Supplementary Materials: Navigating Beyond Instructions: Vision-and-Language Navigation in Obstructed Environments

Anonymous Authors

This document provides additional experimental results, more method details, and further analysis to supplement the main paper, including:

- Sec. 1: more experiments on the obstructed environments.
- Sec. 2: additional statistics and visualizations of the R2R-UNO dataset.
- Sec. 3: qualitative examples of the generated environments and navigation process.
- Sec. 4: a discussion on the performance gap between original and obstructed environments.
- Sec. 5: discussions on limitations and future directions of this work.

1 MORE EXPERIMENTS

In this section, we further investigate the factors impacting agent performance when meeting instruction-reality mismatches and the different situations for high-level instructions in REVERIE [9]. All experiments follow the implementation details in the main paper. The code is available at https://anonymous.4open.science/r/ObstructedVLN-D579.

1.1 Data Augmentation

In the official DUET [4] implementation [1] for training R2R [1], no augmented data is used since the model quickly reaches peak performance in unseen scenarios with limited gold data from R2R. However, such rapid optimization is not ideal when simultaneously training in R2R and R2R-UNO. Given the increased complexity and our curriculum training strategy, agents achieve peak performance in obstructed environments significantly later than in original ones, which can cause over-fitting and prevent optimal performance for both tasks with such limited data for unobstructed environments.

To address this issue, we propose utilizing synthesized instructions from PREVALENT [5] as augmented data to enrich the training data for original environments. While these data may not directly enhance performance on R2R, they assist in maintaining agent performance in R2R until the best performance is achieved in obstructed environments. In Tab. 1, we present an ablation study of this augmented data on the validation unseen splits of R2R and R2R-UNO Block-1 set. For each dataset, we report the best Success Rate (SR), the corresponding step to achieve the best SR, and the ΔSR, representing the success rate gap between these two steps. For example, if DUET reaches its best performance at step c_1 for R2R and c_2 for R2R-UNO, ΔSR For R2R is calculated as $SR_{R2R}^{c_1} - SR_{R2R}^{c_2}$. A lower ΔSR value indicates a closer alignment of these two peaks, which is desirable for achieving optimal performance for both datasets. The results prove that using PREVALENT significantly narrows the performance gap between the two peaks. With a ΔSR of only 1.0% for R2R, the augmented data effectively helps agents maintain their performance in original environments and balance the training across the two tasks.

Table 1: Ablation study of augmented data on R2R and the Block-1 set of R2R-UNO.

Model	Best in R2R			Best in R2R-UNO		
	Step	SR↑	ΔSR↓	Step	SR↑	ΔSR↓
Vanilla	3K	72.0	4.5	40K	68.5	11.3
PREVALENT	16K	**72.9**	**1.0**	54K	**68.9**	**6.4**

Table 2: Navigation performance with different training data on REVERIE and the Block-1 set of R2R-UNO.

Env	REVERIE			R2R-UNO		
	SR↑	SPL↑	RGS↑	SR↑	SPL↑	RGS↑
Original	49.8	**35.0**	**35.2**	34.3	24.6	24.6
Obstructed	48.4	31.7	33.6	**38.4**	**25.7**	**26.3**
Both	**50.2**	32.6	34.5	37.5	25.3	25.6

1.2 Different Situations on REVERIE

Unlike R2R, datasets with high-level instructions will not encounter instruction-reality mismatches since the instructions only describe the destination without specifying the path, making our obstructed environments only additional unseen data for the same task. To validate this, we conducted experiments within the REVERIE dataset [9]. Tab. 2 shows the performance of DUET when trained on different environments of REVERIE, evaluated on the validation unseen splits.[2] We use the Block-1 set of R2R-UNO for the obstructed environments, as R2R and REVERIE share the same paths within the Matterport3D dataset [2]. We further incorporated the Remote Grounding Success (**RGS**) metric [9] to assess the efficacy of object grounding. The observed results are markedly different from those obtained with R2R and R2R-UNO in the main paper. Training with only obstructed environments can achieve results comparable to those using the original REVERIE, demonstrating that original and obstructed environments are the same task when high-level instructions are provided. In the more challenging R2R-UNO dataset, including obstructed environments improves about 4% SR and 1% RGS, which is relatively minimal compared to the 20% performance boost with R2R instructions. Although our obstructed environments do not introduce instruction-reality mismatches with high-level instructions, they still present a more challenging version of the

[1] https://github.com/cshizhe/VLN-DUET

[2] The results exceed those reported in the original DUET paper due to a larger maximum action length.

Table 3: Navigation performance of DUET and HAMT on the val seen and unseen splits of R2R and R2R-UNO datasets. The setting number γ includes all Block-x sets with $x \leq \gamma$. $\gamma = 0$ means only using R2R for training.

Model	Split	γ	Block-1		Block-2		Block-3		R2R	
			SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
HAMT	Val Seen	0	41	36	28	24	21	19	**76**	72
		1	64	**60**	53	49	48	44	75	72
		2	65	59	57	53	51	49	**76**	**73**
		3	**66**	**60**	**61**	**56**	**57**	**53**	75	72
	Val Unseen	0	34	30	22	19	16	14	65	59
		1	**52**	**46**	41	36	35	31	66	**61**
		2	50	44	41	**37**	36	32	65	60
		3	51	43	**43**	**37**	**38**	**33**	**67**	**61**
DUET	Val Seen	0	50	44	34	30	25	22	**80**	**75**
		1	**77**	**71**	70	65	65	62	79	71
		2	**77**	69	**76**	68	**74**	67	79	70
		3	**77**	**71**	75	**71**	**74**	**70**	**80**	72
	Val Unseen	0	44	36	31	25	23	20	**72**	**60**
		1	**69**	**55**	60	50	53	45	**72**	58
		2	67	53	63	52	59	50	71	56
		3	67	53	**65**	**54**	**63**	**54**	**72**	57

original dataset, making the results in obstructed environments significantly lower than the original REVERIE.

1.3 Impact of the Number of Obstructed Edges

To find out the relationship between different numbers of obstructed edges, we train DUET [4] and HAMT [3] models using R2R and different sets of R2R-UNO, as shown in Tab. 3. In this context, each setting γ contains all Block-x sets with $x \leq \gamma$ in R2R-UNO. For example, the model with $\gamma = 3$ is trained using R2R and all three sets of R2R-UNO. The results indicate a positive correlation between the inclusion of data with increased obstructed edges and improved agent performance in both HAMT and DUET models on their respective evaluation sets. Notably, training with all three R2R-UNO sets brings the best performance on Block-2 and Block-3 sets while also maintaining strong results on Block-1 and R2R. The $\gamma = 1$ setting also performs well on Block-2 and Block-3 sets, suggesting that the skills required to navigate obstructed environments and adapt to instruction-reality mismatches are universal for different situations.

1.4 Hyper-Parameters in Learning Strategy

We conduct an ablation study to evaluate the impact of the maximum proportion of obstructed environments α_{max} in our ObVLN and the task-wise sample strategy, as shown in Tab. 4. The $\alpha_{max} = 0$ corresponds to training only with R2R, which performs poorly in obstructed environments and is included in the main paper, so we omit it here. The results reveal the trade-off between performances in original and obstructed environments. Intuitively, increasing α_{max} leads to the enhanced agent performance in obstructed environments at the expense of performance in original ones, and the reverse is also true. Although $\alpha_{max} = 1$ achieves the highest results in the R2R-UNO, it significantly compromises performance in the original environments, which is unacceptable since real-world agents must operate effectively across both environments. Considering the overall performance of both datasets, our method outperforms task-wise sampling.

Table 4: Ablation study of different α_{max} and c values on the val unseen splits of R2R and the Block-1 set of R2R-UNO.

c	α_{max}	R2R		R2R-UNO	
		SR↑	SPL↑	SR↑	SPL↑
0K	0.25	72.0	59.1	66.2	51.0
	0.50	70.9	58.1	67.0	54.5
	0.75	69.2	55.3	68.4	55.0
	1.00	62.4	41.2	**70.8**	**57.7**
20K	0.25	**72.5**	**59.5**	65.1	53.4
	0.50	72.3	58.3	68.5	54.9
	0.75	71.3	57.4	68.6	56.9
	1.00	70.2	55.4	69.0	55.9

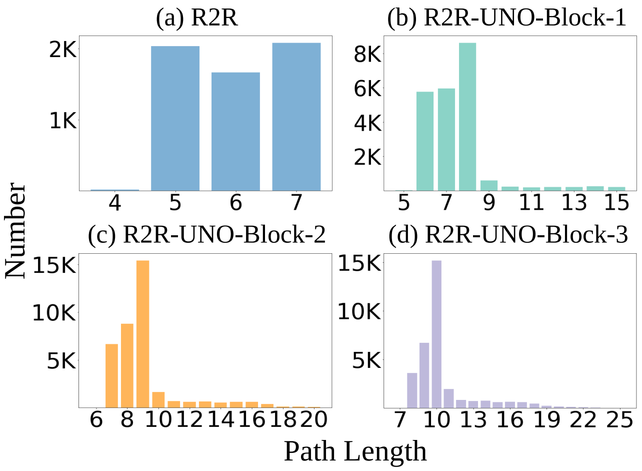

Figure 1: The path length distribution in R2R and three sets of R2R-UNO, including all splits.

2 MORE R2R-UNO STATISTICS

In this section, we provide more visualizations of the R2R-UNO dataset, including the path length distributions, the category distribution of inpainting objects, and the distribution of CLIP [11] scores approximated by Gaussian Mixture Models (GMMs).

Path Length Distribution. Fig. 1 shows the path length distribution of R2R and the three sets of R2R-UNO. It can be observed that most paths include one or two additional nodes following the obstruction of an edge, indicating that detours are always close to the obstructions. This is aligned with real-world obstructions and represents that most cases in obstructed environments are relatively manageable. Despite this, current state-of-the-art VLN methods still perform poorly in zero-shot evaluations within such contexts, emphasizing the significance of our contributions. Additionally, the presence of some significantly lengthened paths ensures that the

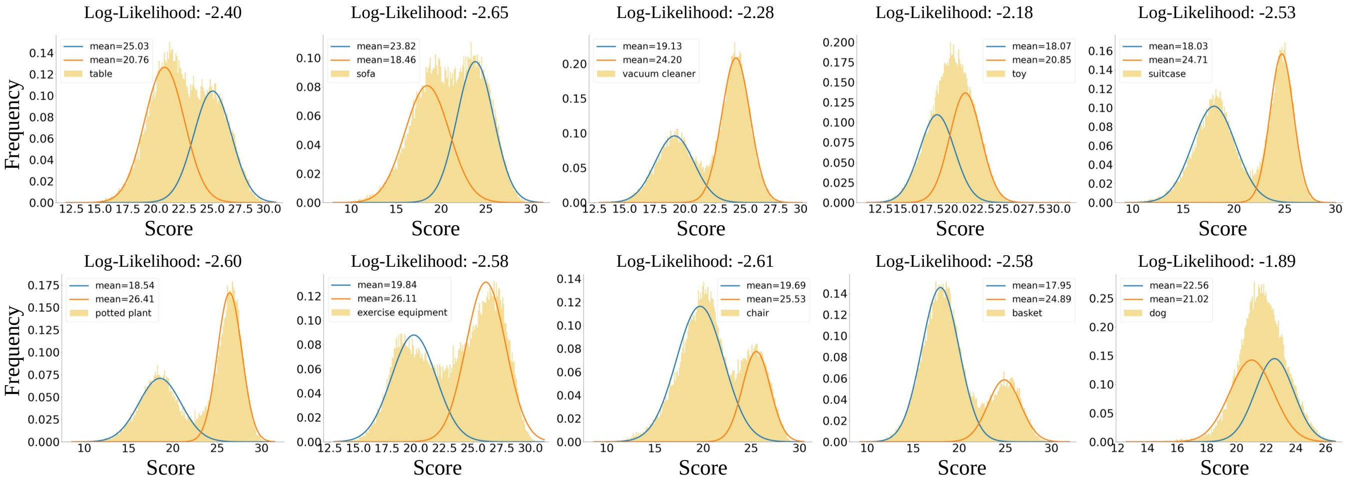

Figure 2: The CLIP score distributions for ten object categories in R2R-UNO and their corresponding curves of GMMs.

R2R-UNO dataset also possesses sufficient complexity and challenges for the instruction-reality mismatches.

Compatibility Score Distribution. We present the compatibility score evaluated by CLIP and corresponding curves of the trained bimodal Gaussian Mixture Models (GMMs) in Fig. 2. Most object categories follow the bimodal Gaussian distribution. However, *'toy'* and *'dog'* are out of this pattern since they are rarely seen in original environments, resulting in a high probability of obstruction generation like an unimodal Gaussian distribution. Despite this, using a bimodal Gaussian distribution can also help select those high-quality candidates, so we treat all categories equally. We employ the log-likelihood values to verify the effectiveness of our GMMs in Fig. 2. The results demonstrate the strength and reliability of our method in integrating diverse obstructions into existing views.

Object Category Distribution. Fig. 3 illustrates the distribution of ten object categories in the R2R-UNO dataset. Each category comes with at least 1,500 instances of obstruction, with eight out of them surpassing 3,000 instances, proving the great diversity of our generated obstructions. The categories 'Basket' and 'Chair' are less frequent due to the greater generation challenges and high occurrence in the inpainting backgrounds.

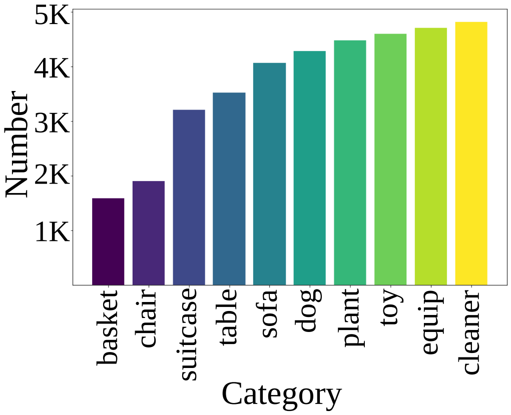

Figure 3: The object category distribution in R2R-UNO. *'plant'* is short for *'potted plant'*. *'equip'* represents *'exercise equipment'* and *'cleaner'* means *'vacuum cleaner'*.

3 QUALITATIVE EXAMPLES

In this section, we provide more qualitative examples of the view projection process in the object insertion module and the agent trajectories in obstructed environments.

3.1 View Projection

Fig. 4 displays qualitative examples of the view projection results and the final panoramic view in R2R-UNO. The inpainted regions of the target view are projected into eight adjacent views and combined with the original images to form novel views, as shown on the left part. We stitch these discrete views together to generate the panorama on the right, which is natural and photo-realistic.

These examples demonstrate the effectiveness of our method in integrating novel obstructions into real scenes to generate obstructed environments.

3.2 Navigation Trajectory

We provide an example to show the adaptability of agents trained with ObVLN in obstructed environments in Fig. 5. When encountering an obstruction along the instructed path, the ObVLN agent efficiently identifies an alternate route and continues following instructions to reach the destination. However, agents lacking training in such instruction-reality mismatches get confused and lose their way, leading to an unsuccessful navigation trajectory. This scenario demonstrates the significance and effectiveness of introducing obstructed environments and the ObVLN method.

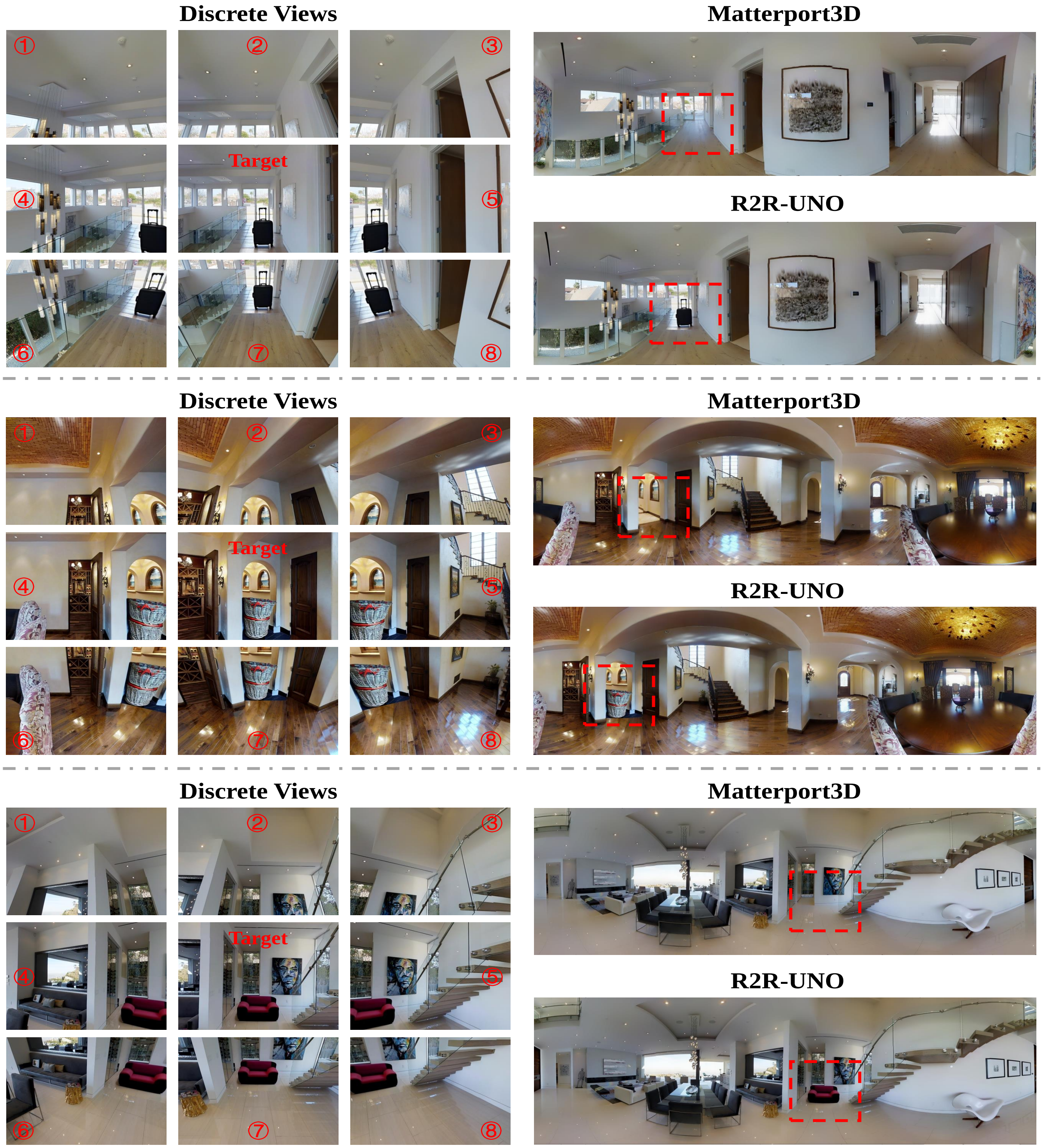

Figure 4: Qualitative examples of the view projection results and the final panoramic views in R2R-UNO. On the left, we display the inpainted target view and the eight adjacent views from the projection. On the right, we perform a comparison between the original Matterport3D panorama and the novel panorama with obstructions in R2R-UNO. The red dash line highlights the changed areas. Note that slight horizontal shifts between the two panoramas may occur as a result of the image stitching process employed (source: https://github.com/OpenStitching/stitching).

Instruction: *"Walk out the door in front of you and turn left. Once you reach the black shelf, turn right and enter the hallway. Turn into the first door on your left and stop once you enter the bathroom."*

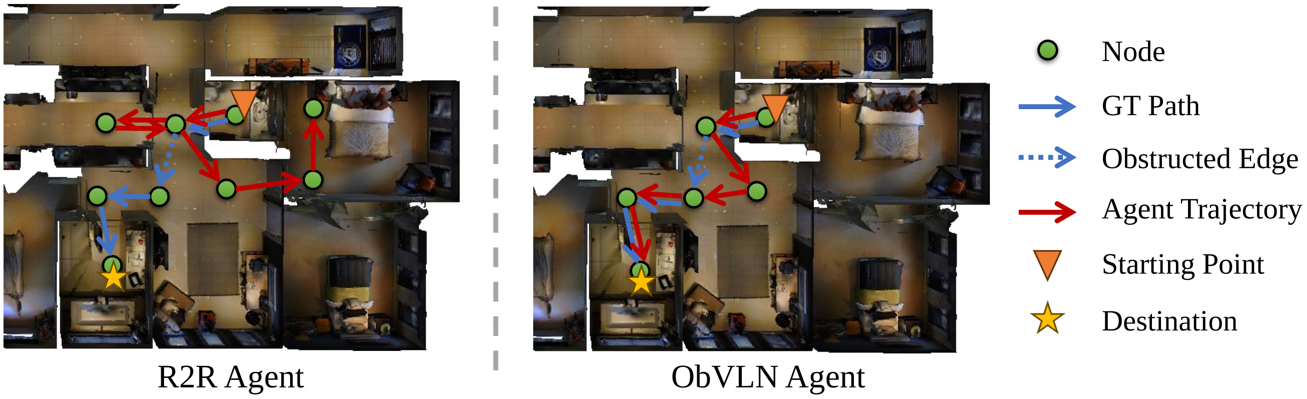

Figure 5: A visual comparison between trajectories of DUET trained with R2R and those enhanced by our ObVLN. When facing the obstruction, the standard DUET (left) fails to find the detour, resulting in an unsuccessful trajectory. In contrast, agents trained with ObVLN (right) effectively find an alternate path and arrive at the destination.

4 PERFORMANCE GAP BETWEEN R2R AND R2R-UNO

In Sec. 4.1 of the main paper, we highlight the challenges faced by state-of-the-art VLN methods in obstructed environments, proved by their poor zero-shot performance on the R2R-UNO dataset. However, the significant performance degradation could also be attributed to factors other than instruction-reality mismatches, thus making the impact of the “perfect instruction assumption” unclear. Therefore, we further discuss these potential factors here to support our claim that instruction-reality mismatches are the primary cause of the observed performance decline. We consider four key factors: 1. Visual changes from inpainting. 2. Path number imbalance. 3. Increased navigation complexity. 4. Instruction-reality mismatches.

Regarding the first factor, the results in Tab. 3 of the main paper demonstrate that even in the absence of visual changes as the visual feedback, DUET maintains a 60% success rate, which is much higher than the 40% when zero-shot evaluation. We also experiment with only modifying visual observations while maintaining the navigation graphs, the vanilla DUET agents can achieve a 71% SR under the Block-1 observations. These findings indicate that visual changes are not the primary cause of the performance drop.

We take the second factor into consideration because R2R paths with different numbers of redundant edges will result in different numbers of modified paths in the R2R-UNO dataset. To eliminate the impact of this shift, we assess the success rate under controlled conditions by marking all successful paths in R2R also as successful in R2R-UNO and vice versa. The SR on the validation unseen split of DUET slightly decreased from 72% to 71%, suggesting the limited impact of this factor.

For factors three and four, we compare the differences in performance gain with and without the obstructed environments as training data between R2R (Tab. 2 in the main paper) and REVERIE (Tab. 2). R2R with both factors gets a substantial 20% SR improvement after training in obstructed environments. In contrast, REVERIE shows a slight 4% SR difference without the instruction-reality mismatches. This difference suggests that the instruction-reality mismatches are the main contributor to the performance drop, emphasizing the significance of our work.

5 DISCUSSION

Limitations. Although obstructed environments are designed to help agents adapt to instruction-reality mismatches caused by real-world dynamics such as unexpected obstacles, the development of these agents is currently limited to simulators rather than actual robots. This requires further advancements in continuous environments like VLN-CE [6]. Another limitation is that our object insertion is conducted on 2D images instead of 3D environments, potentially leading to inconsistencies across multi-view observations and depth information. We experiment with state-of-the-art single-image to 3D reconstruction works [7, 10] to map our 2D changes into 3D spaces but get unsatisfactory results. Furthermore, there is still room for enhancing agent performance in obstructed environments to further approach the performance in original settings.

Future Work. In the future, we plan to extend our methods into continuous VLN benchmarks like VLN-CE [6], and incorporate obstruction detection methods to deploy our agents on real robots. This will enable us to assess their performance in real-world navigation scenarios, especially after the introduction of instruction-reality mismatches. We also aim to perform 3D object insertion to the Matterport3D dataset to render obstacles more naturally and realistically. Besides unexpected obstacles, some other potential causes for instruction-reality mismatches also need investigation, including existing object removal, furniture rearrangement [12], and human-in-the-loop navigation [8]. Additionally, we will investigate more training strategies to further enhance agent performance in obstructed environments.

REFERENCES

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*. 3674–3683.

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *International Conference on 3D Vision (3DV)*. 667–676.

[3] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *NeurIPS* 34 (2021), 5834–5847.

[4] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*. 16537–16547.

[5] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*. 13137–13146.

[6] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*. Springer, 104–120.

[7] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*. 9298–9309.

[8] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. 2024. Habitat 3.0: A Co-Habitat for Humans, Avatars, and Robots. In *ICLR*.

[9] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*. 9982–9991.

[10] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. 2024. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In *ICLR*.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.

[12] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2021. Visual room rearrangement. In *CVPR*. 5922–5931.