Supplementary for Behavioral Analysis of Vision-and-Language Navigation Agents

1. Discussion

We discuss the insights regarding model behaviour, as well as some future directions. Main paper's goal is to develop a framework for skill-based behavioral analysis, despite that, we could provide some speculation on the underlying reason of model behaviour in the hope of benefiting the future development of embodied agents.

–Low room/object sensitivity of HAMT. Low sensitivity of object and room seeking implies a weakness of the agent in spatial relation reasoning and vision-language alignment. We suspect this resulted from lack of specific proxy tasks, and visual features only capturing limited information (as also stated in [22]). We encourage people to design specific architectures, build proxy tasks addressing spatial relation reasoning, and incorporate richer object information or object representation learning modules.

– HAMT vs. EnvDrop (Stop & Turn). Architecture difference (HAMT vs EnvDrop: Transformer vs Recurrent Neural Network) might give HAMT an advantage in both Stop and Turn. Further for Turn, we believe some proxy training tasks unique to HAMT brought the advantage. We suspect Single-Step Action Regression, and Spatial Relationship Prediction are helpful. Former predicts action heading and elevation directly from given instruction, history, and current observation; latter predicts relative spatial position of two views given visual feature angle feature or both. Further analysis could be interesting future work.

– EnvDrop vs. EnvDrop (CLIP). CLIP may provide improved semantics, but not action-grounding benefits. A full-scale component-wise analysis is out-of-scope for this paper, but would be an interesting application of our behavioral analysis framework that our code release could support in the future.

2. Data Correlation Analysis

Our dataset represents a finite, correlated sample from the space of all instruction-trajectories pairs in indoor scenes. There may be correlation within trajectories from the same scan or from interventions drawn from the same trajectory. We conducted Hierarchical bootstrapping and linear mixed effect modeling to account for the correlation in data.

– Hierarchical bootstrapping for CIs. We use hierarchical bootstrap resampling [2] (scenes→trajectories) to correctly simulate a new draw from the underlying population we are studying. Then we obtain confidence intervals from the new draw.

– Linear mixed effect modeling. We model each of our interventions with a linear mixed effect model where each scan and trajectory are modeled as imparting a random slope and intercept along with a overall fixed intervention effect – i.e. modeling the effect for an episode i taken from scan j and trajectory k as

$$\text{effect}_i = \left(w_{\text{fix}} + w_{\text{scan}_j} + w_{\text{traj}_k}\right) * I_i + b_{\text{fix}} + b_{\text{scan}_j} + b_{\text{traj}_k}$$

where $w_{\text{scan}_j}, w_{\text{traj}_k}, b_{\text{scan}_j}$, and b_{traj_k} are modeled as random effects and I_i is a binary variable indicating whether this episode contains an intervention. Models were fit using `lmer` in `R` and significance of fixed effects were evaluated through the `anova` command. We provided analysis for HAMT, ENVDROP-IMAGENET, ENVDROP-CLIP in main paper and appendix.

3. Additional Case Studies for Envdrop-clip and Envdrop-imagenet

We provide complete analysis for the two additional VLN agents we tested: ENVDROP–CLIP and ENVDROP–IMAGENET. These were not included as case studies in the main paper due to space.

3.1. Stop

Fig. 1 and Fig. 2 show average stop probabilities across different trajectory lengths for the truncated implicit stop, intervened explicit stop, and one-step ahead instruction settings for ENVDROP–IMAGENET and ENVDROP–CLIP. Error bars are 95% hierarchical bootstrap confidence intervals. For ENVDROP–CLIP, we find the average stop probability to remain fairly constant for shorter trajectory lengths (until around sixteen) under both implicit and explicit stop instructions, but dropped at longer trajectory lengths (from sixteen to twenty). This suggests agents ground the stop instruction better for shorter trajectories. And the plot also

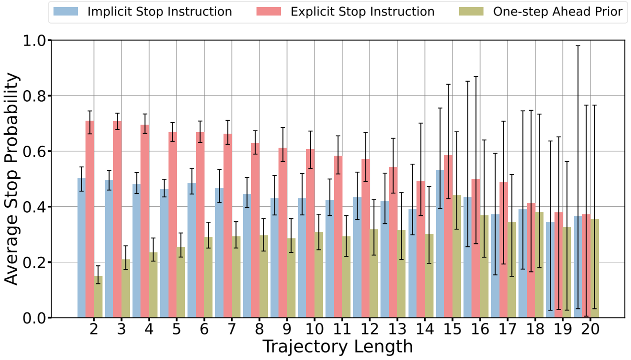

Figure 1. (Envdrop–clip) Average Stop Probability vs Trajectory length instruction for "implicit stop instruction", "explicit stop instruction" and "one-step ahead prior". We find agents respond strongly to both implicit and explicit stop interventions at earlier steps – stopping with high probability across shorter trajectory lengths. (Until around sixteen) Explicit stop instructions produce a stronger effect than implicit.

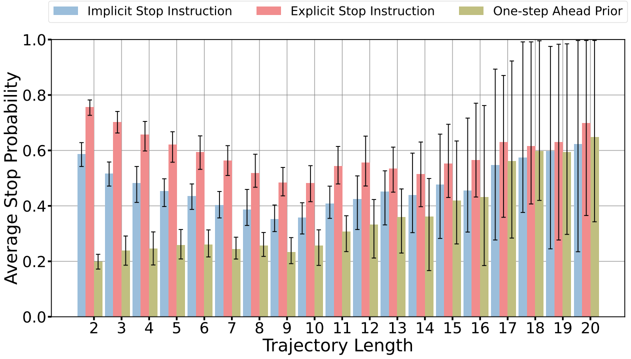

Figure 2. (Envdrop–imagenet)Average Stop Probability vs Trajectory length instruction for "implicit stop instruction", "explicit stop instruction" and "one-step ahead prior". We find agents respond strongly to both implicit and explicit stop interventions – stopping with high probability across all trajectory lengths. Explicit stop instructions produce a stronger effect than implicit.

suggests stop probability is higher for explicit than implicit stop and both are higher than one-step ahead setting.

To evaluate statistical significance of the effect, we again use `lmer` where the observed stop probability is assumed to be an effect of the intervention plus random effects from the environment and source trajectory. We find agents have a higher probability of stopping when given explicit rather than implicit stop instructions (0.66 *vs*. 0.47, effect 0.19 `anova:` $p \approx 0$), and the agent responds to both implicit and explicit stop instructions by increasing stop probability compared to the one-step ahead baseline (effect 0.22, $p \approx 0$). For ENVDROP–IMAGENET, we find the average stop probability to remain fairly constant for implicit and explicit stop instructions across all trajectory lengths. This suggests the agent can ground to both implicit and explicit stop instructions regardless of trajectory length. The stop probability for explicit stop instruction is higher than implicit stop instruction, and both are higher than one-step ahead setting. We find agents have a higher probability of stopping with explicit rather than implicit stop instructions (0.63 *vs*. 0.47, effect: 0.16 $p \approx 0$), and the agents respond to both implicit and explicit stop instructions by increasing stop probability compared to the one-step ahead setting (effect 0.22, $p \approx 0$)

Note ENVDROP–IMAGENET has a tendency to stop more likely for longer trajectory than ENVDROP–CLIP. This might suggest the correlation between trajectory lengths and stop probability for ENVDROP–IMAGENET is stronger.

Summary. We find both ENVDROP–IMAGENET and ENVDROP–CLIP respond strongly to implicit and explicit stops across most trajectory lengths and explicit stop instructions have a stronger effect. In addition, we find ENVDROP–CLIP tends to have a lower probability of stopping at longer trajectories regardless of stop instructions.

3.2. Unconditional Directional Instructions

Fig. 3 and Fig. 4 show the distribution of probabilities over all episodes for each directional intervention as histograms on polar axes. For convenience, we denote the target direction region with a green arc at the center of each plot.

For ENVDROP–CLIP, across all directions, we find the agent either stops (roughly 46% of the time on average) or moves in a roughly forward direction in the no intervention setting. There is a slight bias towards left or right in those settings. However, the agent does not receive any left/right instruction, so this reflects a minor structural bias caused by the filtering process. All left (right) episodes include a neighbor to the left (right) and an agent with a bias towards moving roughly forward may select them at a higher rate than nodes in the backward direction.

For the intervention setting, we see a strong response to directional language for forward, left and right. For these three settings, the agent stops significantly less (roughly 21% of the time on average) and we observe a shift in distribution towards the corresponding direction. Similarly as before, we accumulate the probability mass into directional bins and evaluate the effect of intervention on the accumulated probability. We again use `lmer` the same as before to account for potential correlations in scenes and trajectories. We find the agent exhibits a significantly higher accumulated probability for forward, left, and right direction with directional instruction than without – estimating intervention effects as increased accumulated probability for forward (0.08, $p \approx 0$), left (0.36, $p \approx 0$), right (0.34, $p \approx 0$)

For backward, back left, and back right, the agent does not have a good response to directional language. We find

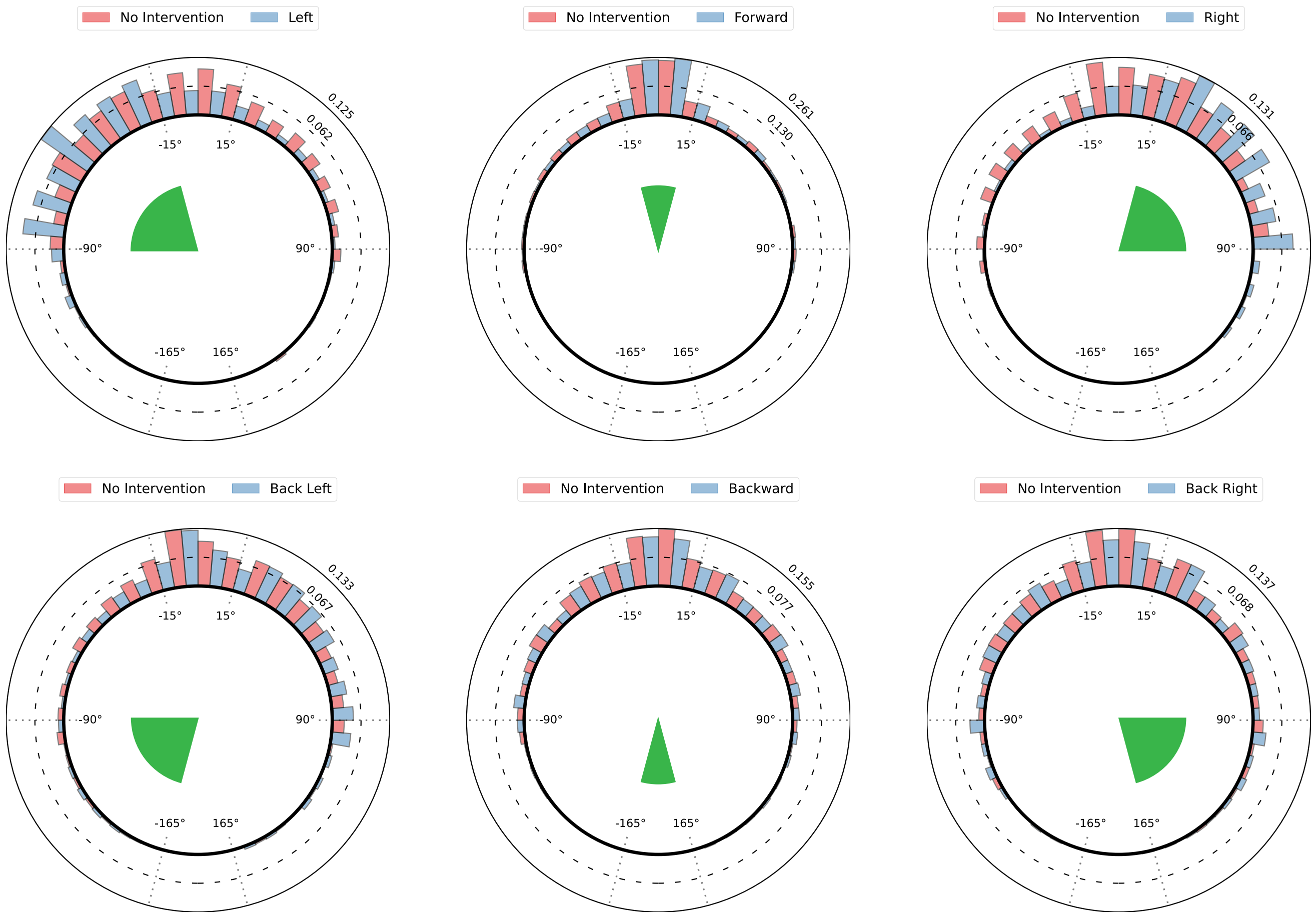

Figure 3. (ENVDROP–CLIP) We plot the next step direction probability distribution of the agent onto polar axis for easier visualization. We provide results for six directions and contrast between "No Intervention" (red) with "Direction" (blue). The number on the outer circle and middle dotted circle are max and $\frac{max}{2}$ respectively. We found the ENVDROP–CLIP agent is responsive to only three directional instructions: forward, left and right. The probability mass of directional interventions shifts toward the area indicated by those three directional instructions compared to "No Intervention".

the agent either stops (roughly 58% of the time on average), moves forward (reflecting forward bias the agent learned during training), or responds to part of the instruction. We created backward, back left, and back right directional language by composing sub-instructions. ("Turn around and walk forward" for backward, "Turn around and go to your right" for back left, and "Turn around and go to your left" for back right). Fig. 3 suggests for all three conditions, the agent may not be able to execute "turn around" or may not be able to compose "turn around" and other directional instructions. Similarly, estimating intervention effects as increased accumulated probability for backward ($3E{-}4$, $p = 0.42$), back left ($-3E{-}3$, $p = 0.33$), back right ($4E{-}3$, $p = 0.38$). We observed overall similar effects for ENVDROP–IMAGENET in Fig. 4, the agent responds to forward, left, and right strongly, but has no respond to backward, back left and back right. The estimated intervention effects are: forward (0.08, $p \approx 0$), left (0.36, $p \approx 0$), right (0.32, $p \approx 0$), backward ($6E{-}4$, $p = 0.27$), back left ($-5E{-}3$, $p = 0.03$), back right ($-3E{-}3$, $p = 0.29$)

Summary. We find both ENVDROP–CLIP and ENVDROP–IMAGENET agents strongly respond to directional language for forward, left and right. But they are not able to respond to backward, back left, back right conditions properly. They only ground to part of the intervention instruction but fail on the whole instruction. (e.g., probability mass distributed to "right" for "turn around and go to right" instead of the correct direction, "back left") This may due to inability to parse "turn around" instructions. Some dataset biases from training are still evident in a bias towards forward actions.

3.3. Object

Fig. 5 and Fig. 6 present distributions over angular distance for the intervention and no-intervention settings for

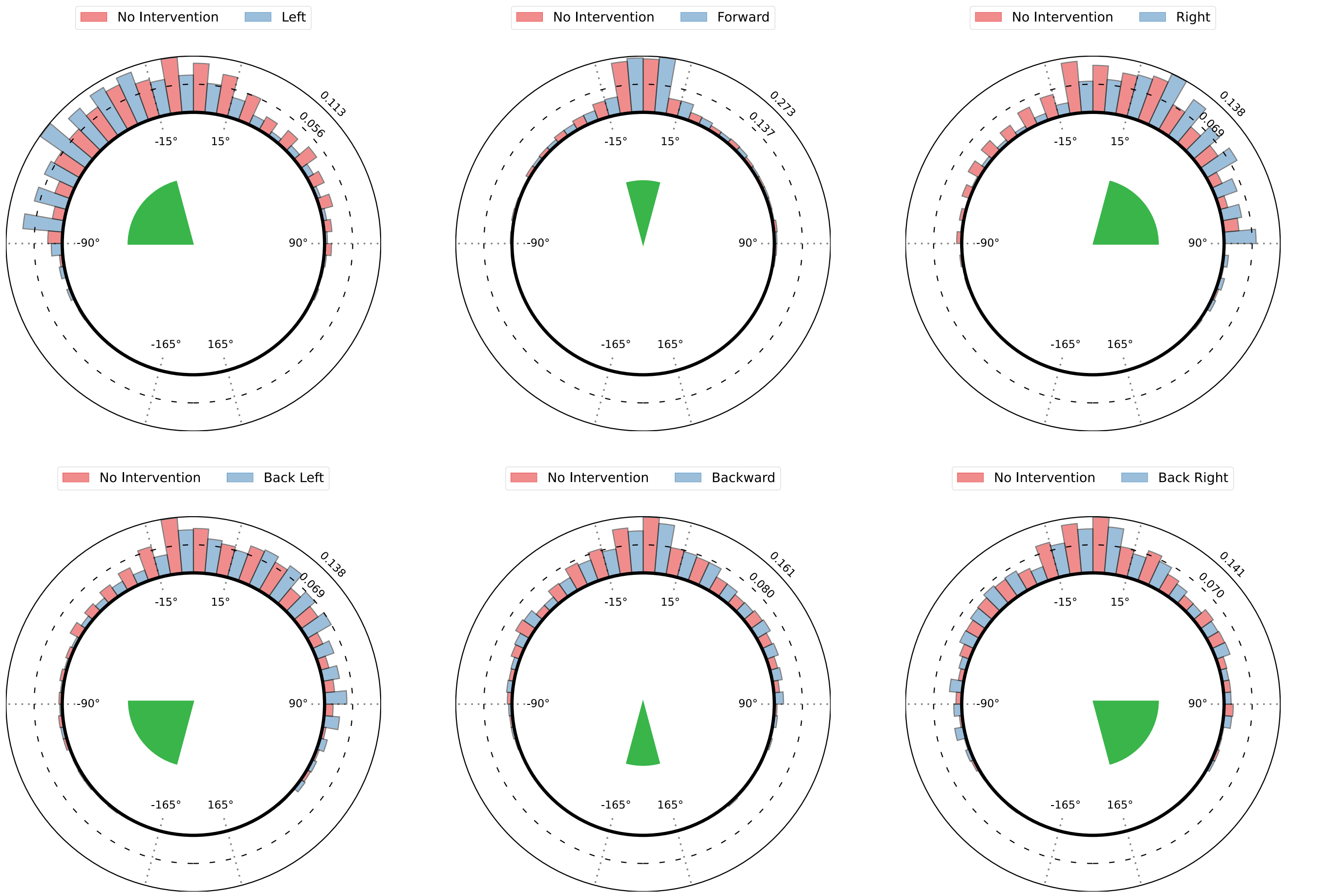

Figure 4. (ENVDROP–IMAGENET) We plot the next step direction probability distribution of the agent onto polar axis for easier visualization. We provide results for six directions and contrast between “No Intervention” (red) with “Direction” (blue). The number on the outer circle and middle dotted circle are max and $\frac{max}{2}$ respectively. We found the ENVDROP–IMAGENET agent is responsive to only three directional instructions: forward, left and right. The probability mass of directional interventions shifts toward the area indicated by those three directional instructions compared to “No Intervention”.

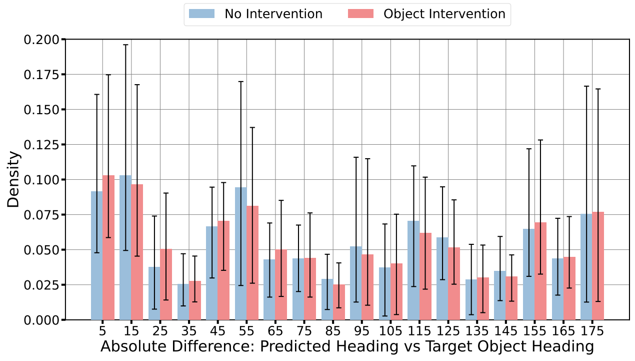

Figure 5. The distribution of the absolute difference between model prediction and target object direction for intervention and no intervention settings. (ENVDROP–CLIP)

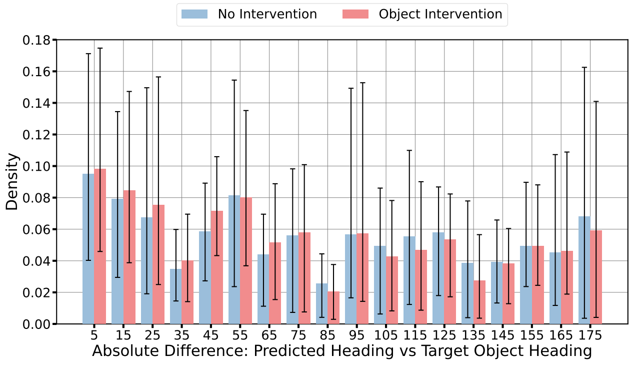

Figure 6. The distribution of the absolute difference between model prediction and target object direction for intervention and no intervention settings. (ENVDROP–IMAGENET)

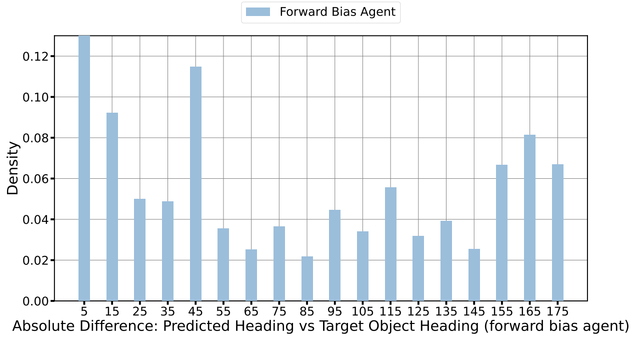

Figure 7. The distribution of absolute difference between model prediction and target object direction for forward bias agent.

ENVDROP–CLIP and ENVDROP–IMAGENET respectively.

For ENVDROP–CLIP, the agent is not significantly responsive to object-seeking instruction. (blue *vs*. red bars in Fig. 5. We again use a `lmer` to evaluate the effect of intervention on the accumulated probability within 15 degree of absolute angular difference. We find weak fixed effect of $4E{-}2$ (`anova`, $p = 2E{-}6$) for intervention vs nonintervention. For ENVDROP–IMAGENET(Fig. 6), we find a weak fixed effect of $3E{-}2$, ($p = 6E{-}7$) for intervention vs non-intervention. However, both ENVDROP–CLIP and ENVDROP–IMAGENET show a wide spread angular error that suggests the target objects are not being grounded accurately. (Recall all trajectories have neighboring nodes that would incur no more than 15 degrees of error.) To explore this error distribution further, we also examine a baseline `Forward bias` (Fig. 7 agent that places probability on neighbors inversely proportional to their relative heading. We find this baseline exhibits a similarly shaped error distribution to the agent – suggesting the agent may be taking forward actions when uncertain about the target object. As in our other experiments, the no intervention setting is more likely to stop than the intervention (50% *vs*. 30% for ENVDROP–CLIP, 54% *vs*. 37% for ENVDROP-IMAGENET.

Summary. We find evidence for only a weak tendency to move towards referenced objects for ENVDROP–IMAGENET, and ENVDROP–CLIP.

3.4. Room-seeking Instructions

1-Hop Results. The probabilities of delta distance for ENVDROP–CLIP and ENVDROP–IMAGENET are displayed in Fig. 8 and Fig. 9 respectively – values greater than zero represent the agent moving *closer* to nodes with the target room type. We observe a weak right-shift in the density suggesting the agents respond somewhat to the intervention. We again use a `lmer` to evaluate the effect of intervention on the delta geodesic distance. For ENVDROP–CLIP, we find the estimated fixed effect as 0.10 (`anova`, $p = 9E{-}5$) for intervention vs no intervention. For

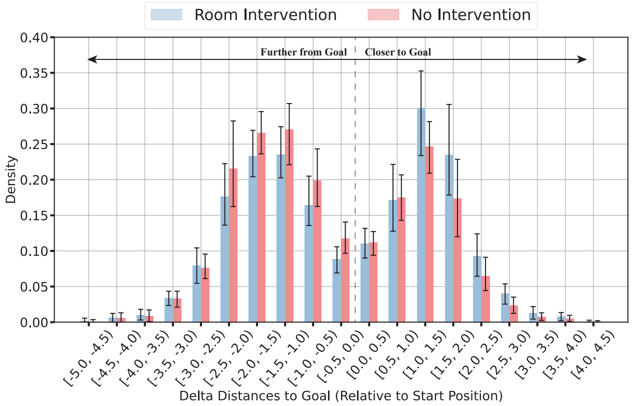

Figure 8. (ENVDROP–CLIP) Distribution of delta distance to nodes of the target room type. The delta distance difference of distance to nodes of target room type (relative to start position) with or without intervention. Positive delta distance means the agent move closer to nodes of target type with intervention than otherwise. The distribution shift towards right with intervention than otherwise, indicates the agent is responsive to room-seeking instruction. (-0.15 *vs*.-0.41, $p = 9E{-}5$)

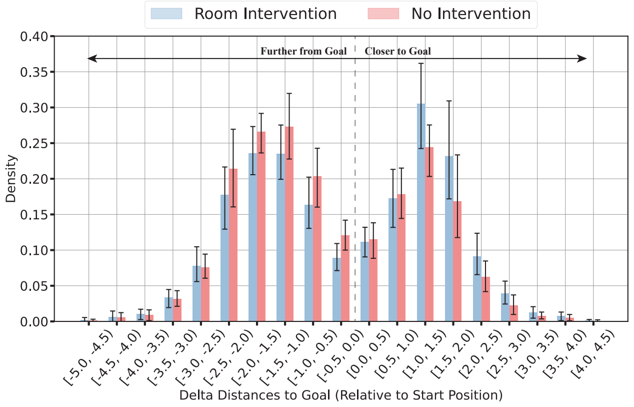

Figure 9. (ENVDROP–IMAGENET) Distribution of delta distance to nodes of target room type. The delta distance difference of distance to the nodes of target room type (relative to start position) with or without intervention. Positive delta distance means the agent move closer to nodes of the target type with intervention than otherwise. The distribution shift towards right with intervention than otherwise, indicates the agent is responsive to room-seeking instruction. (-0.16 *vs*.-0.42, $p = 4E{-}3$)

ENVDROP-IMAGENET, we find the estimated fixed effect as 0.05 ($p = 4E{-}3$). However, Both the agents do not reliably place strong beliefs on neighbors with the target room type – negative median delta distance and significant mass to the left of zero.

k-Hop Results. We report the distance to the nearest node with target room type here.

We show ridgeline plots in Fig. 10 and Fig. 11 for

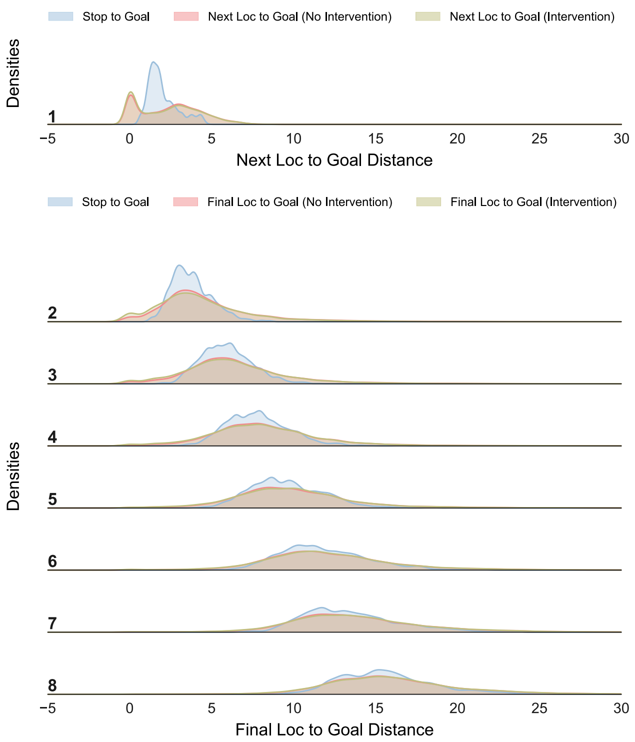

Figure 10. (ENVDROP–CLIP) Distribution of geodesic distance to the nearest node of target room type for k-hop room-seeking experiments. `Stop to Goal` is a baseline agent that always takes the stop action.

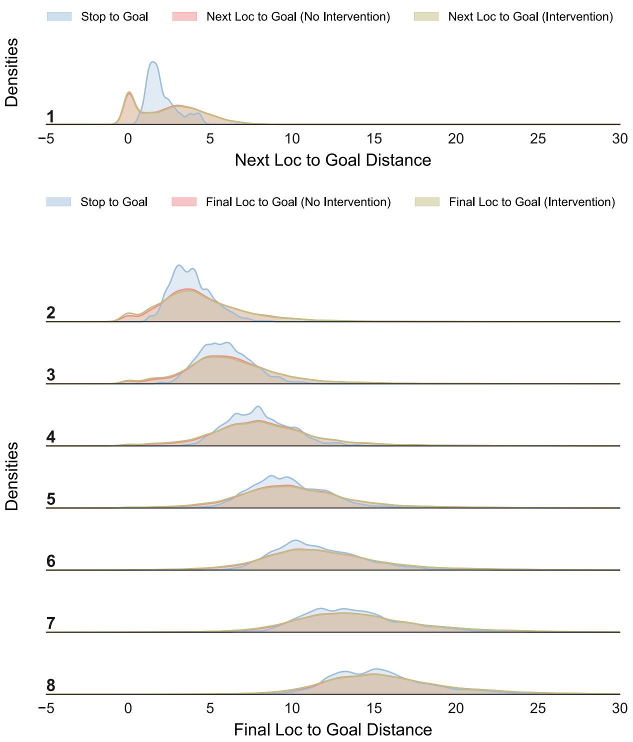

Figure 11. (ENVDROP–IMAGENET) Distribution of geodesic distance to nearest node of target room type for k-hop room-seeking experiments. `Stop to Goal` is a baseline agent that always takes the stop action.

ENVDROP–CLIP and ENVDROP–IMAGENET, respectively. We compare distance to the nearest node of target room type distributions for 1- to 8-hops. For both agents, we find the error increases with target room distance. We again leverage a `lmer` to evaluate the effect of intervention on $d_{geo}(n_{end}, n_{near})$. For ENVDROP–IMAGENET, we find weak effect of ≤ -0.08 (`anova`, $p = 8E{-}3$) for intervention *vs*. non-intervention for 2–8 hops with 95% confidence). For ENVDROP–CLIP, we find similar weak effects ≤ -0.08 ($p = 1E{-}2$) for 2–6 hops with 95% confidence. Overall, this suggests agents have limited ability to search for rooms based on common sense exploration.

Summary. Both the ENVDROP–CLIP [3] and ENVDROP–IMAGENT [4] agents are only weakly sensitive to room type reference instructions when the room is visible (within one hop) but lack the ability to perform common sense exploration to find further away rooms (k-hop). Overall sensitivity is low, suggesting the agent may not rely on room-specifying portions of instructions when navigating.

4. Templates and Examples

Tab. 1 show templates we used in our cases studies. And as we mannully designed the templates from examining RxR [1] dataset, we also provide example instructions in Tab. 2 containing the templates. Note the we make sure each part of templates can be found in training data.

Skill-specific language	**Template**
Stop Instruction	This is your destination.
	This is your end point.
	You reached your destination.
	You are done.
Unconditional Directional Instructions	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*)
Object-seeking Instruction	Walk towards the XX (*Object*)
Room-seeking Instruction	Walk towards the XX (*room type*)

Table 1. Templates for Skill-specific language used for our work.

Template	**Examples**
This is your destination. This is your end point. You reached your destination. You are done.	As you're facing the wall, you're gonna see 4 white coats to your right. Just turn around and take a few steps forward, you're gonna have a small sink to your left. **This is your destination.**
Walk forward.	Starting off in a side room library. **Walk forward.** And you ll see an open living room with chairs, a piano to your left, a desk to your right , wall is also to the right. Continue straight to the middle of the chair and the desk. Once the desk is to your right forward is a window with a mountain range and in front is also another couch, a big long couch. And to the left is a small circle table. Taking one last step and onto the couch the table is still to your left
Turn around	You will start by standing in front of a glass door and on your right is a doorway. **Turn around** and you will see a doorway to the washroom. Walk towards the doorway and inside the washroom. Once you're there, stand in between the sink and the bathtub and once you're there, you're done.
Go to your right	You are facing a large window. You are going to turn all the way around. You are going to exit this room and make a right. You are going to move forward into this room on the large blue rug. And you are going to go to the middle door on your right. The doors will now be open and you are going to take a step outside. You are going to **go to your right.** You are going to move forward down this pathway. And you are going to stop when you are right next to the yellow outlined glass window on the building will be on your right and on your left is just going to be the cement banister between 2 columns and you are done.
Go to your left	You are facing an open door and a massage bed. You are going to go thru the door. And once thru the door you are going to make an immediate right. You are going to step into this room and you are going to **go to your left.** You are going to hop over to the 3rd massage lounger on your right. Then you are going to make a right and go thru the entrance. You are going to continue moving forward, you will see a staircase in front of you. And you are going to stop right when you are near the banister to the staircase, on the left of you is going to be a corner with a statue and to the right of you is going to be a seating area with 2 wicker chairs and you are done.

Turn left	Begin facing some shelves. Turn around and head out the open doors. Head to the dining table and **turn left.** Head down the left side of the dining table until you reach the living area. **Turn left** and go to the random swing from here head to the white chair in the corner of the room on the elevated platform under the odd art and you're done.
Turn right	You're in a living room. **Turn right** and you'll see a small hallway. Go into it. Toward the doors you can see that are horizontally slatted with wood. In the hallway on the left you will see a table with lots of photos on it. Go toward the table. Look at the table look right. You'll see another room in the distance with a large rectangular table with various boxes and a lamp on it. Step toward there, you'll see its a bedroom. Step to the foot of the bed, look right walk over to the single chair to the right of the bed. Step into the corner left of that chair and stop.
Walk towards the XX (*object*)	We are standing inside an empty walk in closet. We are going to head out inside the bedroom. **Walk towards the bed** and outside on the balcony. Stop when you're outside on the balcony overlooking the city. That's it.
Walk towards the XX (*room type*)	You're in a bathroom with wooden floors and wooden walls. There's a bathtub in front of you. Walk around the bathtub. To your right you see a toilet, a cabinet, a sink and a mirror. In front of you there's a doorway exiting the bathroom. Walk towards this doorway. Continue to walk towards the door. Exit outside into the main room. You're now in the main room. It also has wooden walls and wooden floors. There's a kitchen in front of you. **Walk towards the kitchen.** You're now in the kitchen, to your right you can see some cabinets, a sink, a table and you've reached the end.

Table 2. Examples of templates from RxR [1] training dataset.

Method	NE	OE	SR	SPL	nDTW	sDTW
HAMT	7.75	5.48	42.49	39.33	54.01	35.05
HAMT–tf	4.92	3.43	52.94	50.75	72.41	47.98

Table 3. We report scores for teacher forcing part of ground truth (HAMT–tf) vs No teacher force (HAMT). We find by forcing the agent until the end of partial ground truth, there is no performance drop but increase across all metrics than otherwise.

5. Teacher Forcing Effects

We run a small experiment to verify agents continue to behave rationally after being forced through the intervention trajectories. Consider a truncated (τ, I), pair, we replace the I with full instruction I_f. Given full instruction I_f, agents were either forced until the final node of τ then started to take argmax actions, or without teacher forcing along τ at all. Tab. 3 indicates no performance drop occurred due to the teacher forcing process. (Perhaps unsurprisingly, teacher forcing brings a 10% performance increase.)

References

[1] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020. 6, 8

[2] Shiquan Ren, Hong Lai, Wenjing Tong, Mostafa S. Aminzadeh, Xuezhang Hou, and Shenghan Lai. Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, 37:1487 – 1498, 2010. 1

[3] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2021. 6

[4] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of NAACL-HLT*, pages 2610–2621, 2019. 6