Less is More: Generating Grounded Navigation Instructions from Landmarks

Su Wang **Ceslee Montgomery** **Jordi Orbay** **Vighnesh Birodkar** **Aleksandra Faust**
Izzeddin Gur **Natasha Jaques** **Austin Waters** **Jason Baldridge** **Peter Anderson**

Google Research

Abstract

We study the automatic generation of navigation instructions from 360° images captured on indoor routes. Existing generators suffer from poor visual grounding, causing them to rely on language priors and hallucinate objects. Our MARKY-MT5 system addresses this by focusing on visual landmarks; it comprises a first stage landmark detector and a second stage generator–a multimodal, multilingual, multitask encoder-decoder. To train it, we bootstrap grounded landmark annotations on top of the Room-across-Room (RxR) dataset. Using text parsers, weak supervision from RxR's pose traces, and a multilingual image-text encoder trained on 1.8b images, we identify 971k English, Hindi, and Telugu landmark descriptions and ground them to specific regions in panoramas. On Room-to-Room, human wayfinders obtain success rates (SR) of 71% following MARKY-MT5's instructions, just shy of their 75% SR following human instructions—and well above SRs with other generators. Evaluations on RxR's longer, diverse paths obtain 61-64% SRs on three languages. Generating such high-quality navigation instructions in novel environments is a step towards conversational navigation tools and could facilitate larger-scale training of instruction-following agents. The full paper at CVPR 2022 is available at `https://arxiv.org/abs/2111.12872`.

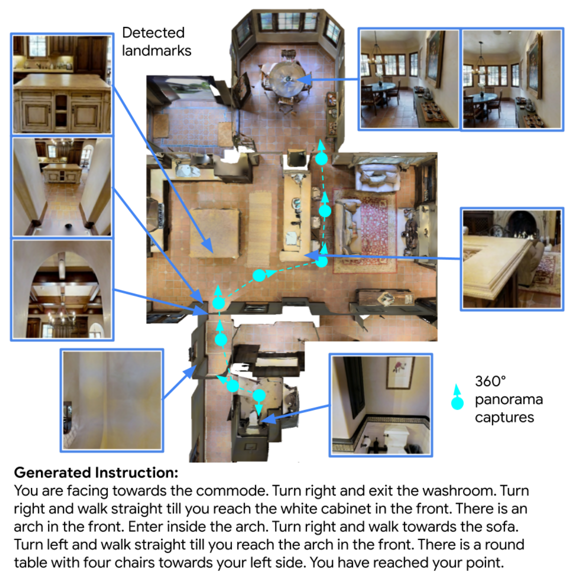

Figure 1. We generate grounded navigation instructions from a sequence of 360° images captured along a route in a previously unseen building. Our two-stage approach first detects landmarks and then generates instructions conditioned on these landmarks.

Introduction

First of all, we make progress towards the desired capability of ***generating instructions directly from visual input***. This allows for much stronger generalizability: Instruction generators for indoor wayfinding assume the access to preexisting floorplans and landmark databases [11], but recent work attempts to generate novel instructions directly from visual inputs [6, 10, 13]. Progress toward this goal will enable navigation aids that are conversational rather than map-based—and it could provide a virtually unlimited supply of high-quality synthetic navigation instructions for training instruction-following robots. Describing navigation paths is also a key capability for human-robot communication, equipping robots to answer questions such as *where did you go?* or *where should I meet you?*.

We seek to generate accurate and fluent navigation instructions–in multiple languages–directly from visual representations and actions taken to traverse a path. Previous work assumed that the input to the instruction generator is a sequence of 360° panoramic (henceforth, *pano*) images captured at intervals on a path, typically training on instructions from Room-to-Room (R2R) [1] using Matterport3D environments [2]. These models' instructions have proven valuable as additional training data for vision-and-language navigation (VLN) agents [6]. However, *people* struggle to follow them [16]: human wayfinding success rates on R2R are 36% for Speaker-Follower [6] and 42% for EnvDrop [13] in unseen environments. The generated text is stylistically correct, but frequently references non-existent objects and confuses spatial terms such as left/right.

Secondly, we ***identify and filter for relevant visual grounding from visual inputs***. A challenge for visually-

		Model	Landmarks	Training Data	WC	NE ↓	SR ↑	SPL ↑	Quality ↑	Visual Search % Start ↓	Visual Search % Other ↓	Time (s) ↓
R2R (en)	1	SpkFol [6]	Full Panos	R2R	24.6	6.0	42.0	35.8	4.1	39.8	23.6	54.2
	2	EnvDrop [13]	Full Panos	R2R	24.5	6.0	41.7	35.3	4.0	40.7	23.5	54.0
	3	SpkFol-RxR [6]	Full Panos	RxR	61.8	3.9	57.8	48.7	4.2	36.0	23.7	67.5
	4	Marky-mT5	Outbound	RxR	57.5	3.6	64.9	54.1	4.2	36.2	23.8	72.5
	5	Marky-mT5	Predicted	RxR	58.2	2.9	70.8	59.8	4.3	35.5	23.2	70.1
	6	Human	-	-	25.6	2.8	74.9	66.4	4.5	37.8	23.0	52.2

Table 1. R2R Val-Unseen human wayfinding performance (**N** = 783 for each model). Combining the larger RxR dataset with landmark modeling and our bootstrapped landmark dataset, we almost eliminate the gap between model-generated and human-written instructions on paths of R2R-level difficulty – achieving a 70.8% success rate vs. 74.9% for human instructions and 42% for previous models. (outbound: outbound: defined as the view from the current pano in the direction of the next pano)

oriented instruction generators is dealing with irrelevant visual inputs. In many other image-to-text generation tasks (e.g., image captioning), much of the visual information in the input is reflected in the output text. This is not the case when generating navigation instructions. Human annotators look at less than 30% of the environment [9], and the instructions reference only a fraction of the objects that they look at. This makes learning a precise mapping between visual inputs and text outputs much harder. Perversely, access to more information can degrade performance [5], as models happily learn spurious correlations that cause hallucinations during inference. To solve this, we exploit the spatiotemporal grounding in the Room-across-Room (RxR) dataset [9]. Instead of *writing* instructions, RxR annotators *spoke* while traversing paths. Every RxR instruction thus comes with *pose traces* that align the words spoken (and later transcribed) with what annotators were looking at. We use these pose traces and instructions to derive a new *silver* annotated dataset[1] that contains bounding boxes over visual landmarks combined with their multilingual descriptions (English, Hindi, and Telugu). Specifically, we bootstrap landmark annotations using text parsers to identify landmark phrases in instructions. We then use powerful image-text co-embedding models [8] combined with weak supervision from pose traces to ground those landmarks in the environment.

Modeling & Evaluation

Our two-stage MARKY-MT5 (land**mark** and **m**ultilingual **T5** [15]) system enhances instruction generation by improving how visual landmarks are selected and mentioned. Given a path-connected sequence of panoramic views, the first stage **landmark detector** (trained on the data automatically bootstrapped from human annotations which informs how humans select landmarks, i.e. *silver landmarks*) infers a sequence of landmarks (i.e. *predicted landmarks*) that a person might select for describing the path. E.g., in Fig. 1 eight landmarks are selected, each represented by an image. This sequence, plus interleaved descriptions of navigation actions, is passed to the second stage **instruction generator** – a multimodal extension of the multilingual T5 (mT5) model [15] similar to VL-T5 [4] – to produce the instruction in Fig. 1.

The quality of the generated instructions is evaluated with a) ***large-scale human evaluation*** (with over 20k navigation sessions) to gauge human followability; b) comparing MARKY-MT5 generated instructions and human-written ones for the same paths on SotA navigation agent.

[1]The term *silver data* refers to high-quality annotations–not created by people–that are derived by combining models and constraints [7, 12, 14].

Findings & Conclusion

First, ***landmarks matter***. On R2R, MARKY-MT5 increases success rate (57.8% → 70.8%) and SPL (48.7% → 59.8%), and lowers navigation error (3.9m → 2.9m) compared to prior work without landmarks – represented by SpkFol-RxR trained on the same dataset (Tab. 1, row 5 vs. 3). Further, ***compared to human-written instructions***, using a combination of RxR data, silver landmarks and modeling improvements, we almost eliminate the gap between model-generated and human-written instructions on paths of R2R-level difficulty – with a 71% success rate vs. 75% for human instructions and 42% for previous models (Tab. 1). However, on the more challenging RxR-style paths, a gap remains – human wayfinders obtain a 62% success rate using MARKY-MT5 vs. 78% for human instructions. MARKY-MT5 generated instructions are also ***indistinguishable from human-written ones for a state-of-the-art VLN agent*** [3] – we achieve near identical success rates (56.5% vs. 55.7%) and NDTW (62.9% vs. 63.3%) for human and generated instructions. Finally, an appealing property of our two-stage approach is that ***diverse instructions*** can be generated by sampling landmark predictions.

Despite the accomplishments, the strength of our approach–***focusing on visual landmarks–is also a limitation***. MARKY-MT5 is blind to other context when generating, making it susceptible to pragmatic failures, e.g. generating *'Leave the room'* in a room with multiple exits. Addressing this could lead to further gains.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *3DV*, 2017. 1

[3] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, 2021. 2

[4] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *Proceedings of ICML 2021*, 2021. 2

[5] Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *NeurIPS*, 2019. 2

[6] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, 2018. 1, 2

[7] Udo Hahn, Katrin Tomanek, Elena Beisswanger, and Erik Faessler. A proposal for a configurable silver standard. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 235–242, Uppsala, Sweden, July 2010. Association for Computational Linguistics. 2

[8] 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. 2

[9] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *Proceedings of EMNLP*, 2020. 2

[10] Shuhei Kurita and Kyunghyun Cho. Generative language-grounded policy in vision-and-language navigation with bayes' rule. In *ICLR*, 2021. 1

[11] Vivien Mast and Diedrich Wolter. A probabilistic framework for object descriptions in indoor route instructions. In Thora Tenbrink, John Stell, Antony Galton, and Zena Wood, editors, *Spatial Information Theory*, pages 185–204, Cham, 2013. Springer International Publishing. 1

[12] Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. The CALBC silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). 2

[13] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL*, 2019. 1, 2

[14] Qingrong Xia, Zhenghua Li, Rui Wang, and Min Zhang. Stacked AMR parsing with silver data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4729–4738, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[15] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. 2

[16] Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alex Ku, Jason Baldridge, and Eugene Ie. On the evaluation of vision-and-language navigation instructions. In *EACL*, 2021. 1