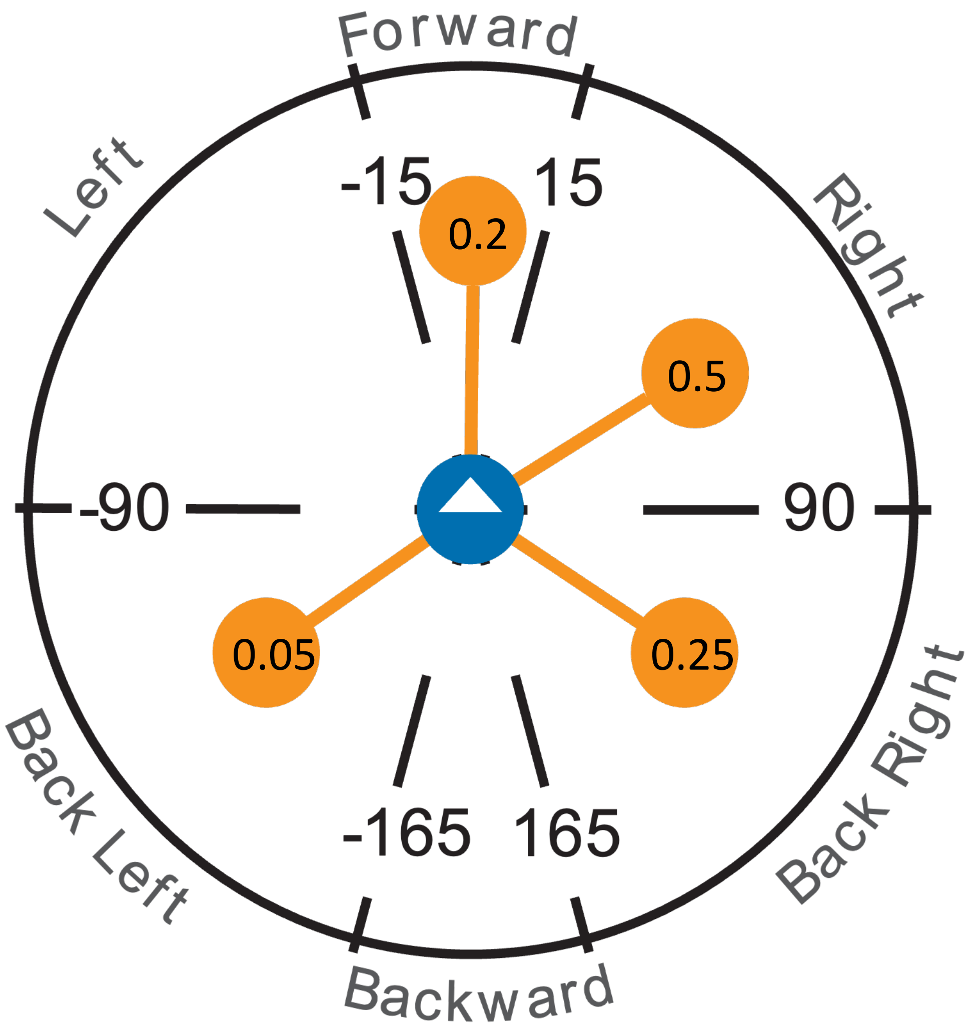
Turn right and walk forward. (*right*)

Turn around and go to your right. (*back left*)

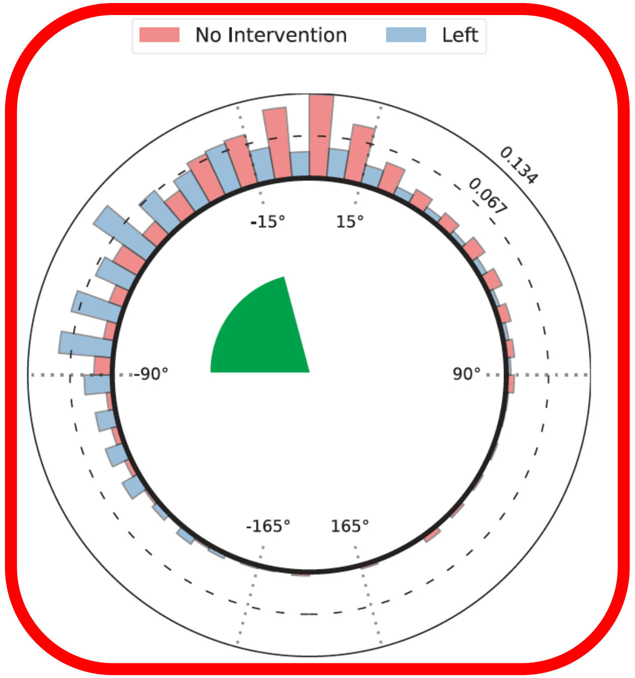
Turn around and go to your left. (*back right*)

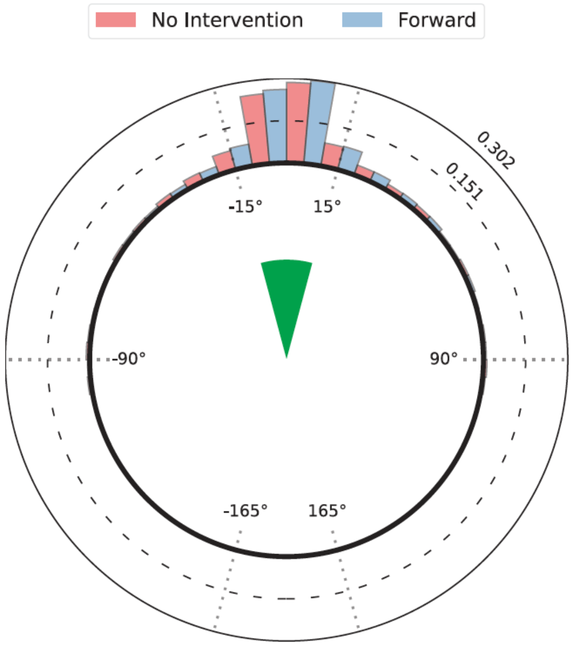
HAMT: Unconditional Directional Instructions (Mapping)

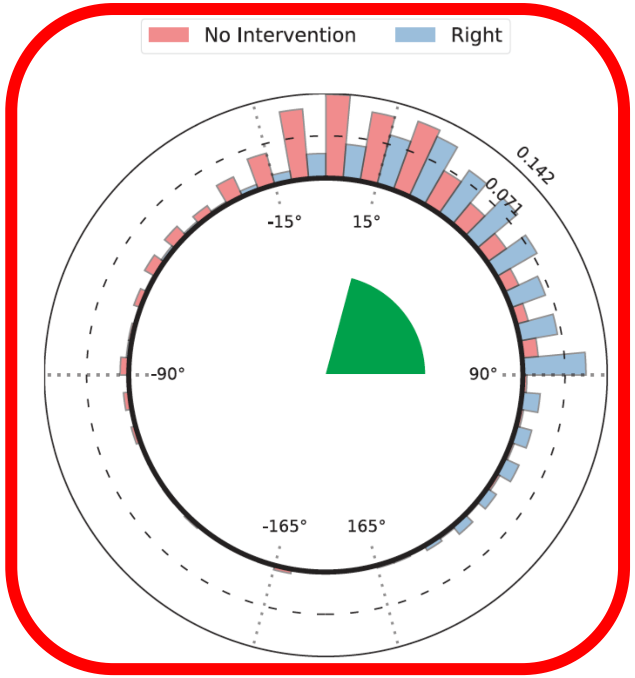
Behavior Analysis Quantity ➡ Angular difference relative to heading

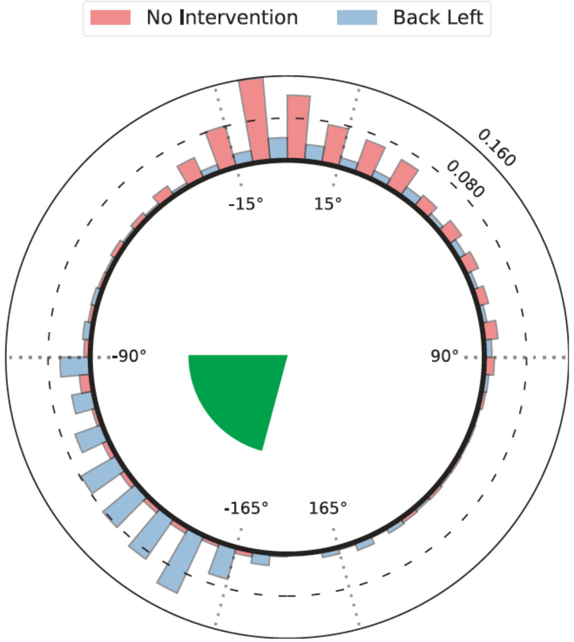


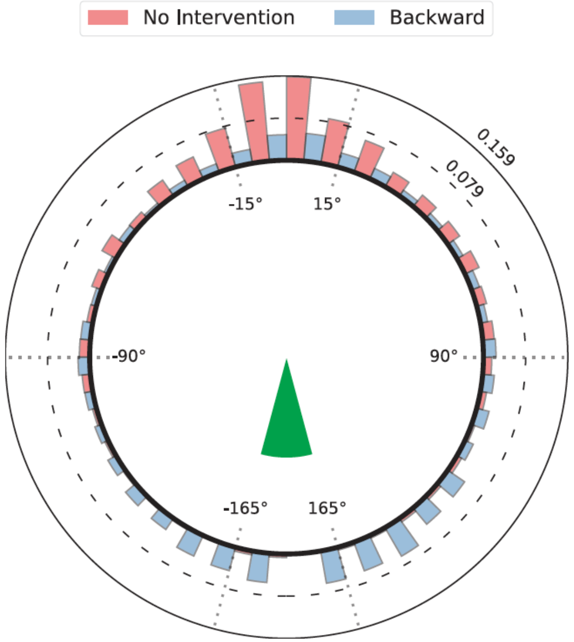
HAMT: Unconditional Directional Instructions (Analysis)

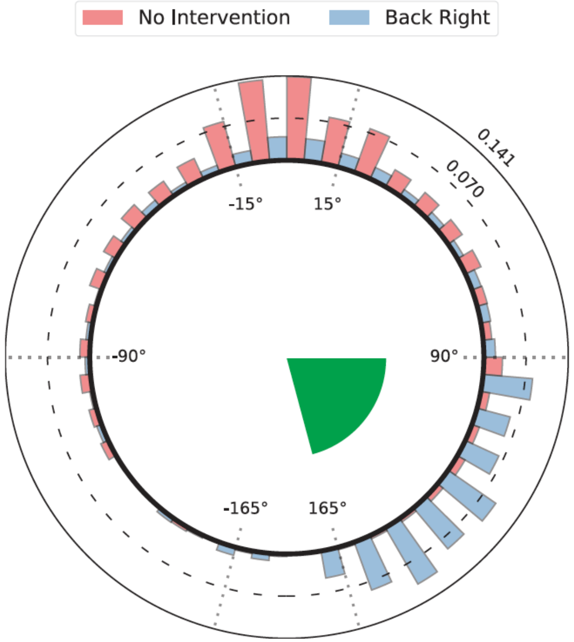




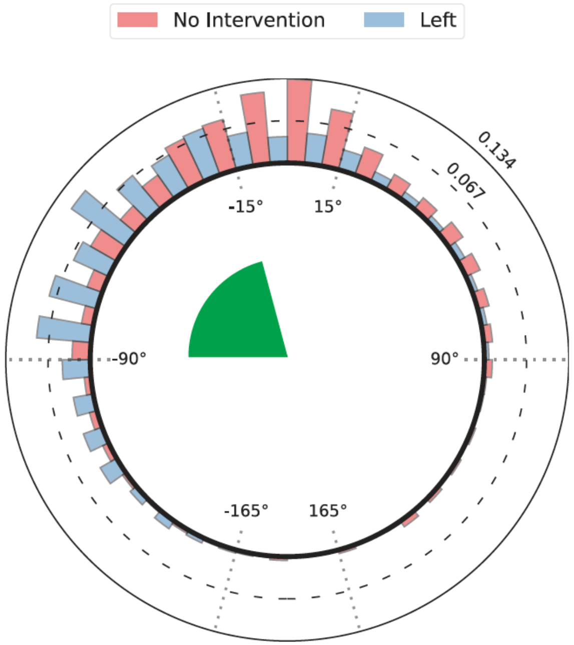


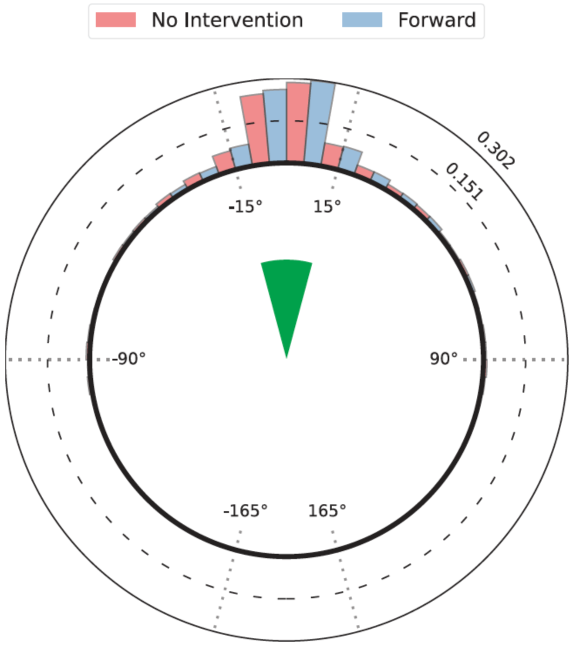


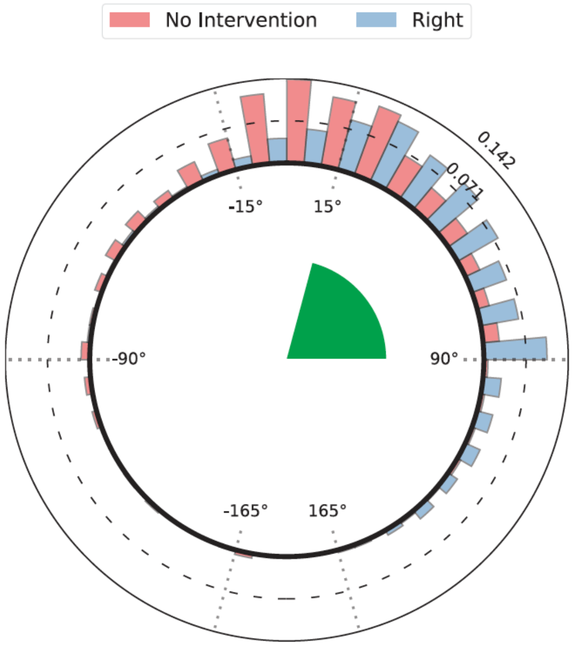


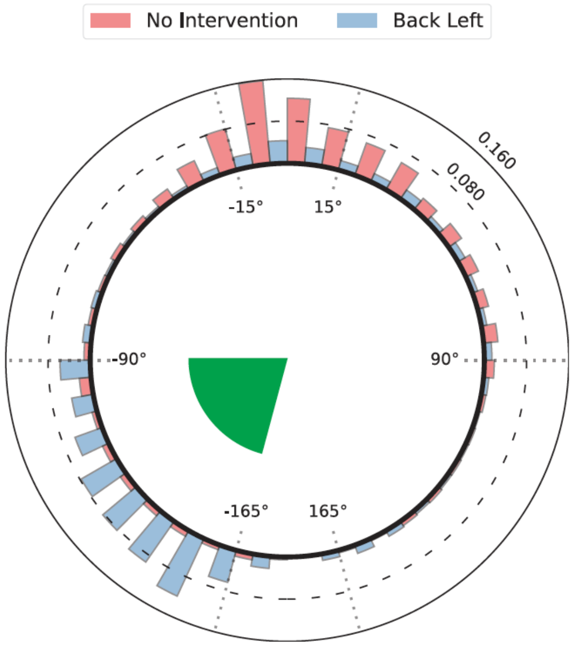


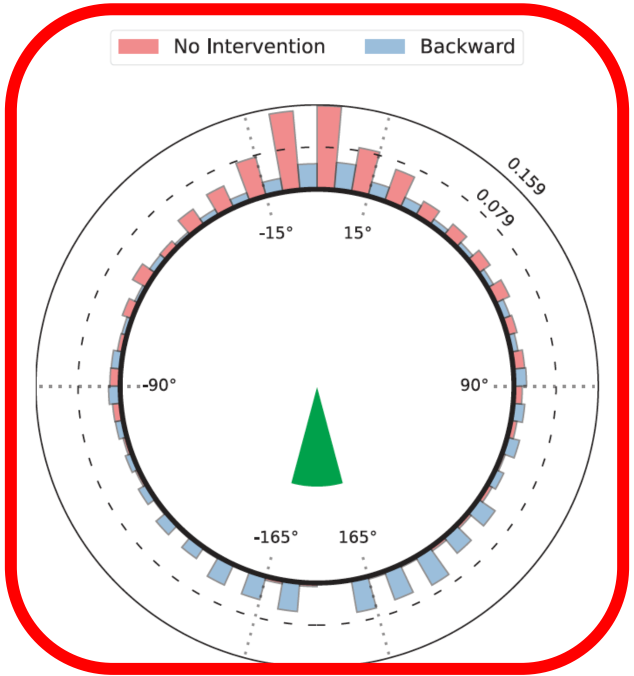
HAMT: Unconditional Directional Instructions (Analysis)

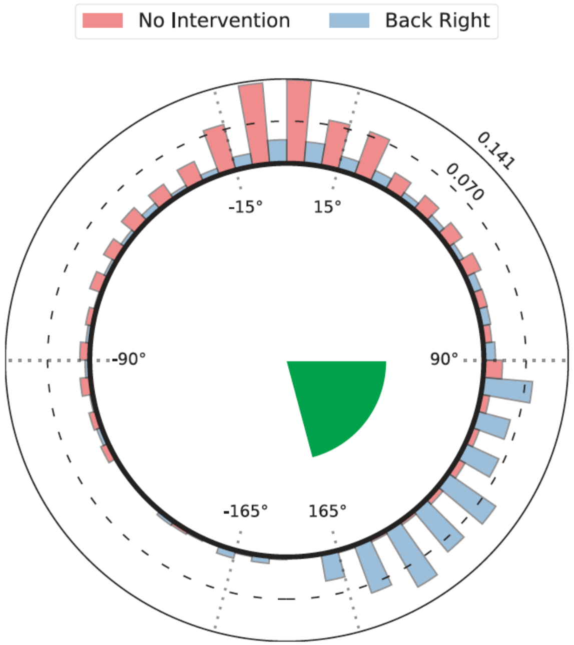




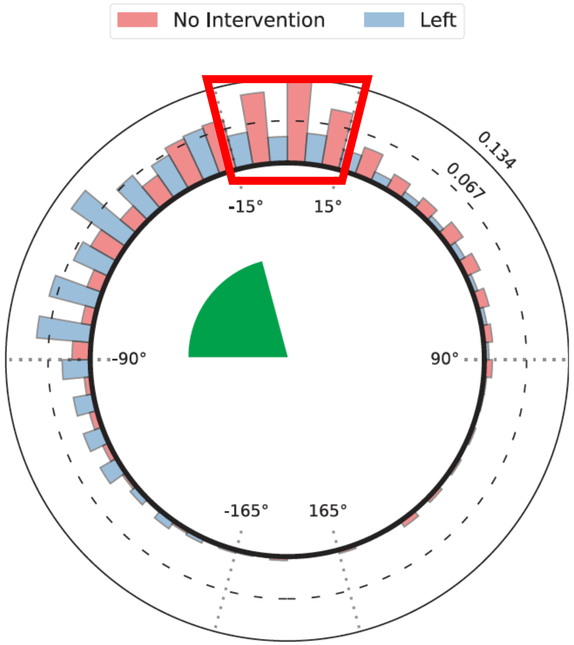


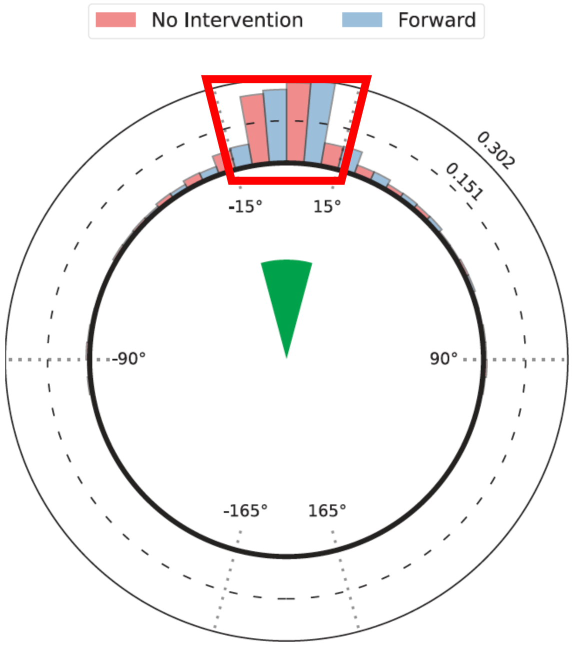


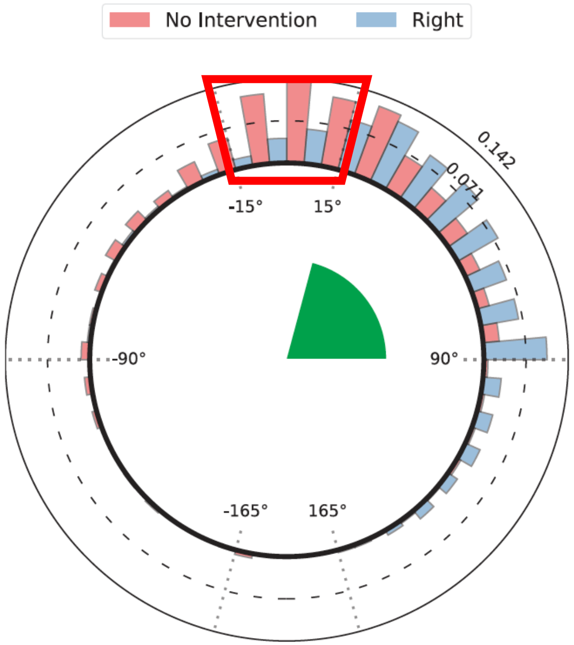


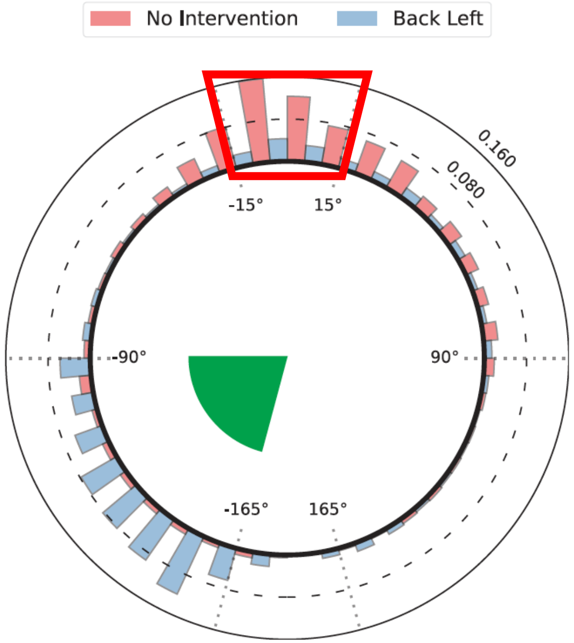


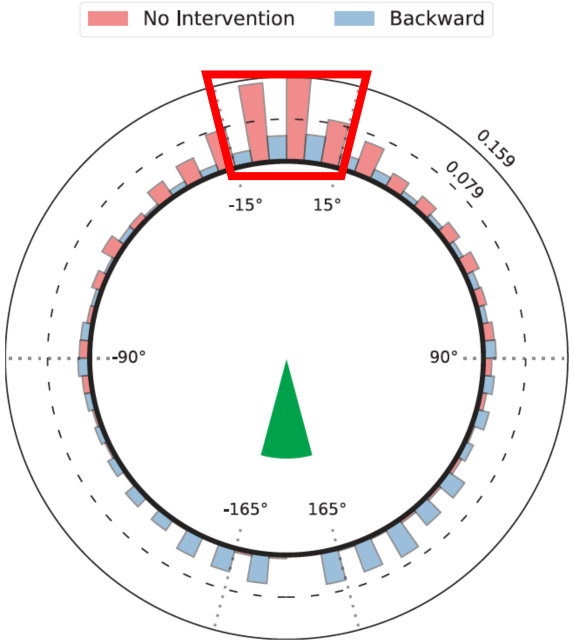
HAMT: Unconditional Directional Instructions (Analysis)

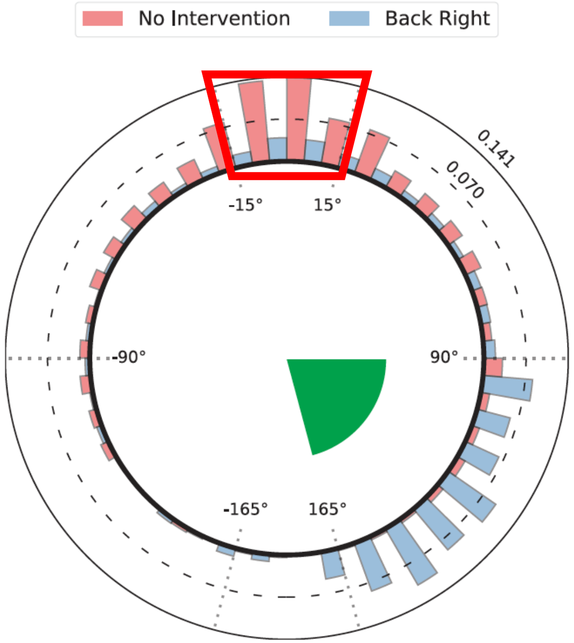








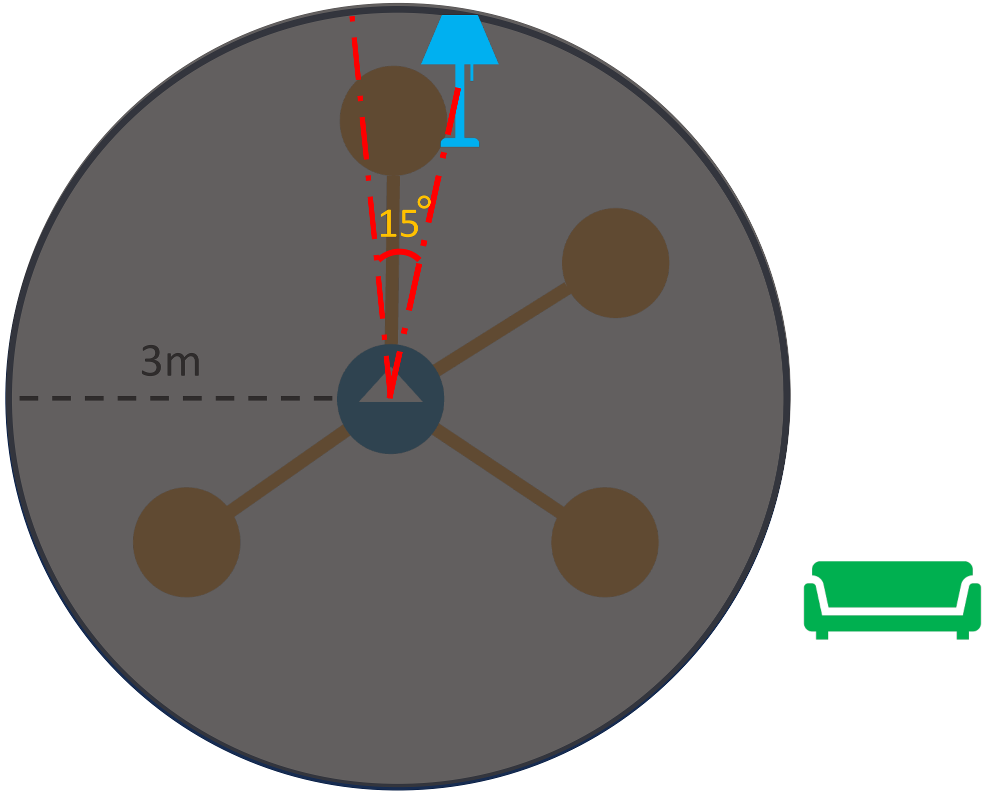




HAMT: Object Seeking (Filtering)

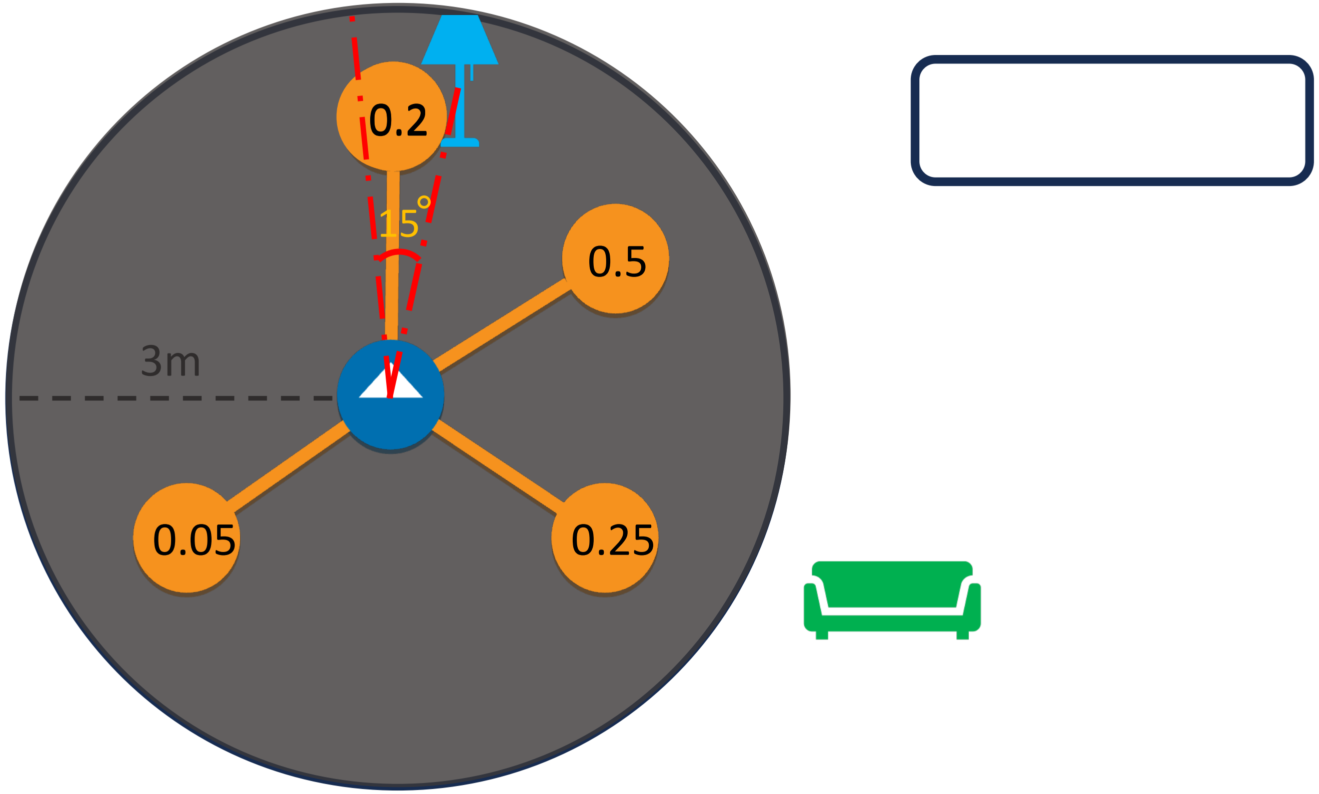
REVERIE (Qi et al.)

Walk towards the XX (*Object*)



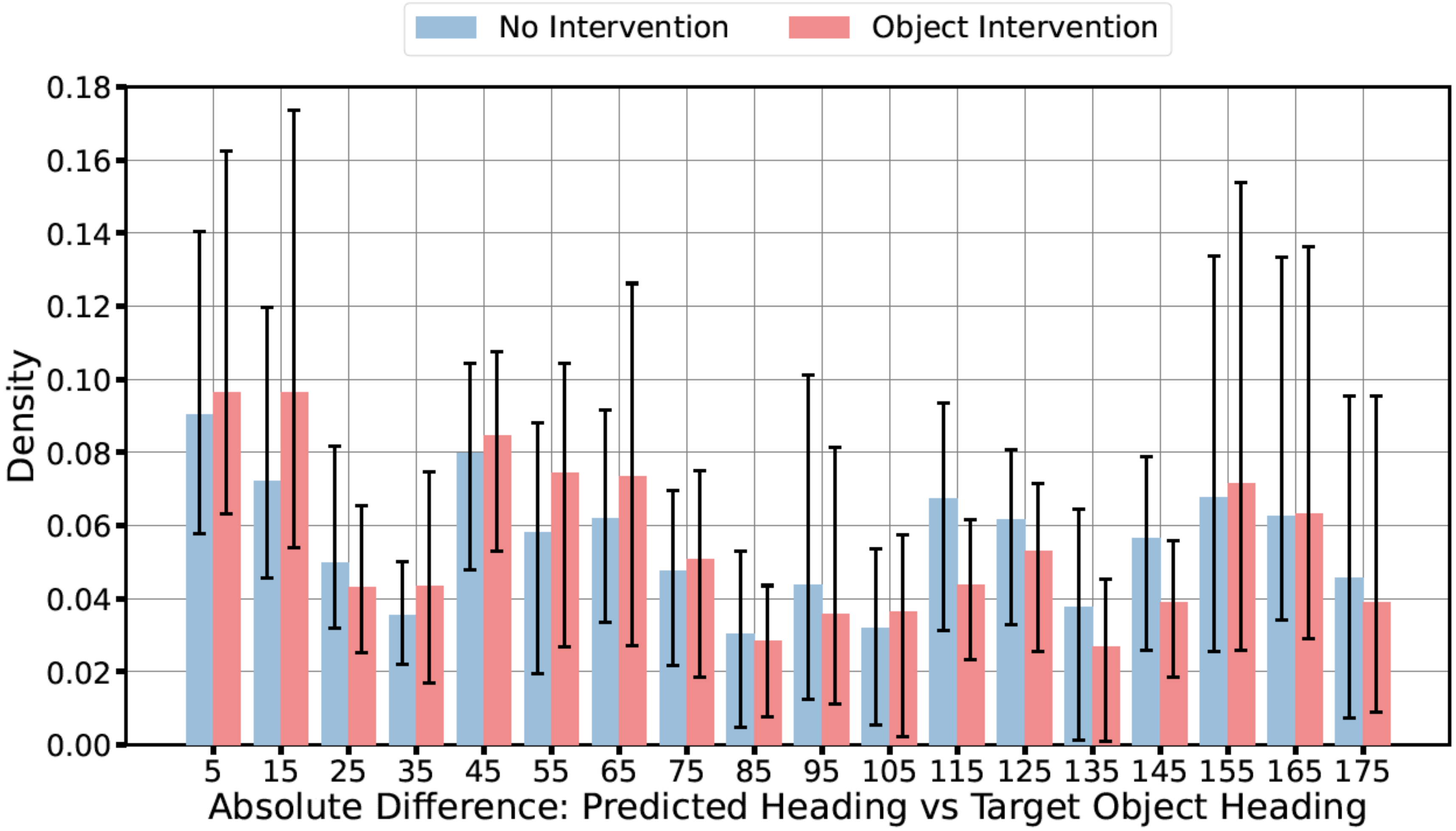
HAMT: Object Seeking (Mapping)

REVERIE (Qi et al.)

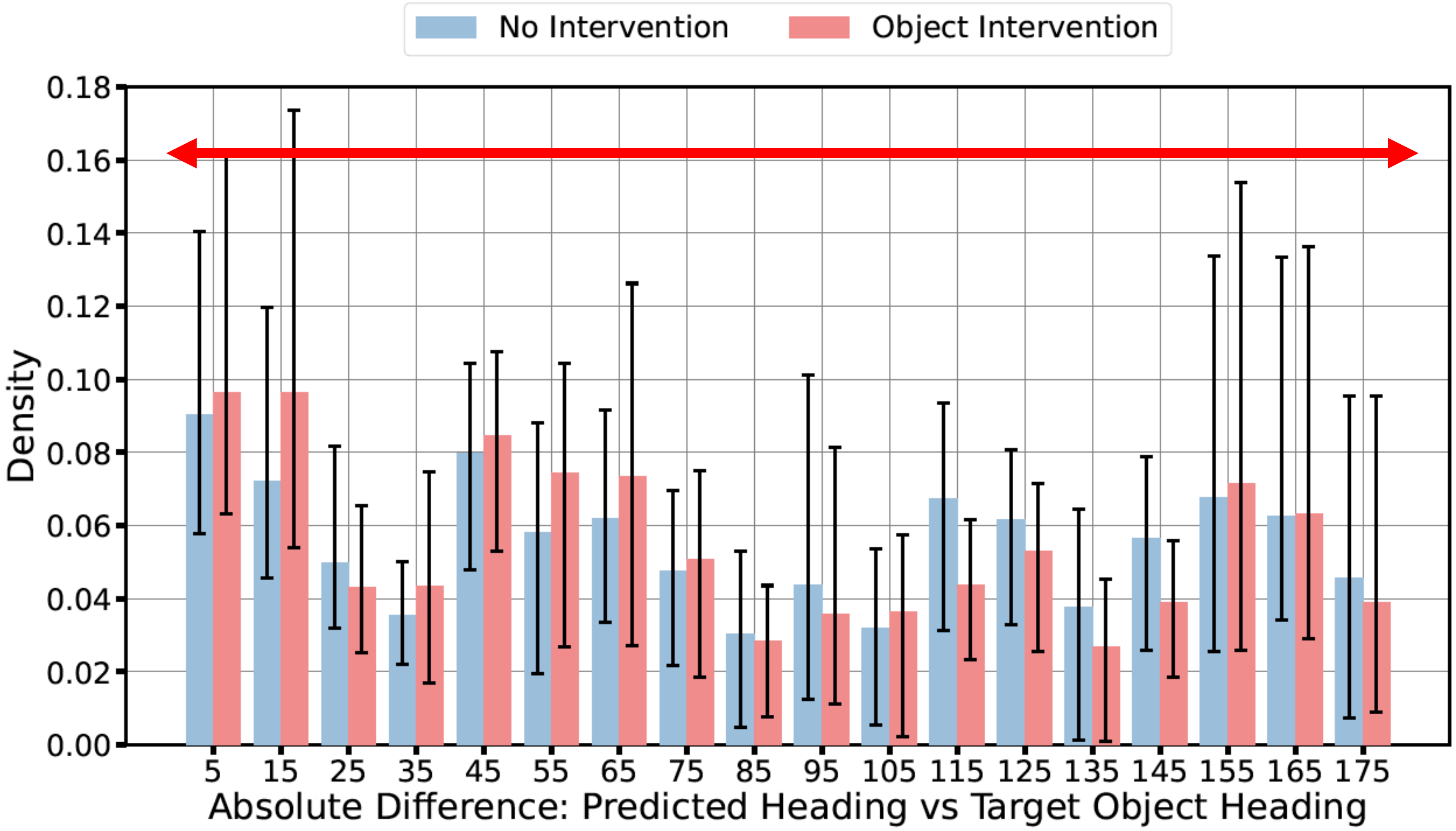


Behavior Analysis Quantity → Absolute angular error relative to object

HAMT: Object Seeking (Analysis)



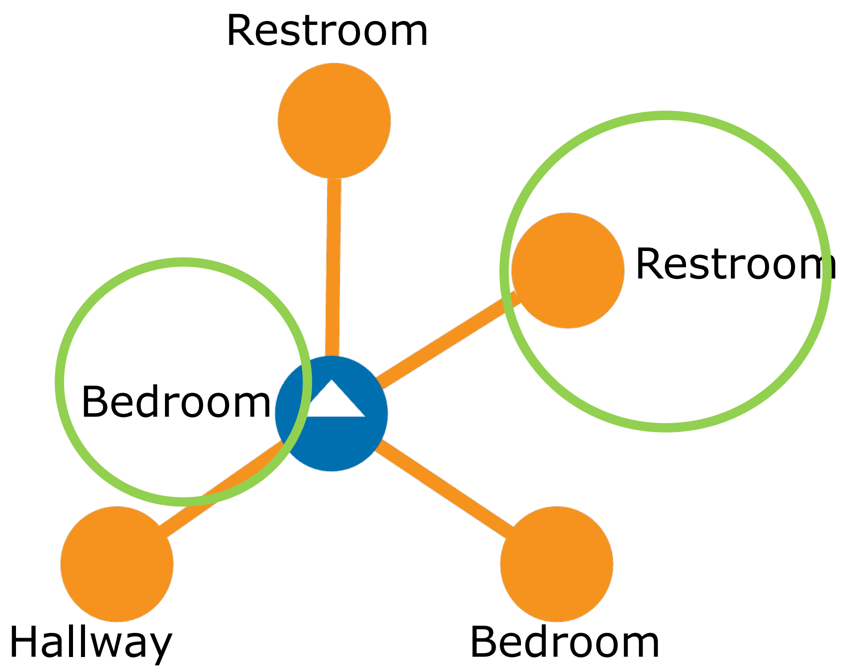
HAMT: Object Seeking (Analysis)



HAMT: Room Seeking (Filtering)

MATTERPORT3D (Chang et al.)

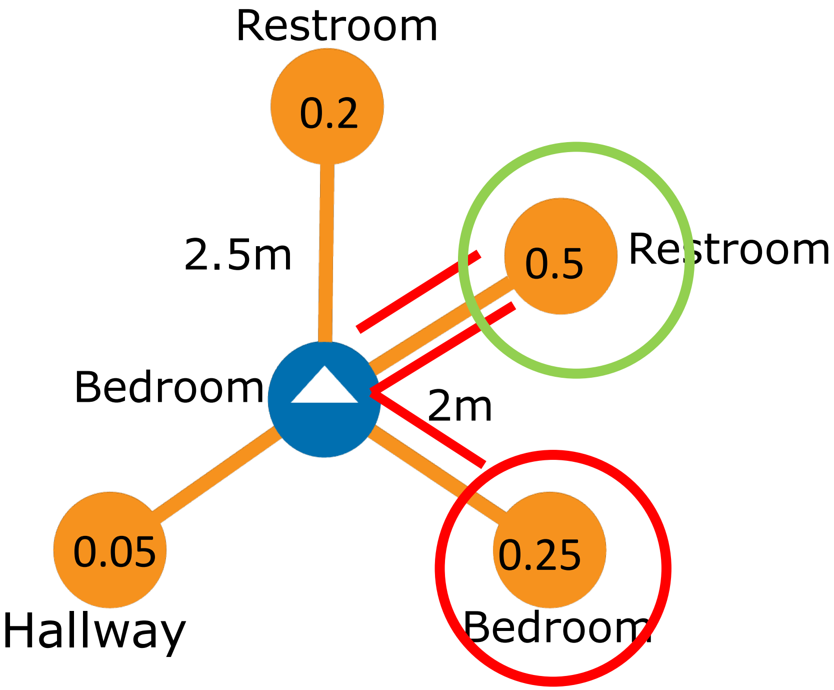
Walk towards the XX (*room type*)



HAMT: Room Seeking (Mapping)

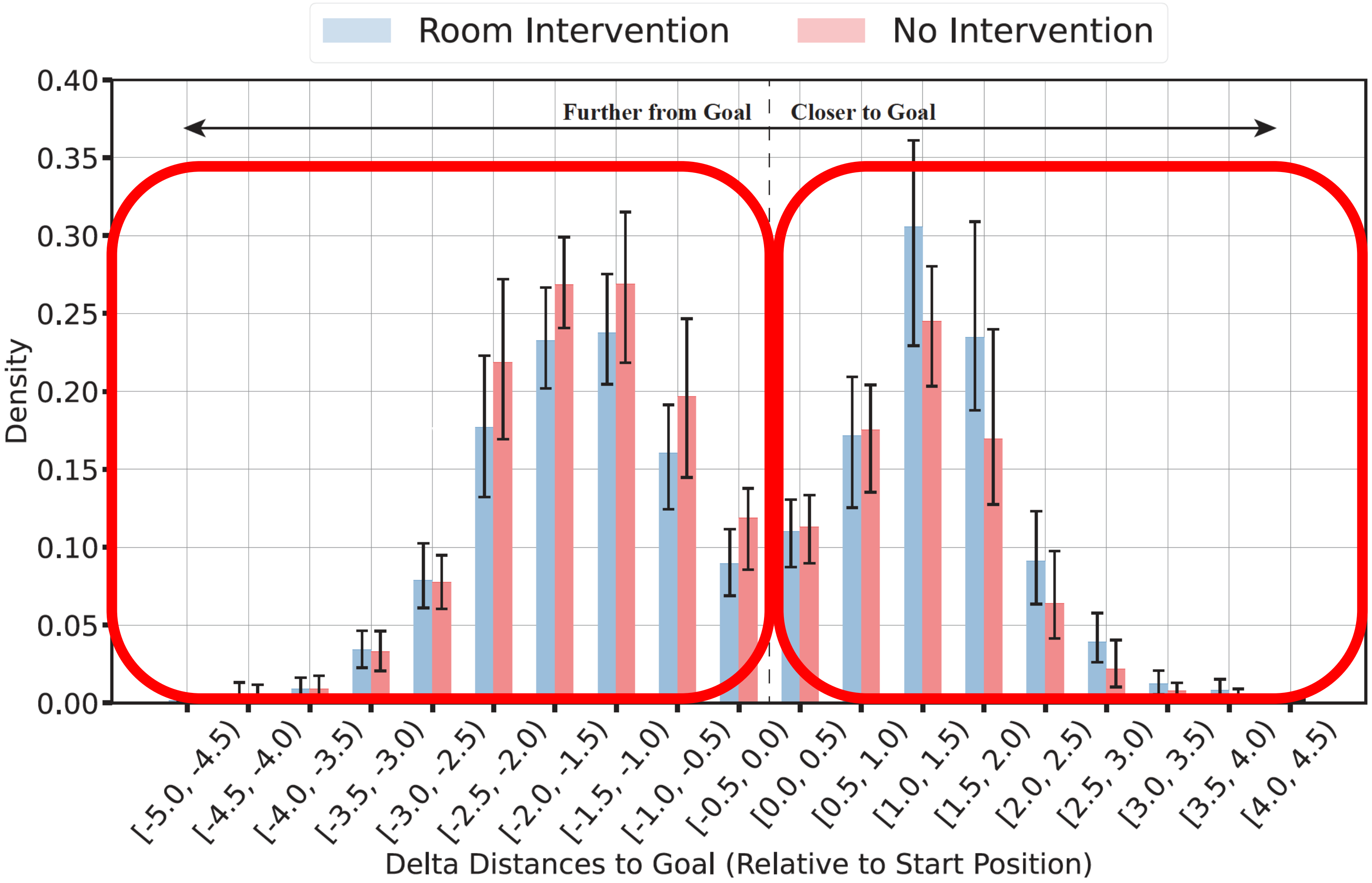
MATTERPORT3D (Chang et al.)

Length(   ) — Length(   ) :

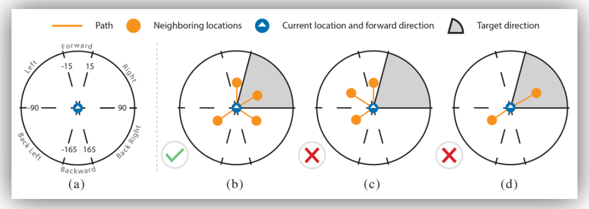


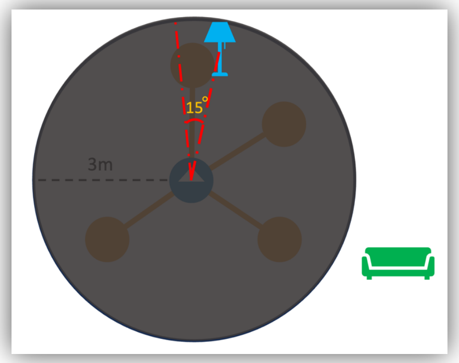
Walk towards the restroom.

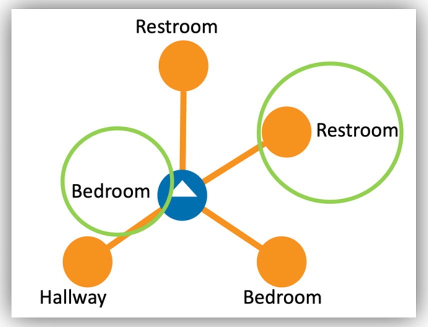
HAMT: Room Seeking (Analysis)



Competency Score



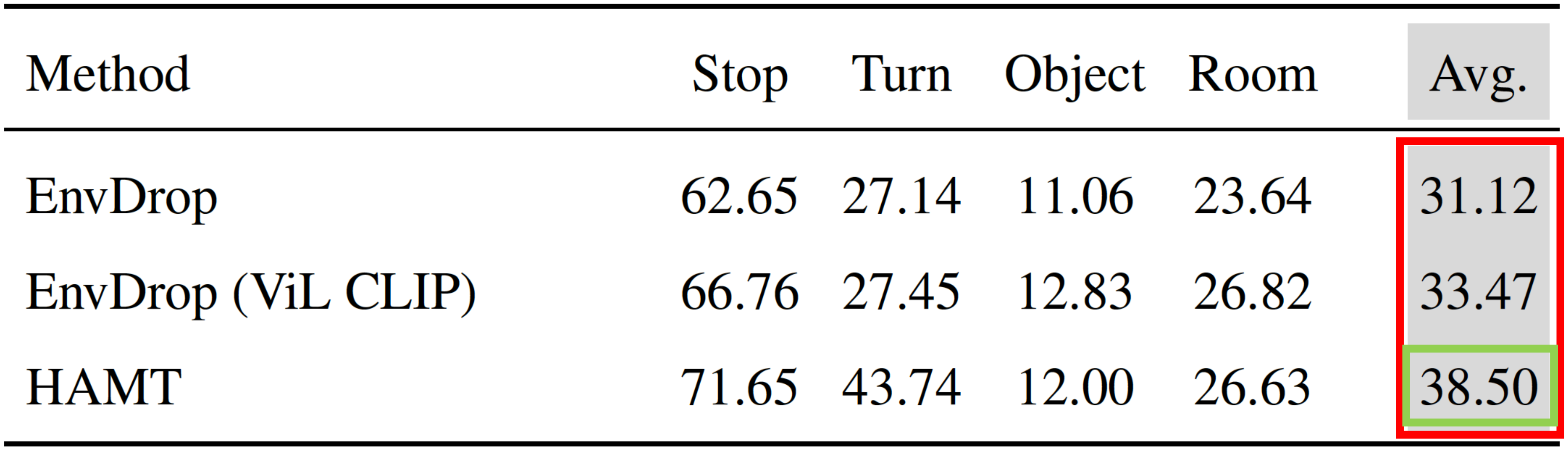




Target specified by Intervention Instruction

Competency Score

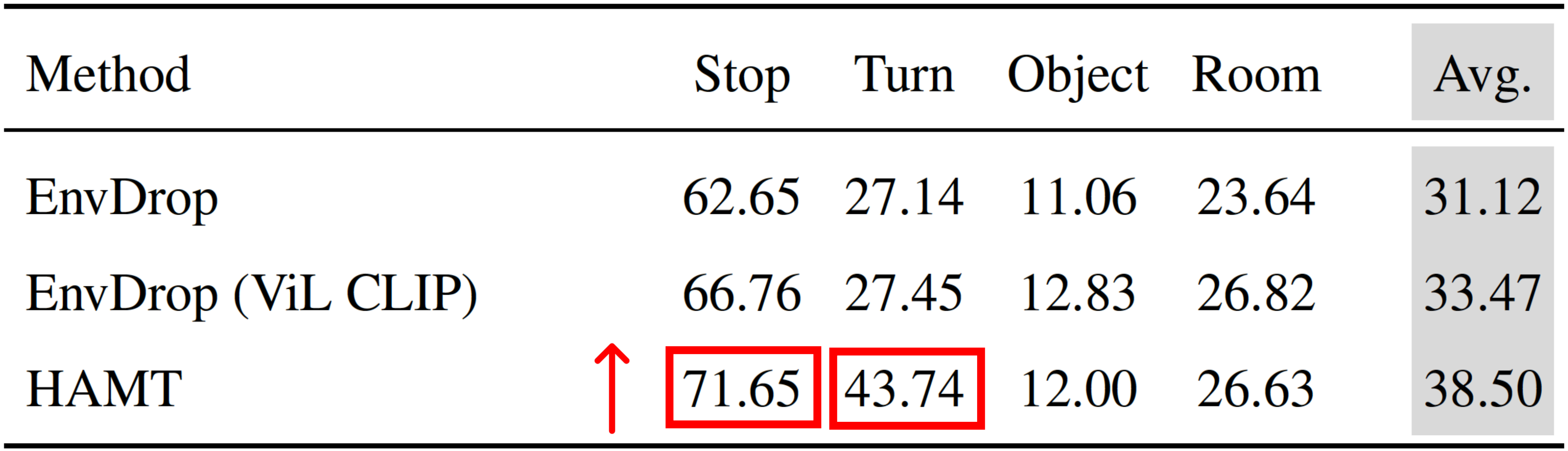
Ordered by increasing RxR performance



Method	Stop	Turn	Object	Room	Avg.
EnvDrop	62.65	27.14	11.06	23.64	31.12
EnvDrop (ViL CLIP)	66.76	27.45	12.83	26.82	33.47
HAMT	71.65	43.74	12.00	26.63	38.50

Competency Score

Ordered by increasing RxR performance



Method	Stop	Turn	Object	Room	Avg.
EnvDrop	62.65	27.14	11.06	23.64	31.12
EnvDrop (ViL CLIP)	66.76	27.45	12.83	26.82	33.47
HAMT	71.65	43.74	12.00	26.63	38.50

Contact & Project Information

GitHub Repository

https://github.com/Yoark/vln-behave

Project Page

https://yoark.github.io/paper_sites/vln-behave/

For more information or any questions, feel free to reach out:

Zijiao Yang
yangziji@oregonstate.edu