SUPPLEMENTARY MATERIAL

Less is More: Generating Grounded Navigation Instructions from Landmarks

1. Bootstrapping a Landmark Dataset

Extracting landmark phrases. As outlined in Sec. 3, the first step to create silver landmark data is extracting temporally-ordered lists of landmark phrases from RxR's human instructions. For this, we use a 3-layer distilled [7] mBERT-based [2] dependency parser pretrained on multilingual Wikipedia data [3] and outputting Universal Dependencies [5]. Specifically, the steps are:

1. Extracting all entity mentions (along with their part-of-speech) from an instruction;
2. Centering around the mentions as the head N/NP, absorbing all of their non-clausal dependents. E.g. for ... *at the large brown* ***chair*** *that sits next to the window* ..., the mention is *chair*, and the absorbed dependents are *large* and *brown* (but not the dependent clause *that sits* . . .);
3. Consolidating the head N/NP and its absorbed dependents into a single text span and recording the indices of the first and last characters. In the example above, we produce *large brown chair*.

While the method identifies most desired landmark text spans, inevitable errors do occur, sometimes due to the imperfection of the human-written instructions, e.g.

- Case 1. Non-object NPs. For *". . . take two steps towards the door"*, for instance, *steps* will be extracted as a candidate. Similarly, *left* in *". . . take a left"*.
- Case 2. Ill-formed/incomplete sentences. For example, in Fig. 3, the single-word sentence *"Fancy."* results in *fancy* being parsed as a landmark text span.

To address Case 1, we manually compiled a *stoplist* (Fig. 1) of common non-object NPs that may appear in indoor environment (e.g. [*side, step, one, endpoint*, . . .]) for each language (English, Hindi, Telugu). Case 2 is unavoidable with the tools currently available to us. Fortunately, they occur quite rarely.

Fig. 2 demonstrates landmark phrase extraction on an example instruction.

Evaluation. In Sec. 3 of the paper, we include results from a small-scale evaluation of 100 randomly-sampled English instructions from the silver landmark dataset. To provide ground truth landmark groundings, each automatically-extracted landmark phrase was manually aligned by the paper authors to a subsequence of frames from the corresponding pose trace video. The interface used for this manual-alignment is illustrated in Fig. 3. To compute the precision scores in the paper, whenever our automatic approach grounds a landmark phrase to a frame, it is a *true positive* if the frame is in the human-selected subsequence, and a *false positive* otherwise. Results are averaged over all landmark phrases in an instruction and then all instructions.

	BLEU	CIDEr
2-class	**5.3**	**7.7**
10-class	4.8	7.2
100-class	4.4	6.4

Table 1. Impact of training the landmark detector on clustered landmark classes. Automatic evaluation scores are reported for Marky-mT5 without pretraining, using the RxR Val-Unseen split.

Further analysis of the silver data. Fig. 4 provides further examples of landmark annotations from our bootstrapped silver landmark dataset. Fig. 5 presents the distribution of landmark phrases, including the 20 most frequent and least frequent landmark phrases. The distribution naturally has a long tail, although many unique landmark phrases are semantically similar (e.g. *brown chair* vs. *white chair*). We also empirically confirm the intuition that people tend to pick landmarks close to the *outbound direction* (i.e. the direction towards the next route segment). Fig. 6 illustrates the distribution of landmark centers in equirectangular image coordinates aligned to the outbound direction of each pano. On the horizontal dimension (heading), the landmarks are clustered close to the outbound direction, whereas on the vertical dimension (pitch), landmarks are clustered on the horizon and slightly below.

2. Landmark Detection

Training. As noted in Sec. 4, the CenterNet landmark detector was trained using a single class to represent a landmark (i.e., 2 classes in total). In initial experiments, we also investigated training the detector using a richer set of landmark classes, by clustering landmark phrases (using k-means) into classes based on their MURAL-large [4] text embeddings. However, this slightly reduced automatic evaluation scores for the full model (Tab. 1), perhaps because detection confidence scores are uncalibrated across classes.

Inference. During inference with the landmark detector, we must determine how many landmarks to return. As described in Sec. 4, we used a 1:1 ratio of landmarks to path length, defined as the number of panos in the path. To determine this ratio, we experimented with the following values $\{1.0, 1.2, 1.5, 1.75, 2.0\}$ and computed automatic evaluations on the full model as reported in Tab. 2.

left	corner
right	way
front	top
back	few steps
start	two steps
end	place
point	bit
end point	bottom
endpoint	edge
one	area
side	set
other	center
room	2 steps
middle	sides
turn	left corner
destination	start point
final destination	three steps
goal	color
wall	stop
step	space
steps	left move
one step	1 step

बाएं	कदम
बाए	एक कदम
दाहिने	कोने
दाए	मार्ग
दाएं	ऊपर
सामने	कुछ कदम
वापस	दो कदम
शुरू	स्थान
समाप्त	अंश
बिंदु	नीचे
अंतिम बिंदु	किनारा
एक	क्षेत्र
पक्ष	समूह
अन्य	केंद्र
कक्ष	पक्ष
मध्य	पक्षों
मोड़	बायाँ कोना
गंतव्य	प्रारंभ बिंदु
अंतिम गंतव्य	तीन कदम
लक्ष्य	रंग
तरफ	विराम
दिशा	स्थान
मुड़िये	बायीं चाल
बाज़ू	1 कदम
हल्का	2 कदम
उदर	
दीवार	

ఎడమ	వెనకవైపున	ఒక్క అడుగు
ఎడమవైపు	వెనకవైపుకు	మూలలో
ఎడమవైపున	వెనకవైపు	మార్గం
ఎడమవైపుకు	వెనకవైపున	టాప్
కుడి	వెనకవైపుకు	కొన్ని దశలు
కుడివైపు	ప్రారంభం	రెండు దశలు
కుడివైపున	ముగింపు	స్థలం
కుడివైపుకు	పాయింట్	బిట్
ముందర	ముగింపు స్థానం	దిగువన
ముందరవైపు	ఒకటి	అంచు
ముందరపున	వైపు	ప్రాంతం
ముందరపుకు	వైపున	సెట్
ముందు	వైపుకు	కేంద్రం
ముందువైపు	ఇతర	2 దశలు
ముందుపున	గది	వైపులా
ముందుపుకు	మధ్య	ఎడమ మూలలో
తిరిగి	మలుపు	ప్రారంభ స్థానం
వెనుక	గమ్యం	మూడు దశలు
వెనక	చివరి గమ్యం	రంగు
వెనెక	లక్ష్యం	ఆపు
వెనుకవైపు	గోడ	స్థలం
వెనుకవైపున	అడుగు	ఎడమ కదలిక
వెనుకవైపుకు	దశలు	1 అడుగు
వెనకవైపు		

Figure 1. Stoplists for excluding hard-to-localize noun phrases (NPs) when extracting landmark phrases. Left: English; Center: Hindi; Right: Telugu. Each stoplist was compiled by a native speaker examining samples of parsed landmark phrases.

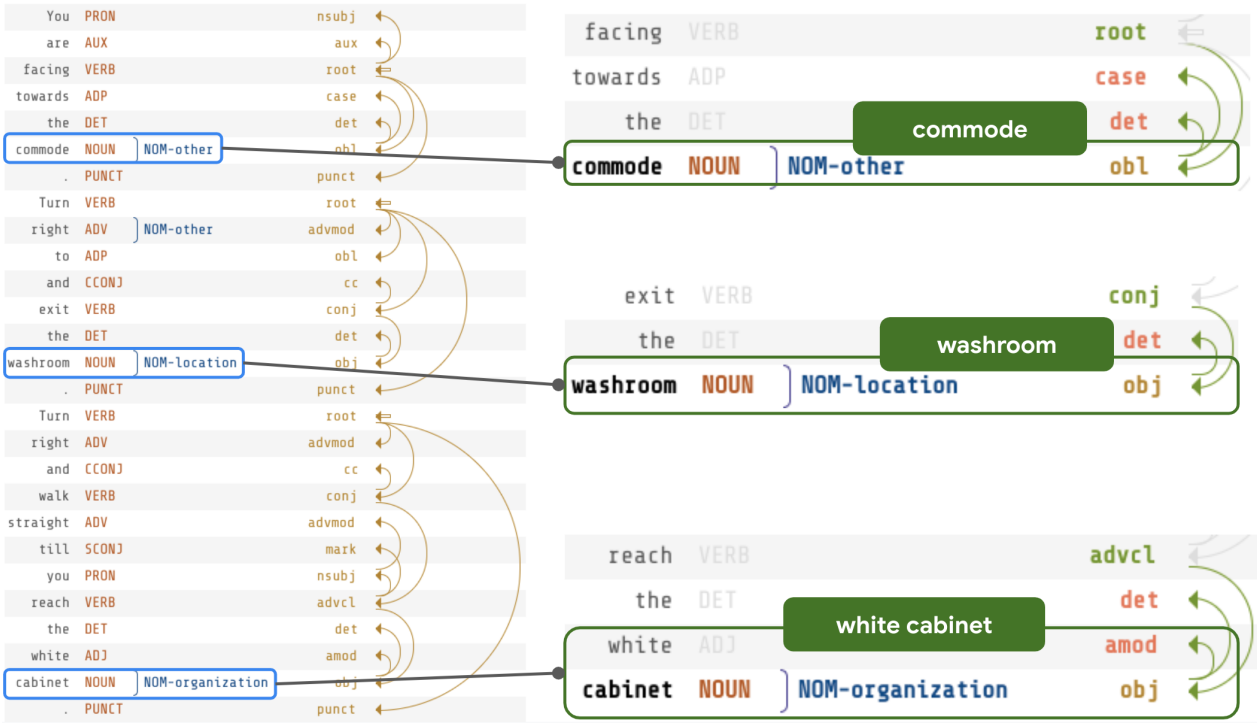

Figure 2. Landmark phrase extraction for the instruction segment "*You are facing towards the commode. Turn right to and exit the washroom. Turn right and walk straight till you reach the white cabinet.*". First the object-denoting heads are identified, then we filter through each one's dependency links to find full NP phrases for each head with constraints (e.g. only including adjectival, numeral, or adverbial dependencies, etc.).

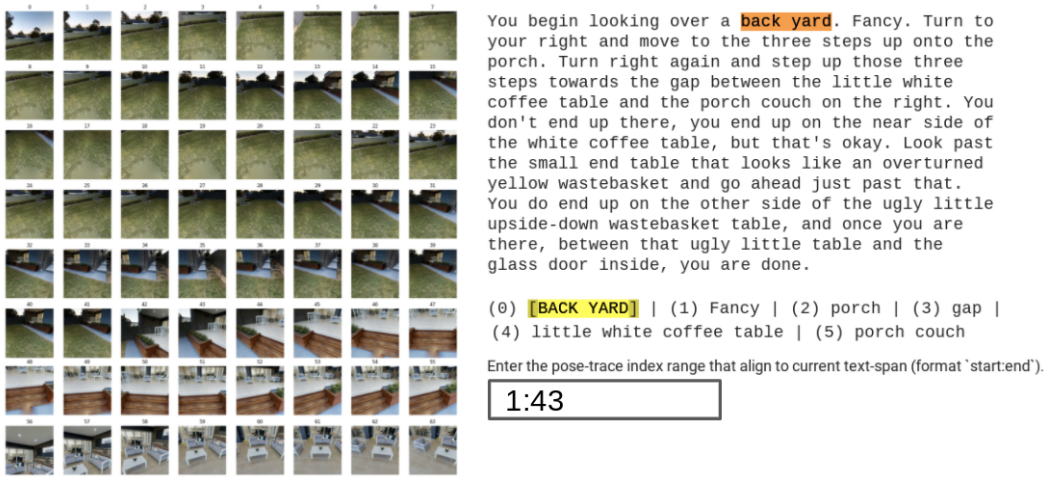

Figure 3. Annotation interface used to collect ground-truth landmark groundings for a small-scale evaluation of the silver landmark dataset. For each automatically extracted landmark phrase (e.g., *back yard*, right), the annotator inputs a range identifying the frames in the pose trace video (left) where that landmark can be seen.

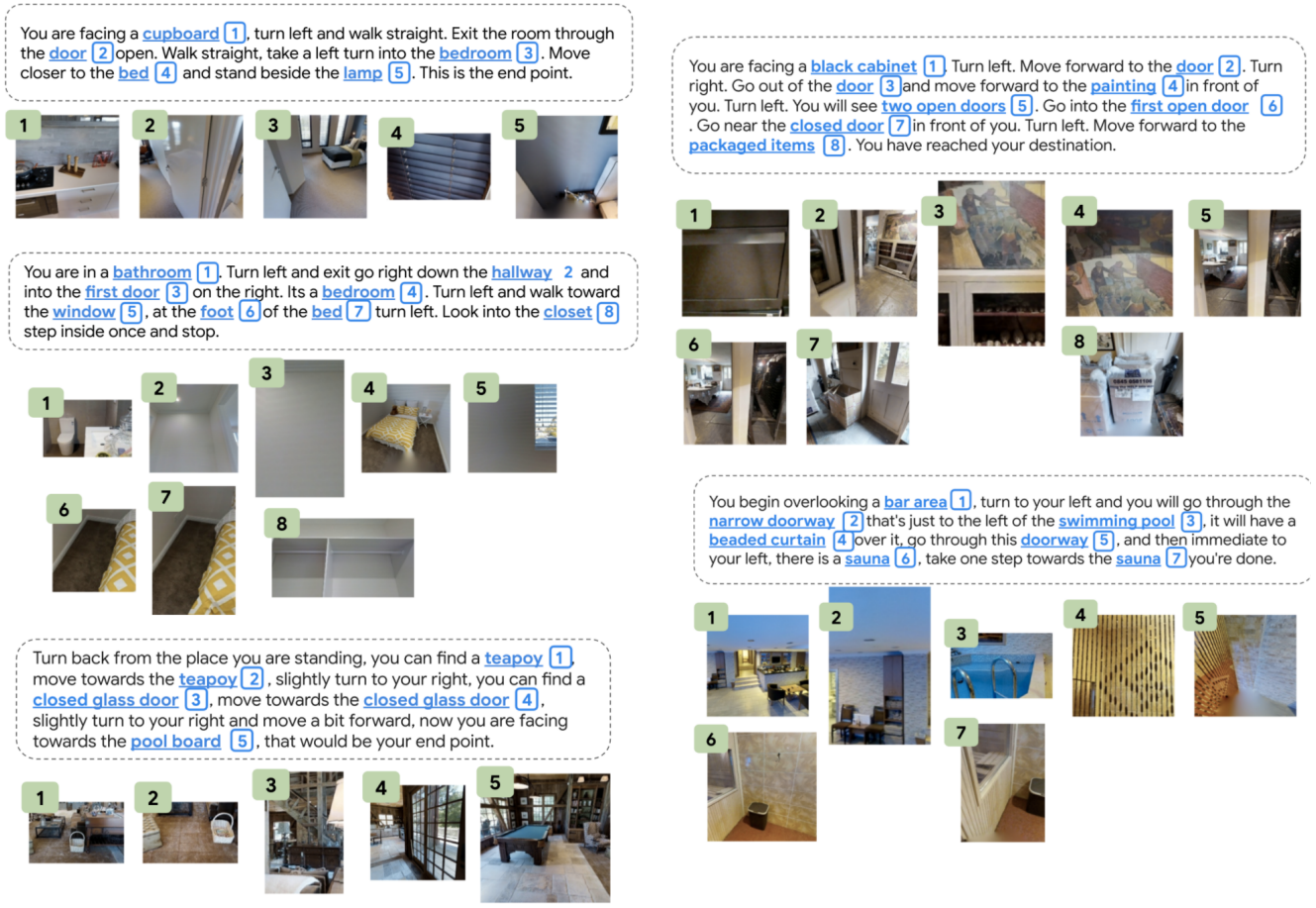

Figure 4. Further examples of bootstrapped silver landmark annotations, illustrating both correct groundings (e.g., *packaged items*, top right [8] and *beaded curtain*, bottom right [4]) as well as failure cases (e.g., *lamp*, top left [5]).

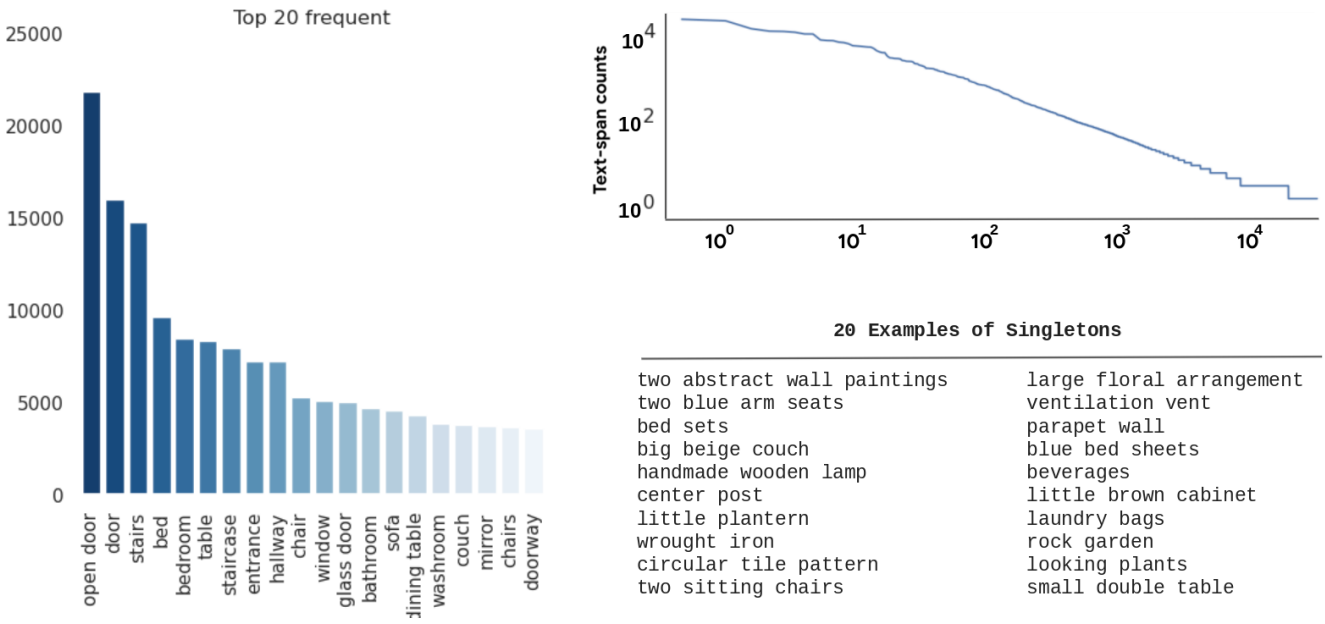

Figure 5. **Left**: Frequency distribution of the top 20 landmark phrases. **Right top**: Long-tail distribution over all landmark phrases sorted in descending order (both x- and y-axes are labeled in log10-scale: landmark indices for x-axis, instance counts for y-axis). **Right bottom**: 20 samples of landmark phrases occurring only once. Common indoor/household items described generically (e.g. *table* or *door*) come on top of the rank over more specified ones (e.g. *large brown dining table* or *closed double door*). On the bottom end of frequency are specifically described uncommon objects, e.g. *handmade wooden lamp*, *parapet wall*, *first two arched doorways*, etc.

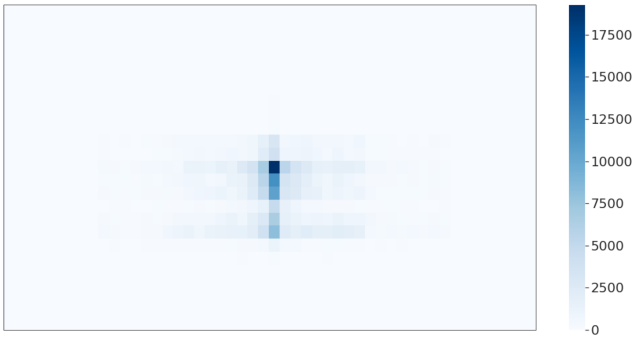

Figure 6. Distribution of landmark centers in equirectangular image coordinates aligned to the outbound direction (i.e., the direction to the next pano on the route). As expected, landmarks are clustered around the outbound heading direction. With regards to pitch, most landmarks are found on the horizon or slightly below. The reason is fairly simple in the context of indoor environments: most landmarks are found on the floor or at table height, relatively few are found on walls and ceilings.

3. Instruction Generation

Pretraining and Finetuning. During pretraining, models are trained with Cross Entropy Loss and optimized with Adafactor [6] with a learning rate of 1. Batch size is 128. Pretrained models were trained for 1.45M steps. Each pretrained model (CC3M, CC12M, CC3M+CC12M) was finetuned and the final pretrained version of the downstream model was selected based on SPICE.

#landmarks / path length ratio	BLEU	CIDEr
1.0	**5.8**	**7.5**
1.2	5.8	7.2
1.5	5.3	6.1
1.75	4.9	5.0
2.0	4.4	3.7

Table 2. Impact of varying the #landmarks / path length ratio. Automatic evaluation scores are reported for Marky-mT5 with Rewrite auxiliary training and CC3M/12M pretraining.

4. Experiments

Human Wayfinding In the PanGEA interface, each annotator is shown a virtual environment in a window on the left, paired with the textual navigation instruction being evaluated on the right (refer top pane in Fig. 7). Hovering their mouse on the window, the annotator will see a green-square indicator showing them the next locations available for them to move to. After double clicking on the green-square, they are taken to the next location — illustrated in Fig. 7 bottom pane, which presents a chain of first person snapshots taken while moving. Upon arriving at the loca-

		Model	Landmarks	WC	NE ↓	SR ↑	SDTW ↑	NDTW ↑	Quality ↑	Visual Search % Start ↓	Visual Search % Other ↓	Time (s) ↓
RxR (en)	1	Marky-mT5	Outbound	75.0	5.2	52.8	40.0	57.0	4.2	36.7	25.4	101.8
	2	Marky-mT5	Predicted	81.6	4.2	61.3	46.5	61.2	4.3	36.1	24.6	107.6
	3	Marky-mT5	Silver	91.2	4.0	65.0	49.0	60.8	4.3	36.3	25.2	118.4
	4	Human	-	98.6	2.7	77.5	62.2	71.0	4.6	35.3	24.3	113.5
RxR (hi)	1	Marky-mT5	Outbound	82.1	5.6	50.8	32.6	46.4	4.2	42.5	27.4	194.8
	2	Marky-mT5	Predicted	78.5	4.7	59.4	40.9	52.5	4.2	38.4	27.6	192.4
	3	Marky-mT5	Silver	72.4	4.6	60.7	42.3	54.4	4.3	39.1	27.1	176.0
	4	Human	-	75.0	2.9	77.1	59.4	67.8	4.6	37.0	26.3	171.5
RxR (te)	1	Marky-mT5	Outbound	43.2	5.2	56.3	39.6	52.9	4.1	37.5	27.9	143.8
	2	Marky-mT5	Predicted	52.1	4.3	63.9	44.6	55.1	4.1	38.3	27.2	162.5
	3	Marky-mT5	Silver	51.9	4.4	64.2	45.0	55.9	4.1	38.1	27.5	154.5
	4	Human	-	53.7	2.6	80.9	62.6	68.9	4.4	37.1	26.5	157.4

Table 3. RxR Val-Unseen human wayfinding performance (**N** = 1,517 for each model), reported separately by language (English, Hindi and Telugu).

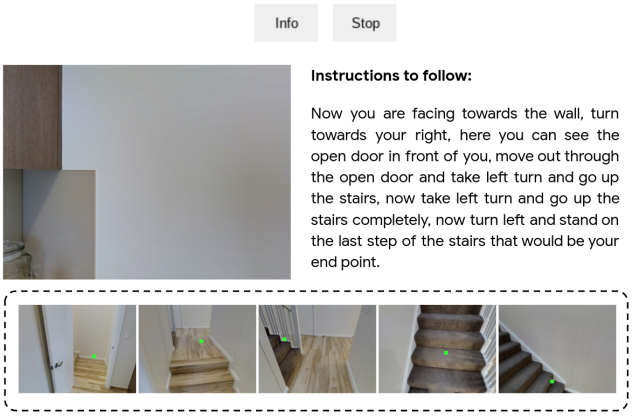

Figure 7. Top: the PanGEA interface used in human wayfinding evaluations; Bottom: illustration of a series of first person snapshots taken along a navigation path.

tion the annotator believes is the destination, they may hit the `STOP` button to finish the task.

Afterwards, the annotator responds to a multi-choice question to provide a subjective rating of instruction quality, classifying the instruction as containing:

- *No mistakes, very very easy to follow*;
- *Few mistakes, easy to follow*;
- *Some mistakes, but still not hard to follow*;
- *Many mistakes, hard to follow*;
- *Way too many mistakes to follow*.

From the top to the bottom, the answer determines the *Quality* metric (a Likert score from 5 to 1). All annotators were fluent in the languages in the instructions given to them for wayfinding. The annotators were paid hourly wages that are competitive for their locale, and they have standard rights as contractors.

Results In Tab. 2 of the main paper we report human wayfinding results on paths from the RxR Val-Unseen split, aggregated overall all languages. In Tab. 3 we report these results separately for each language (English, Hindi and Telugu). Results – and most importantly, the patterns holding between the different settings – are consistent across all languages.

Finally note that, for both R2R and RxR evaluation (Val-Unseen), each path is only evaluated once per language. Therefore, for R2R we have 783 items/paths, for RxR, we have 1,517 paths, and with 3 languages, 4,551 items.

Error analysis on Marky-mT5 instructions. Our human wayfinding results are representative of a step-change in instruction quality compared to previous models. To better understand the types of errors that still remain, we perform a manual error analysis on 110 randomly-sampled English instructions generated for RxR Val-Unseen paths by the full Marky-mT5 system using predicted landmarks.

To perform the analysis we added an option in the PanGEA wayfinding interface to toggle the visibility of the ground-truth path, so that we could better assess how well that path was described by the generated instruction. We classify instruction errors and weaknesses into the following six categories:

- **Landmark Errors**
  - *Full Hallucination*. The instruction describes a landmark that does not exist in the environment;
  - *Weak Description*. A landmark description that is flawed, but not completely wrong (e.g. *blue towel* described as *blue napkin*);
  - *Wrong Orientation*. An instruction refers to a landmark but orients it incorrectly with respect to the route, e.g. saying it is *on the left side* when

	% of all errors
Landmark Errors:	
Full Hallucination	11.4
Weak Description	**62.9**
Wrong Orientation	7.9
Path Errors:	
Wrong Action	7.1
Missing Action	3.6
Weak Granularity	7.1

Table 4. Of generated instructions with errors, 25.7% of errors are issues with actions or landmark orientation (wrong, missing, or convoluted [i.e., weak granularity]). Another 11.4% of errors are full hallucinations but the overwhelming majority, 62.9% are an issue with some aspect of the description of a landmark.

it is actually *on the right side*.

- **Path Errors**
  - *Wrong action*. A mistake in an action/step (e.g. *turn right* when one needs to *turn left*);
  - *Missing action*. Skipping a an action/step (e.g. the instruction neglects to mention a crucial right hand turn);
  - *Weak granularity*. Some segment of the instruction is too coarse to describe the multiple steps that are needed in the path trajectory (e.g. merely using *go forward* to describe a route that passes through two doorways and a dining hall).

We provide examples of each error type in Fig. 8 and Fig. 9, where the blue/purple (arrowed) balls indicate the ground truth path. The error analysis was performed by the paper authors with a critical eye; any weakness in the instructions was annotated as an error.

The results are summarized in Tab. 4. Of the instructions annotated, 16% were judged to be error-free, with some errors or weaknesses identified in the remaining 84%. Out of the 6 error types, *Weak Description* was by far the most common (62.9% of all errors). In contrast, the proportion of *Full Hallucination*, *Wrong Orientation* and *Wrong Action* errors–which were common in previous models–was relatively low, at 11.4%, 7.9% and 7.1% respectively.

Anecdotally, we noticed that the cost of different errors varies. Human wayfinders can often overcome minor flaws in the description of a landmark (e.g. if a ***pink*** *bedspread* is misidentified as a ***white*** *bedspread*). However if ambiguity is involved, e.g. *move towards* ***the open door*** when there more than one door is in view, confusion results. Full hallucination and wrong orientation can certainly be misleading but if the navigator survey the environment carefully in the context of the neighboring segments in the instruction, they are also often resolvable. The three types of path errors are less recoverable, as they often result in missteps that take the wayfinder to an entirely wrong path.

	Model	Landmarks	Aux	PT	SR ↑	SPL ↑	NDTW ↑	SDTW ↑
1	SpkFol-RxR	Full Panos			29.6	25.9	41.6	23.4
2	MARKY-MT5	Full Panos			50.7	46.9	60.1	43.1
3	MARKY-MT5	Outbound			53.6	50.1	62.9	46.7
4	MARKY-MT5	Silver			55.9	52.1	64.1	48.6
5	MARKY-MT5	Silver	✓		56.3	52.3	64.2	48.9
6	MARKY-MT5	Silver	✓	✓	56.4	52.5	64.2	48.9
7	MARKY-MT5	Pred.	✓	✓	55.7	51.8	63.3	47.7
8	Human				56.5	52.7	62.9	48.4

Table 5. Automatic evaluations of generated instructions on RxR Val-Unseen based on HAMT [1] wayfinding performance. Settings and row numbers correspond to Tab. 3 in the main paper.

Results on automatic evaluation. Beyond the strong quality of MARKY-MT5-generated instructions in human evaluation (**????**), they also perform at the similar level in automatic/model evaluation. Employing the state-of-the-art HAMT [1] VLN agent (Tab. 5), in particular, we received 55.7% vs. 56.5% success rate and 63.3% vs. 62.9% nDTW between model-generated (with predicted landmarks) vs. human-written instructions, demonstrating that MARKY-MT5 produces instructions followable by both human and model agents.

References

[1] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, 2021. 6

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, June 2019. 1

[3] Mandy Guo, Zihang Dai, Denny Vrandecic, and Rami Al-Rfou. Wiki-40b: Multilingual language model dataset. In *LREC 2020*, 2020. 1

[4] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 1

[5] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal Dependency Parsing from Scratch. In *Proceedings of CoNLL*, 2018. 1

[6] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 4

[7] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 1

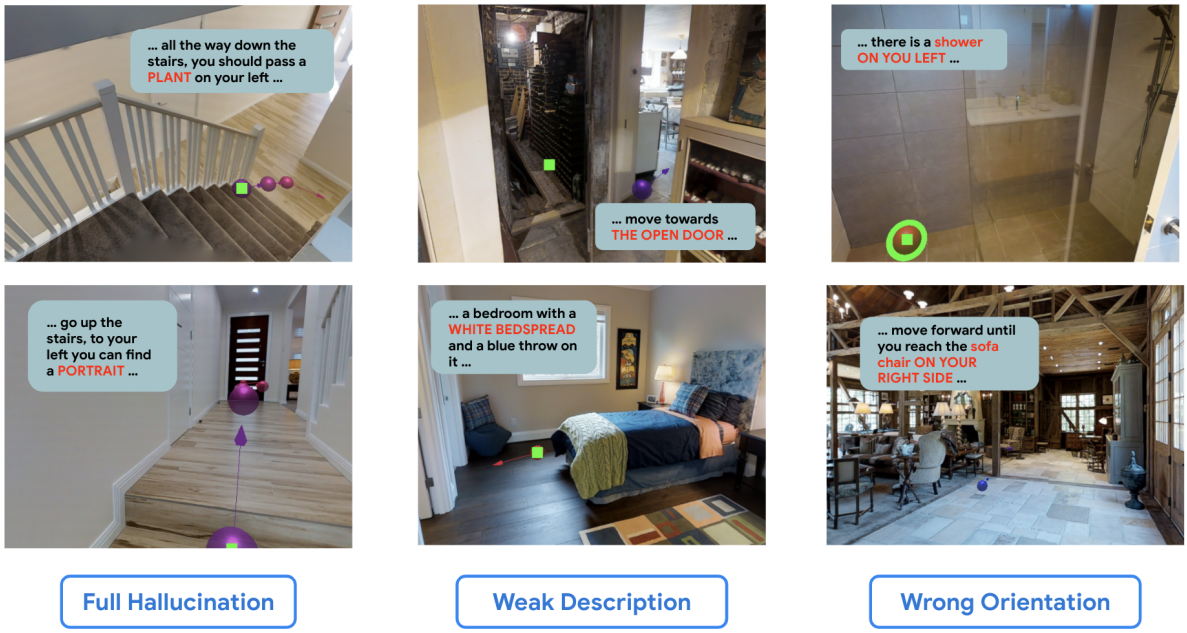

Figure 8. Examples of *Landmark Errors*. **Full Hallucination**: a *plant* and *portrait* are mentioned in the top and bottom panes respectively but these landmarks cannot be found in the visual scene. **Weak Description**: the top pane exemplifies an ambiguous landmark – there are two open doors, and it's not clear which one to move towards; the bottom pane illustrates a flawed description (specifically, wrong color). **Wrong orientation**: On the top pane, the *shower* is to the *right hand side* of the wayfinder rather than left; in the bottom pane, the *sofa chair* should be on the *left side* instead.

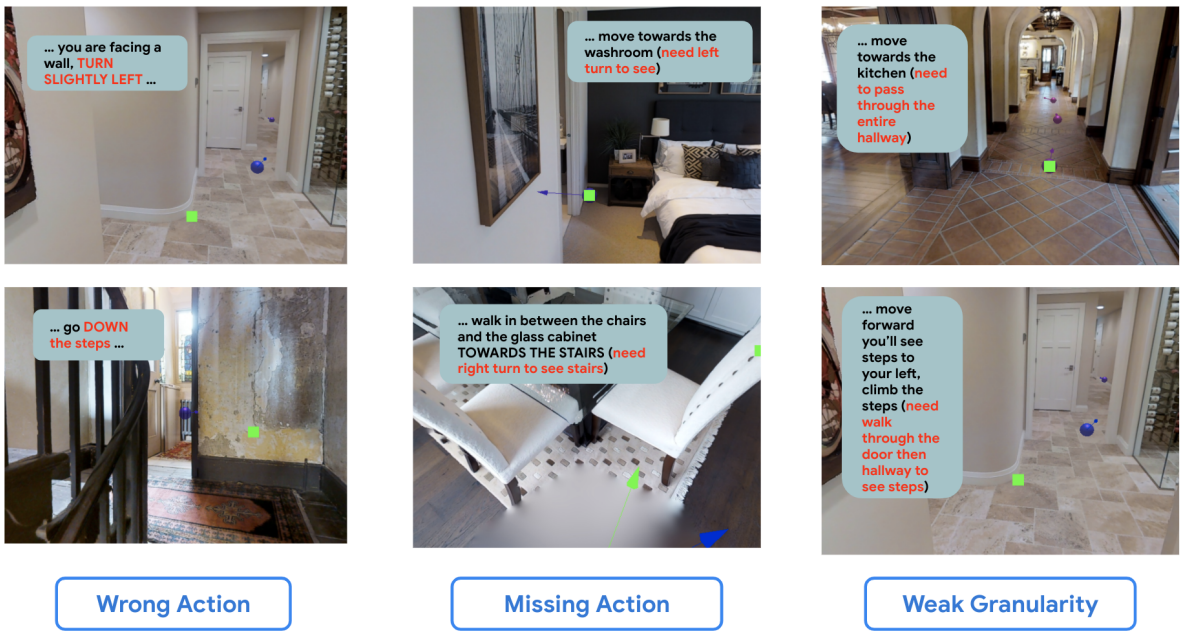

Figure 9. Examples of *Path Errors*. **Wrong action**: the examples show a right turn mistaken to be a left turn, and going up stairs mistaken as going down. **Missing action**: for the top pane, the *washroom* is not visible before making an additional left turn; for the bottom pane, *stairs* require a right turn to see. **Weak granularity**: in the examples, overly coarse instruction segments are given for long path segments.