A. Appendix

A.1. Limitations and Future Work

As we discuss in the main paper, our approach achieves strong but not state-of-the-art results on the R2R dataset, which we attribute to domain differences between R2R and RxR (noting that the Marky instruction generator we use for data augmentation was trained on RxR). One way to address this limitation would be by re-training Marky on R2R data, although this would face some hurdles since R2R lacks the annotator pose traces that were used by Marky when training on RxR.

To better understand the failure modes of our approach, in Figure 4 we plot a distribution indicating the step in the trajectory where an agent makes its first error. We analyze the RxR Val-Unseen split and compare MARVAL to the previous state-of-the-art approaches, EnvEdit and HAMT, as well as human instruction-following demonstrations from the RxR dataset. MARVAL makes fewer errors than the prior approaches, especially at the start of the trajectory, but also *fewer errors than human followers*. Since human followers still significantly outperform MARVAL overall – in terms of navigation error (0.79m vs. 4.49m), success rate (94.5 vs. 64.8) and path-fidelity metrics such as NDTW (81.8 vs. 70.8) – this suggests the main focus for future agent improvement should be on recovering from errors, which human wayfinders clearly do extremely well in order to still reach the goal in 94.5% of episodes.

A.2. Implementation Details

Pretraining In all experiments we train with a batch size of 128 using the AdaFactor optimizer. During pretraining, we use dropout of 0.1 and a learning rate that exponentially-decays from 0.1. We monitor one-step action prediction accuracy on ground-truth trajectories using held-out instructions from unseen environments (RxR Val-Unseen and R2R Val-Unseen). We pretrain until convergence and then select the best snapshot based on one-step action prediction accuracy on RxR Val-Unseen.

Finetuning During finetuning, we use a constant learning rate of 0.001 and dropout of 0.2. Since the human-annotated datasets used for finetuning (RxR and R2R) are relatively small, during finetuning we update only the WordPiece embeddings in the agent and keep all other transformer weights frozen. This makes finetuning more stable and less prone to overfitting (especially on the smaller R2R dataset). We finetune for a maximum of 150K iterations while monitoring standard VLN path-fidelity metrics such as success rate (SR). We select the best snapshot based on SR on Val-Unseen.

A.3. Additional Pretraining Results

In Table 5 we report pretraining results on the *Val-Seen* splits, complementing the *Val-Unseen* results included in the main paper. We observe the same trends in the seen environments as in the new, unseen environments (Val-Unseen), although the relative improvement from using a larger model (row 8 vs. 7) is larger.

A.4. Performance by Language

In Table 6 we report results by language on the RxR Val-Seen and Val-Unseen splits in comparison to the previous state-of-the-art EnvEdit model. Improvements are across-the-board and results on each language are similar.

A.5. Marky Synthetic Instruction Examples

Figures 5 and 6 provide examples of Marky-generated (synthetic) instructions instruction and their associated trajectories.

A.6. MARVAL Instruction-Following Examples

Figures 7 and 8 provide examples MARVAL successfully following instructions from the RxR Val-Unseen split (i.e., in a previously unseen environment). In Figure 9 we include a failure case.

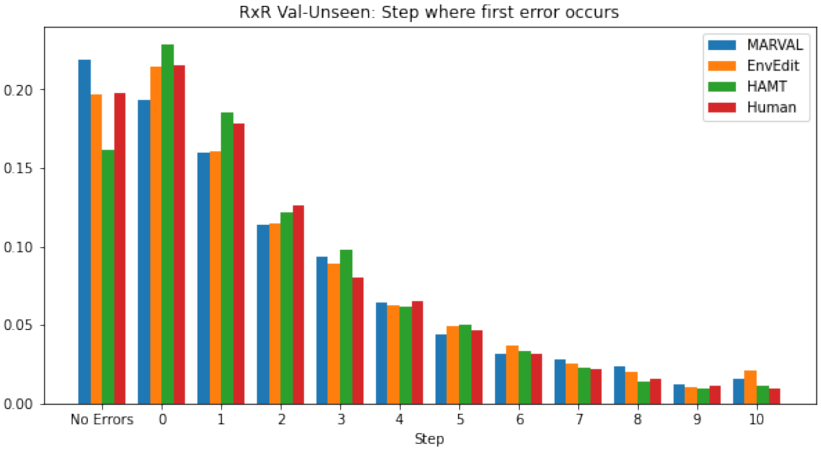

Figure 4. Error analysis indicating at which trajectory step each agent makes its first mistake. Surprisingly, MARVAL makes less errors (produces more perfect trajectories) than the prior work *and human followers*. Since human followers still significantly outperform MARVAL overall, this suggests the main focus for future agent improvement should be on recovering from errors.

		Pretraining Data							RxR Val-Seen				R2R Val-Seen		
	Size	**R2R**	**RxR**	**S-MP**	**M-MP**	**M-Gib**	**SE3DS**	**Iterations**	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**SPL**
1	Base	✓	✓					630K	12.14	22.3	39.3	18.8	7.45	37.9	36.3
2	Base	✓	✓	✓				1.68M	11.52	26.7	42.4	22.9	5.19	52.6	50.5
3	Base	✓	✓	✓	✓			2.94M	7.23	48.5	59.3	42.4	6.19	48.7	45.7
4	Base	✓	✓		✓			2.00M	4.42	64.8	73.2	58.5	3.86	67.2	64.2
5	Base	✓	✓	✓	✓		✓	1.94M	5.15	62.0	69.9	55.6	4.61	59.2	56.2
6	Base	✓	✓	✓	✓	✓		3.00M	5.09	61.4	68.7	51.2	4.77	57.5	53.0
7	Base	✓	✓	✓	✓	✓	✓	2.80M	5.22	60.0	68.3	53.1	4.70	59.6	55.1
8	Large	✓	✓	✓	✓	✓	✓	4.80M	3.85	71.0	75.9	64.3	4.05	66.1	62.6
9	Large	✓	✓		✓	✓	✓	5.14M	**3.62**	**72.7**	**77.0**	**65.9**	**3.73**	**68.2**	**64.9**

Table 5. Comparison of pretraining approaches reporting Val-Seen results on RxR and R2R.

	Val-Seen				Val-Unseen			
Agent	**NE ↓**	**SR↑**	**NDTW↑**	**SDTW↑**	**NE↓**	**SR↑**	**NDTW↑**	**SDTW↑**
English (en-IN):								
EnvEdit* [34]	3.82	68.01	71.45	59.09	4.42	62.03	67.89	54.15
MARVAL	3.10	75.77	78.79	68.86	4.49	64.83	70.67	57.64
English (en-US):								
EnvEdit* [34]	4.22	66.19	69.52	56.95	4.33	61.70	67.56	52.94
MARVAL	3.53	72.22	76.19	64.62	4.46	64.47	70.28	56.46
Telugu (te-IN):								
EnvEdit* [34]	3.72	65.59	70.70	57.37	4.49	61.85	67.84	53.75
MARVAL	2.82	76.38	79.34	69.43	4.56	63.85	69.92	56.30
Hindi (hi-IN):								
EnvEdit* [34]	4.01	68.56	71.90	59.92	4.20	64.54	69.74	56.41
MARVAL	2.95	76.61	80.05	69.52	4.30	65.85	72.02	58.88

*Results from an ensemble of three agents.

Table 6. Breakdown of the results on RxR for each language for our best model and the best performing previous model from [34]. Predicted paths for EnvEdit were provided by the authors.

Marky Instruction: You are facing a potted plant. Turn around, exit the room. Turn right, walk straight. Turn right, you can see stairs. Get down through the stairs. Turn right, get down through the stairs. Stand on the second step of the stair from the top. This is the end point.

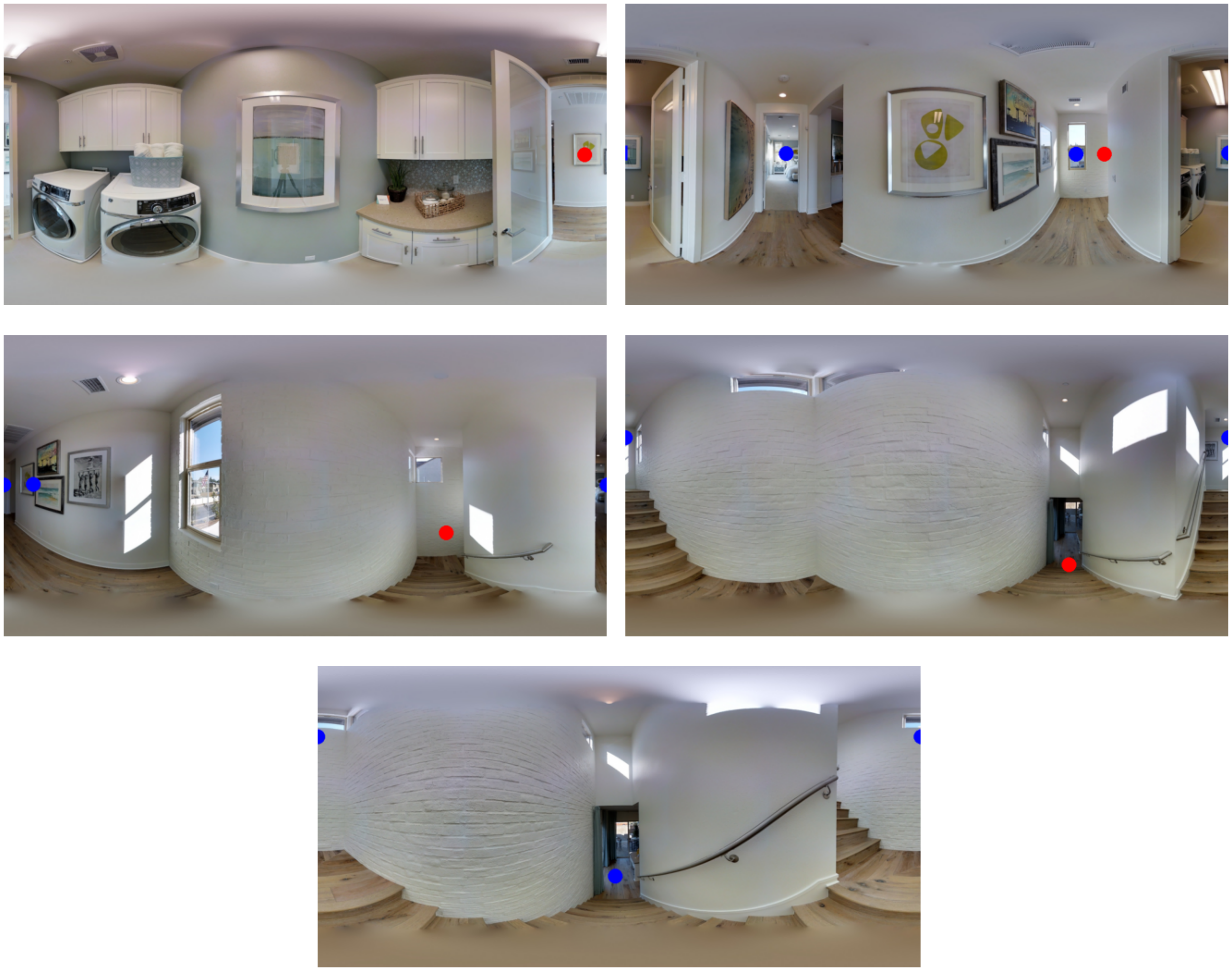

Figure 5. Example Marky (synthetic) instruction for a sampled trajectory. The images are 360° panoramas rotated so that the direction faced by the agent is the in center. Blue dots indicate directions the agent can move in the underlying navigation graph. The correct action at each step is colored in red. Note that the potted plant mentioned in the instruction is on the countertop.

Marky Instruction: You begin facing a living space, take a step forward and then behind the brown chair that's in front of you, there is a archway, go through that archway, and then go down the hallway to the left, turn into the first room on the right, it is a bathroom, take a step inside and you're done.

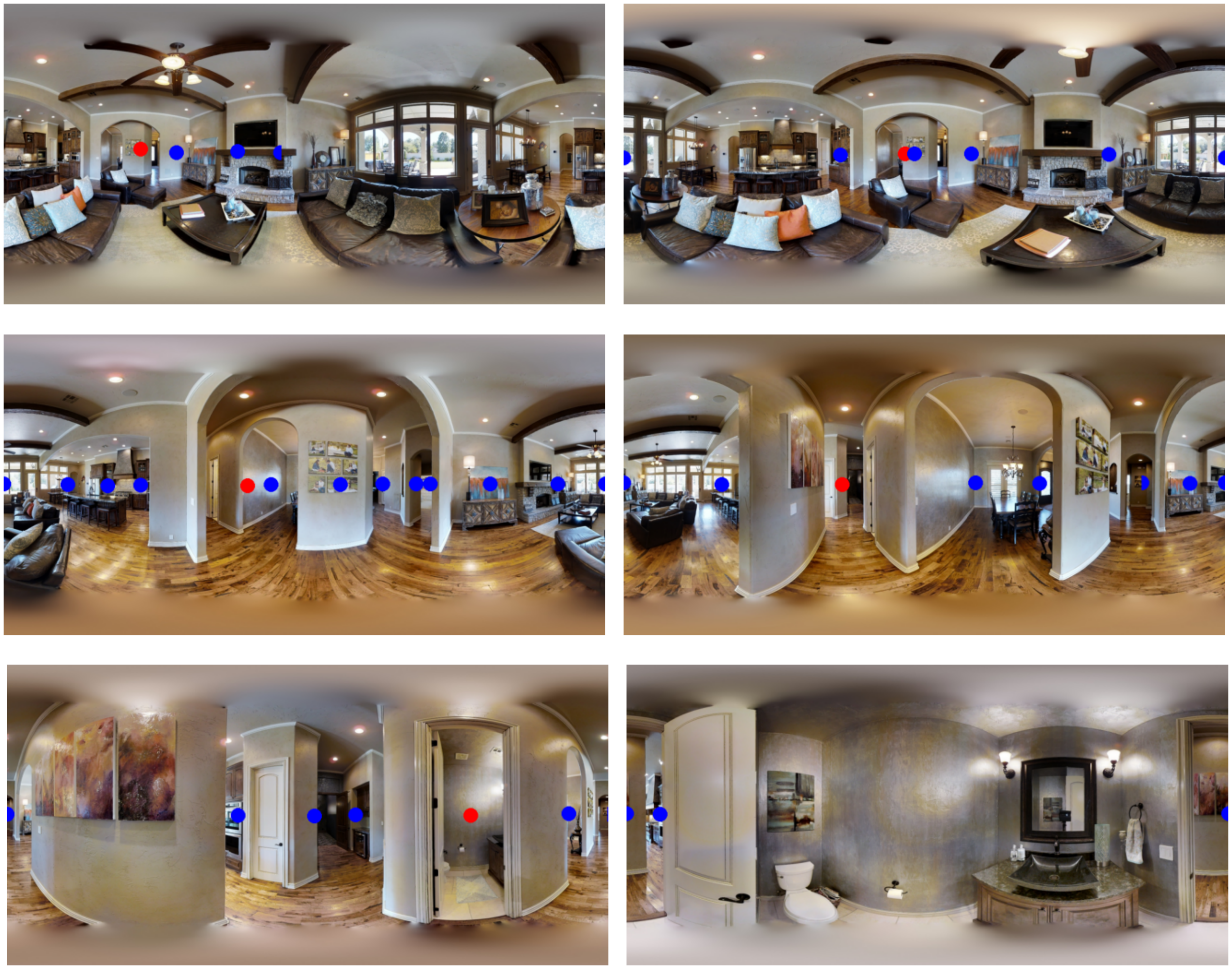

Figure 6. Example Marky (synthetic) instruction for a sampled trajectory. The images are 360° panoramas rotated so that the direction faced by the agent is the in center. Blue dots indicate directions the agent can move in the underlying navigation graph. The correct action at each step is colored in red.

Instruction: You're facing towards a wooden shelf, turn slight left and move forward you can see a gas stove in front of you walk near the gas stove you can see a open door in front of you, turn right and exit the room by moving one step forward, turn slight left and you can see a stair case in front of you walk near the stair case, turn slight right and you can see an open door in front of you walk near the open door, turn slight right and you can see a washing machine in front of you walk near the washing machine and stand near the washing machine and this will be your end point.

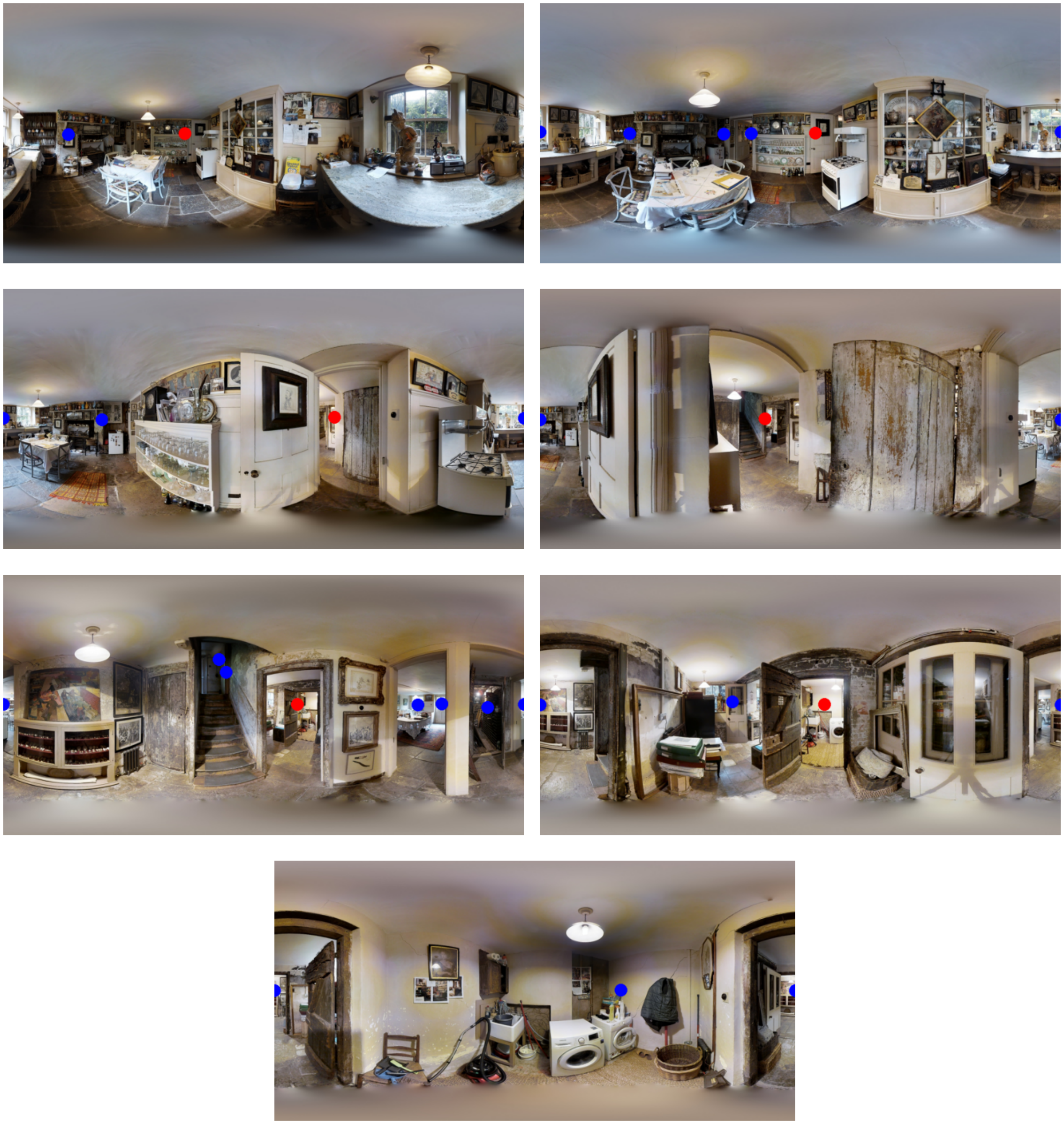

Figure 7. Inference example of MARVAL *successfully* following an English instruction from the RxR Val-Unseen split through a sequence of panos. The panos are rotated so that the direction faced by the agent is the in center. Blue dots indicate action candidates that the agent could move to. The selected action at each step is colored in red.

Instruction: You're starting in the corner of a living room. Turn around behind you and find the clock hanging on the wall in the hallway. Take two steps toward that. Turn right and go straight between the blue couch on your right and the kitchen counter on your left. You'll take three steps and stop at the first, right corner of the long dining table across from the kitchen. You should be looking through windows into the backyard in front of you. To the right, is an open patio, and to your left should be four framed, black and white pictures. You're done.

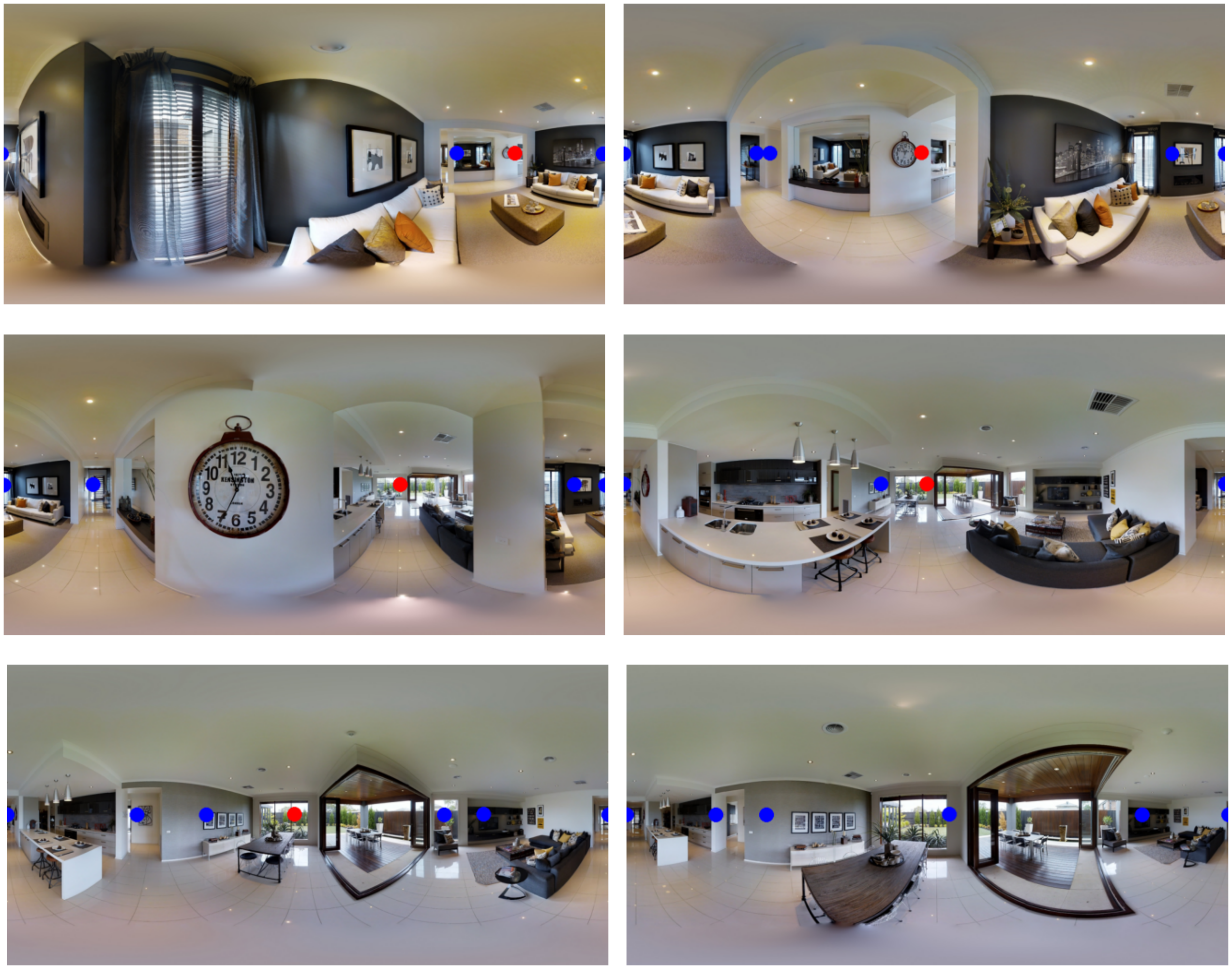

Figure 8. Inference example of MARVAL *successfully* following an English instruction from the RxR Val-Unseen split through a sequence of panos. The panos are rotated so that the direction faced by the agent is the in center. Blue dots indicate action candidates that the agent could move to. The selected action at each step is colored in red.

Instruction: You are in the living room near a sofa chair and facing the open arch which is behind the sofa chair, move towards that arch, slightly turn right and walk down the narrow walkway, now move towards the closed wooden glass door which is in the front, turn left side and move towards the single chair which is on the right side, walk little forward from that single chair, now turn left side, and walk forward towards the small wooden cupboard in the front, slightly turn right side and enter the room, stand on the edge of the table which has book and a pen, that is your destination.

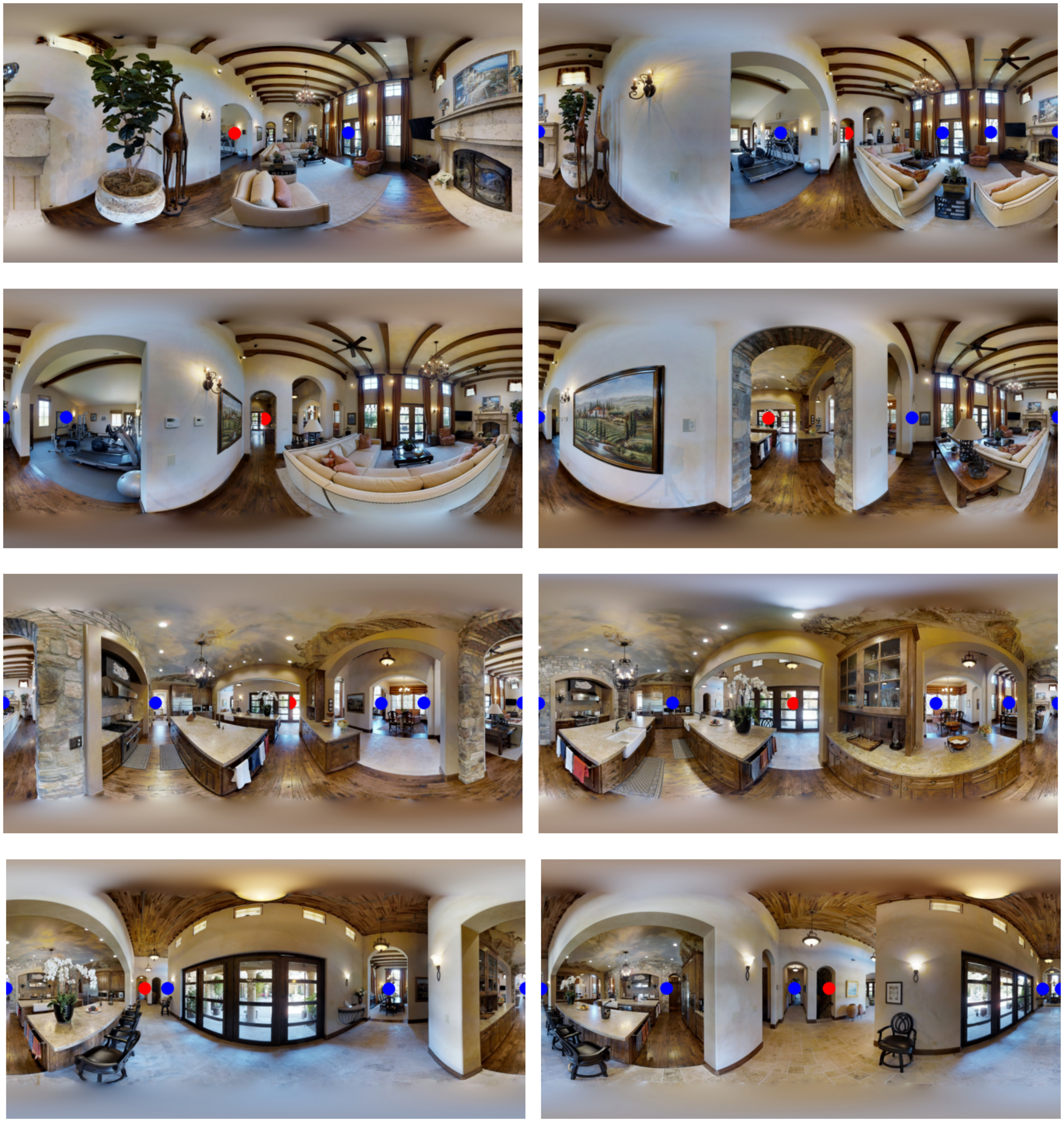

Figure 9. Inference example of MARVAL *failing* to follow an English instruction from the RxR Val-Unseen split. The panos are rotated so that the direction faced by the agent is the in center. Blue dots indicate action candidates that the agent could move to. The selected action at each step is colored in red. Here, MARVAL follows around 75% of the instruction correctly, up until the instruction mentions the 'single chair' (seen in the pano bottom right). However, at this point MARVAL makes an error and does not recover, failing to find the 'table which has book and pen'.