Appendix

In this Appendix, we provide additional elaboration on aspects omitted in the main paper.

- Appendix A: Elaborate derivation of back-door and front-door adjustments.
- Appendix B: In-depth comparison of the four VLN datasets and corresponding metrics.
- Appendix C: Additional experimental results and in-depth discussions on GOAT.
- Appendix D: Analysis of failure cases and comprehensive discussions of limitations.
- Appendix E: Additional qualitative panoramic visualizations from diverse datasets.

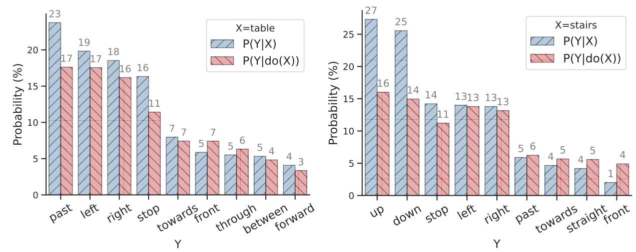
Code Availability: We assure readers that we will make our code and model checkpoints publicly available.

A. Causal Inference Principles

A.1. Back-door Adjustment

In the realm of causal inference [52], the back-door adjustment method serves as a cornerstone, enabling researchers to estimate causal effects from collected data. It hinges on understanding causality, allowing the assessment of the impact of an independent variable X on a dependent variable Y while minimizing the influence of confounders Z. It is essential to distinguish between *"observation"* – passive observation of natural relationships (typically formulated as $P(Y|X) = \sum_z P(Y|X,z)P(z|X)$) – and *"intervention"* – active manipulation of variables to establish causality, denoted as $P(Y|do(X))$. The *do*-operator signifies an intervention where X is forcibly set to a specific value x, thus blocking back-door causal paths originating from X.

To illustrate, consider $P(Y|X)$ and $P_m(Y|X)$ as probabilities before and after intervention on the causal graph, respectively, where $P(Y|do(X)) = P_m(Y|X)$. Calculating the causal effect relies on the observation that P_m, the manipulated probability, shares two crucial properties with P. First, the marginal probability $P(Z = z)$ remains unchanged under intervention since the process determining Z is not affected by removing the arrow from Z to X, denoted as $P_m(z) = P(z)$. Second, the conditional probability $P(Y = y|X = x, Z = z)$ is invariant, because the process by which Y responds to X and Z remains consistent, regardless of whether X changes spontaneously or by deliberate manipulation, *i.e.*, $P_m(Y|X,z) = P(Y|X,z)$. Additionally, the independence between Z and X under the intervention distribution leads to another rule: $P_m(z|X) = P_m(z)$.



(a) (b)

Figure 10. A toy experiment of the differences between the likelihood before (*i.e.*, $P(Y|X)$) and after intervention (*i.e.*, $P(Y|do(X))$) in the R2R training dataset. Only several cases are visualized to avoid clutter.

Considering these equations together, we can derive:

$$P(Y|do(X)) := P_m(Y|X) \tag{25}$$
$$= \sum_z P_m(Y|X,z)P_m(z|X) \tag{26}$$
$$= \sum_z P_m(Y|X,z)P_m(z) \tag{27}$$
$$= \sum_z P(Y|X,z)P(z). \tag{28}$$

Eq. (28) is called the *back-door adjustment formula*. It computes the association between X and Y for each value z of Z, then averages over those values. This procedure is referred to as "adjusting for Z". This final expression can be estimated directly since it consists only of conditional probabilities. To better understand the concept of intervention and meanwhile demonstrate its effectiveness, we conducted a toy experiment based on Eq. (29) and Eq. (30) using direction-and-landmark keywords extracted from instructions in the R2R training dataset:

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \tag{29}$$

$$P(Y|do(X)) = \sum_z \frac{P(Y,X,z)P(z)}{P(X,z)} \tag{30}$$

As depicted in Fig. 10, it is evident that $P(Y|do(X))$ diverges from $P(Y|X)$, supporting our hypothesis that keywords within the instructions function as confounders. To illustrate, consider Fig. 10(a) where X represents a `table`. Previously biased probabilities associated with actions like `past`, `left`, and `right` become more balanced. In other words, when the agent encounters a table, its likelihood to move forward, left, and right becomes evenly distributed, mitigating the erroneous tendency introduced by dataset biases. Likewise, in Fig. 10(b), the intricate probabilities surrounding `stairs` are unraveled, leading to a narrowing down of `up` and `down` probabilities to harmonize with other feasible actions like `stop`, `left`, and `right`. Therefore, by introducing the *do*-operator to realize the active adjust-

Task	Dataset	Train		Val Seen		Val Unseen		Test Unseen		Avg.	Avg.
		Instr	House	Instr	House	Instr	House	Instr	House	Edge	Word
Fine-grained	R2R [4]	14,039	61	1,021	56	2,349	11	4,173	18	5	29
	RxR-En [29]	26,464	61	2,939	56	4,551	11	4,085	18	8	78
Goal-oriented	REVERIE [53]	10,466	60	1,423	46	3,521	10	6,292	16	5	18
	SOON [81]	2,779	34	113	2	339	5	615	14	9	39

Table 8. Dataset statistics. This table provides an overview of each split, including the number of instructions and houses, along with the average edge and average word count for each dataset.

ment $P(Y|do(X))$ rather than merely passive observation $P(Y|X)$ during data fitting, the spurious correlations and underlying biases are alleviated.

A.2. Front-door Adjustment

While the back-door adjustment formula is effective to control for observable confounders, the front-door adjustment method steps in when the confounders cannot be directly observed. In essence, the front-door adjustment method tackles unobservable confounders by identifying alternative pathways that mediate the relationship between the input and the outcome. This nuanced approach is particularly valuable when dealing with intricate causal structures where certain variables are beyond direct measurement.

Concretely, an observable mediator M is inserted between the input X and the output Y, creating a front-door path $X \rightarrow M \rightarrow Y$. First, it's important to highlight that the influence of X on M can be identified, as there are no back-door paths from X to M. Thus, we can obtain

$$P(M|do(X)) = P(M|X). \tag{31}$$

Furthermore, it's crucial to recognize that the impact of M on Y is identifiable. This is because the back-door path from M to Y – specifically, $M \leftarrow X \leftarrow Z \rightarrow Y$ – can be blocked by conditioning on X:

$$P(Y|do(M)) = \sum_{x'} P(Y|M,x')P(x') \tag{32}$$

where x' denotes the possible value of the whole inputs, rather than the current input $X = x$. Both Eq. (31) and Eq. (32) are obtained through the adjustment formula. Subsequently, the front-door adjustment formula can be obtained by chaining these two partial effects:

$$P(Y|do(X)) = \sum_{m}(P(Y|do(M))P(M|do(X))) \tag{33}$$

$$= \sum_{m}\sum_{x'} P(Y|m,x')P(x')P(m|X). \tag{34}$$

The integration of adjustment formulas, incorporating both the back-door and front-door criteria, encompasses diverse scenarios. By leveraging graphs and their underlying assumptions, we can more effectively discern causal relationships and derive causal representations from purely observational data. Motivated by the substantial potential of causal inference, this paper primarily focuses on approximating these adjustments for implementation in deep learning-based methods for VLN. To the best of our knowledge, this is the first work to explain VLN's hidden bias problem from the causal perspective and make an attempt to remove the effect caused by confounders via intervention.

B. Datasets

B.1. Comparison of Various VLN Datasets

The statistical overview and comparison of the four VLN datasets are presented in Tab. 8.

1. Fine-Grained VLN Datasets, including R2R [4] and RxR [29], offer detailed, step-by-step navigational instructions. Specifically, R2R is proposed to guide agents across rooms based on language instructions. RxR, an extension of R2R, augments the complexity with more intricate instructions and paths. To align with other VLN datasets, we focus on RxR's English subsets (en-IN and en-US). What sets the VLN challenge apart is the agent's necessity to follow varied language commands in previously *unseen* real environments. This demands a high level of generalization capability, enabling adaptation to diverse situations.

2. Goal-Oriented VLN Datasets such as REVERIE [53] and SOON [81] emphasize object localization tasks, where agents must find specific objects based on remote referring descriptions. With additional object annotations, these goal-oriented datasets describe the target object and its location with concise instructions. The dataset splits of SOON outlined in DUET [9] are used for ensuring a unified evaluation approach. The goal-oriented VLN task enhances the high-level reasoning abilities of embodied agents, and provides valuable applications in real-world scenarios.

B.2. Evaluation Metrics

For fine-grained VLN tasks, the agent is expected to follow a specific path to reach the target location. The primary metric used to evaluate performance is the `Success Rate` (SR), indicating how often the agent completes the task within a certain distance (usually 3m) of the goal. Additionally, `Navigation Error` (NE) measures the average distance between the predicted and ground-truth locations. `Oracle Success Rate` (OSR) assesses whether any node in the predicted path is within a threshold of

CVPR #1647

Id	Method	SR↑	SPL↑	NE↓	OSR↑
1	**Full Model**	**77.82**	**68.13**	**2.40**	**84.72**
2	BACL w/o Text	77.31	66.37	2.51	84.55
3	BACL w/o Vision	75.95	65.31	2.65	83.78
4	FACL w/o Text	77.01	67.32	2.51	84.46
5	FACL w/o Vision	76.97	67.07	2.51	83.23
6	FACL w/o History	77.18	66.01	2.56	84.42
7	Dict. w/o Update	76.46	66.20	2.65	83.78
8	w/o AGF	77.61	67.02	2.48	84.59
9	Inter. Only Final	76.29	66.80	2.58	84.55

Table 9. More ablation studies on the R2R val-unseen split.

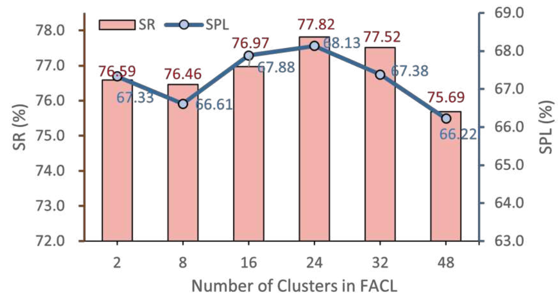


Figure 11. Effect of numbers of clusters in FACL.

the target location. Success weighted by Path Length (SPL) is used to balance both success rate and trajectory length. Since RxR includes paths that approach the goal indirectly, two additional metrics are considered. Normalized Dynamic Time Warping (nDTW) penalizes deviations from the reference path to measure the match between two paths. Success weighted by normalized Dynamic Time Warping (sDTW) refines nDTW, focusing solely on successful episodes, thereby capturing both success and fidelity. For goal-oriented tasks, the primary focus is on the agent's proximity to the goal. In addition to the above metrics, Remote Grounding Success Rate (RGS) is used to assess the accuracy of selecting the object from a set of candidates at the final position. Remote Grounding Success Rate Weighted by Path Length (RGSPL) is introduced to account for both success rate and path length.

C. Additional Experimental Results

C.1. Confounder Factors Variation

To explore the contribution of various modalities to causal learning, we further conducted detailed ablation studies on each modality. As demonstrated in #2 – #6 in Tab. 9, different ablations lead to varying degrees of performance degradation, substantiating our hypothesis regarding confounders in VLN systems. Specifically, in BACL, the ablations of textual and visual intervention result in decreases in SR by 0.51% and 1.81%, and SPL by 1.76% and 2.82%, respectively. This suggests that visual intervention plays a crucial role, which is reasonable given that the primary distinction between seen and unseen environments lies in visual observation. In FACL, the ablations of textual, visual, and historical intervention lead to reductions in SR by 0.81%, 0.85%, and 0.64%, and SPL by 0.81%, 1.06%, and 2.12%, respectively. The adjustment to history has relatively more significant performance gains. Overall, these findings emphasize the importance of comprehensive interventions across cross-modal inputs, yielding more unbiased features and more generalized decision outcomes.

C.2. Update of Confounder Dictionary

In Tab. 9 #7, we investigate the impact of updating confounder dictionaries during training. This involves the textual confounder dictionary in BACL, supported by RoBERTa's end-to-end training, and random sampling from k-means clusters in FACL. The dictionaries are updated either when the model achieves a new best performance in the val-unseen split or every 3,000 iterations. The results demonstrate that updating the dictionary features aligns the representations of confounders more effectively with the evolving model weights and also enhances diversity, leading to improvements in overall performance (↑ SPL 1.93%).

C.3. Adaptive Gate Fusion

In Tab. 9 #8, we analyze the effects of the AGF module, designed to adaptively fuse causality-enhanced features and original context features using a gate-like structure. "W/o AGF" signifies the direct use of causality-enhanced features without combining them with context features. The results demonstrate that the adaptive fusion process enables the model to effectively incorporate both types of features, leading to comparatively higher performance (↑ SPL 1.11%).

C.4. Intervention Location

In Tab. 9 #9, we validate the effectiveness of extending the assumption of causal learning to hidden features, rather than focusing solely on outputs. The results indicate that incorporating intervention modules only before the final Softmax layer enhances generalization capabilities to some extent. However, applying these interventions in shallower layers yields superior performance (SPL 68.13 *vs.* 66.80). This extended assumption renders the application of causal learning in deep learning methods more flexible and practical.

C.5. Number of K-Means Clusters in FACL

Given that the confounder addressed by FACL is unobservable, we employ the K-Means algorithm to cluster the global features extracted by the trained CFP module from the entire training dataset. During the fine-tuning phase, we periodically sample features from these clusters to integrate

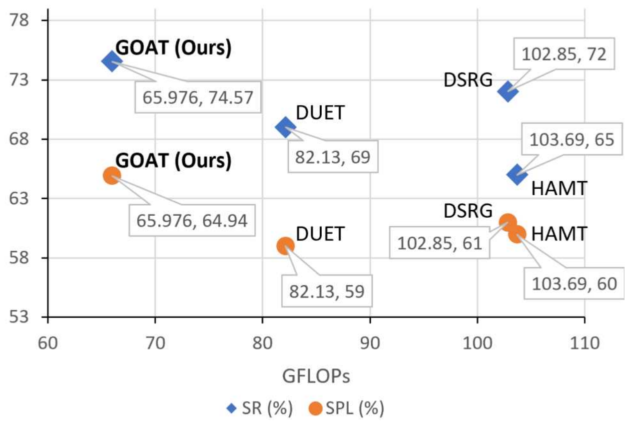


Figure 12. Comparison of GFLOPs and Accuracy.

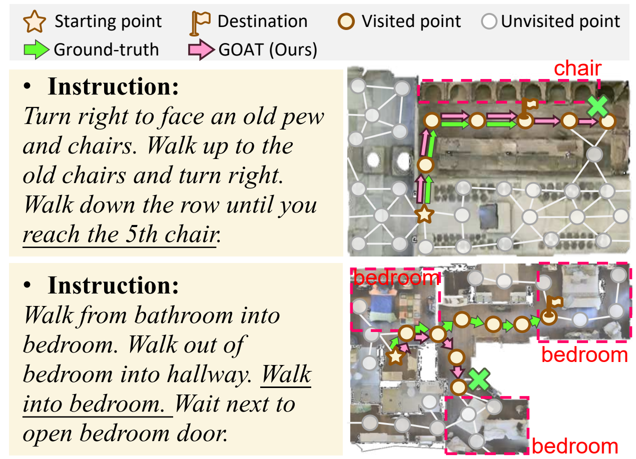


Figure 13. Illustration of Failure Cases.

into training. The experimental results of determining the number of categories for clusters are shown in Fig. 11. It presents that the choice of 24 clusters in FACL yields optimal performance in both SR and SPL. This strategic clustering approach ensures a comprehensive coverage of potential categories, enhancing the model's ability to discern confounders. Notably, too few clustering categories may overlook crucial distinctions, whereas an excessive number introduces redundant computational overhead and irrelevant noise, ultimately hampering training performance.

C.6. Efficiency and Effectiveness Comparison

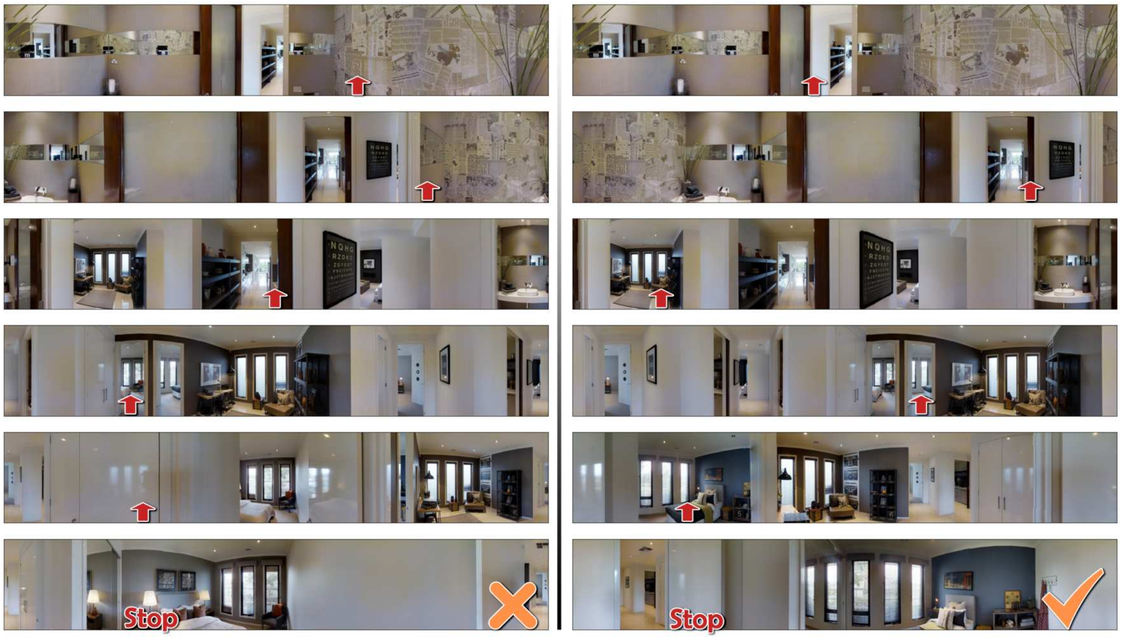
It is necessary to consider both efficiency and effectiveness since VLN is prompted to be applicable in real-world robots in the future. To assess computational complexity, we employed the Python toolkit `thop`, comparing GFLOPs with other transformer-based methods. For fair comparison, we conducted single-step forward inference with a batch size of 8, instruction length of 44, and historical global graph node of 6 across all methods. As shown in Fig. 12, GOAT strikingly balances efficiency and effectiveness, outperforming previous approaches in both SR and SPL while maintaining lower GFLOPs. This reduction in computational cost is attributed to the adoption of a lighter framework with fewer transformer layers. This discovery illustrates that in scenarios with restricted task-specific datasets, adopting a lighter framework can enhance generalization while significantly reducing computational costs.

D. Analysis of Failure Cases and Limitations

Despite GOAT's remarkable performance, we also examined specific failure cases to shed light on its limitations. For instance, as depicted in Fig. 13, GOAT struggles with instructions involving numerical references. In the first case, it misidentified the `5th` chair but arrived at the `8th` chair instead. This phenomenon is aligned with the problem of current large models that are not sensitive to numbers and arithmetic tasks. Addressing this issue could benefit from approaches like the chain-of-thought method [72], which has shown promise in handling numerical tasks. Moreover, when instructions are initially ambiguous (*e.g.*, in the second case, there are actually `two bedrooms` that fit the description), GOAT might select the wrong option. Utilizing datasets like [47, 59], which focus on human-agent interaction, could improve the agent's decision-making in response to ambiguous instructions. Incorporating such datasets could empower the agent with a more robust and practical interactive capacity, reducing the likelihood of erroneous predictions. Finally, the limitation to the integration of causal learning with deep-learning methods, including the approximation process inherent in calculating expectations, and distinct modalities exhibiting varying preferences for specific probability estimations, requires ongoing efforts in future research to enhance the interpretability of these discrepancies.

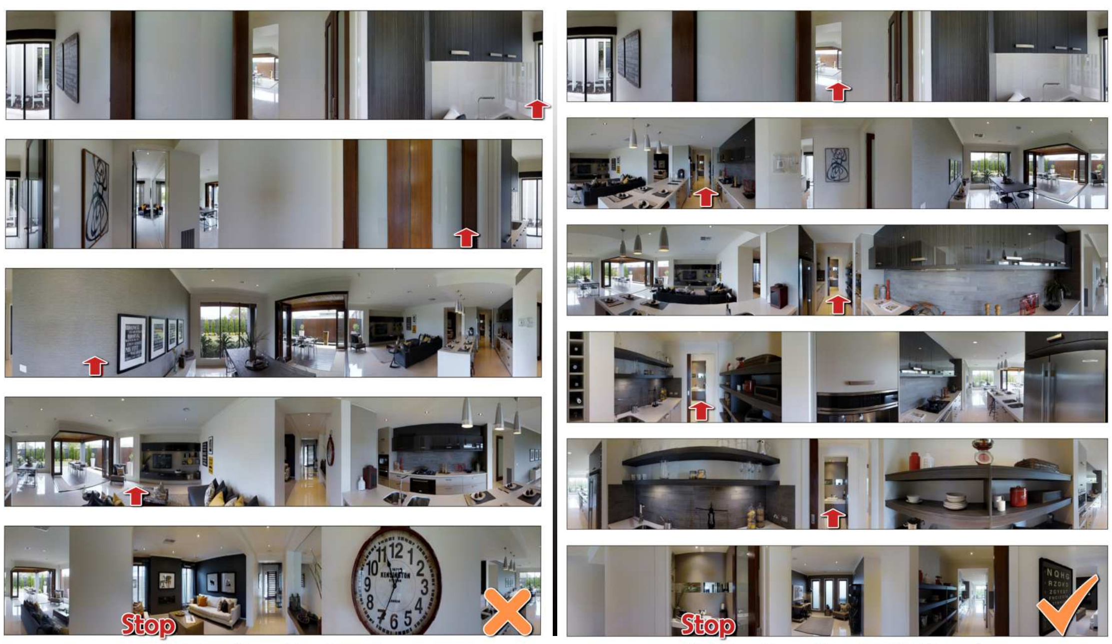
E. Additional Qualitative Examples

Due to space limitations in the text, we present only the top view to depict the navigation scenarios. In this section, we provide predicted panoramic paths on four datasets in Fig. 14, 15, 16, 17, respectively, to enhance readers' comprehension of the tasks' objectives and the effectiveness of our approach. Specifically, the red arrows indicate the forward directions, and the corresponding instructions are provided below the visualized trajectories.



(a) Predicted by DSRG (b) Predicted by GOAT (Ours)

Instruction:
Turn around and exit the bathroom. Once out turn left and head towards the sitting area. Once you reach that area turn left and enter the door to your right, beside the desk. Stop once you are in the doorway of the room.



(a) Predicted by DSRG (b) Predicted by GOAT (Ours)

Instruction:
Walk out of the washroom past the double closet doors and walk into the next room. Walk into the kitchen area and continue along the counter tops past the sink. Continue through the open door at the end of the counter tops.

Figure 14. Visual examples in the R2R validation-unseen split with navigation instructions presented at the bottom.

Figure 15. Visual examples in the REVERIE validation-unseen split with navigation instructions presented at the bottom.

(a) Trajectories Predicted by GOAT (Ours)

Instruction: *We're facing towards a corner of a wall, turn slightly to the right and walk towards the hallway, once you're facing the hallway wall, turn towards the left... On your left there's a staircase, and on your right there's a doorway, walk past the staircase, and walk past the doorway, but before you get into the doorway, you'll face towards a small staircase, walk down that staircase, and then continue straight, and then once you land, sorry, turn towards the left and you'll come across a kitchen... On your left there's a fridge, and on your right there's a counter, and next to the fridge there are cabinets and there's a stove, walk towards the middle between the cabinet and the stove, you'll be facing towards a dishwasher, walk towards it, turn towards the right, the counter is now on your right and on the left there are two sinks, and an island, walk past that island where the sinks are, on your right there's a long wooden table which is surrounded by chairs, it is a dining table, on top of the dining table there's a vase with some beautiful white flowers, on the left there's a clock, walk past the top chair that you see which is in front of the dining table, you're no facing towards the patio area, and that's your destination.*

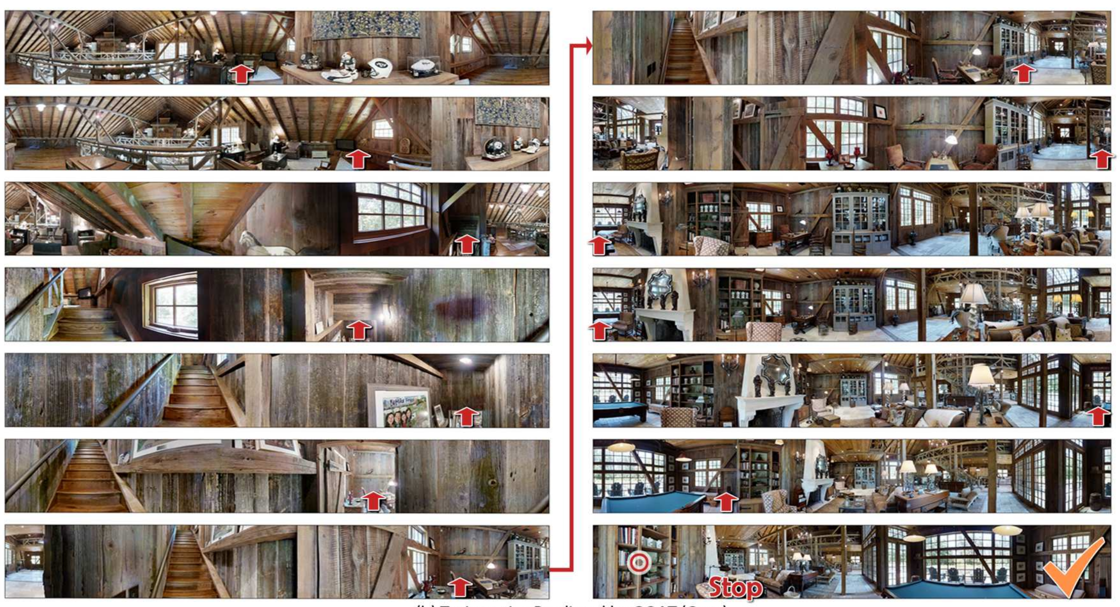
(b) Trajectories Predicted by GOAT (Ours)

Instruction: *You begin in a bathroom looking towards a wall with a painting on it. Turn towards your right and exit the toilet room, and then turn to your right again and exit the entire bathroom into a bedroom. Turn to your right and approach the mirror closet doors, and then turn to the right and exit the bedroom into the hallway. Once you're out there, turn left. You should see a desk with two bar stools tucked under it. Head to the left side of that. That was actually a counter, not a desk, but potato potahto. Continue forward to the left of the small circular table on the end of the black couch, and then turn right and head between the counter and the back of the couch. Continue going forward towards the staircase or the- sorry that is a shuttered door, that you can see in the far distance. Once you are standing in front of the door mat in front of that shuttered door, which says... Metricon... Structure twenty-five years... Quaran- Guarantee. Apologies, the mat is upside-down as you can see. Anyways, once you're standing in front of that shuttered door, with the bright sunlight outside, you're done.*

Figure 16. Visual examples in the RxR validation-unseen split with navigation instructions presented at the bottom.

(a) Trajectories Predicted by GOAT (Ours)

Instruction:
I'd like to find a picture on the wall, opposite to the door of the toilet in the corridor of the bedroom, which is connected with a bedroom, a toilet and a small living room.The picture is rectangular, black and white.



(b) Trajectories Predicted by GOAT (Ours)

Instruction:
Find a tall, woody, gray cabinet which is next to a big window, between a chair. The bookshelf is settled in a spacious and bright living room in front of a dining room and a kitchen. It is located in the first floor.

Figure 17. Visual examples in the SOON validation-unseen split with navigation instructions presented at the bottom.