ASSISTER: Assistive Navigation via Conditional Instruction Generation

Zanming Huang*, Zhongkai Shangguan*, Jimuyang Zhang, Gilad Bar, Matthew Boyd, and Eshed Ohn-Bar

Boston University, Boston MA 02215, USA
{huangtom,sgzk,zhangjim,gbar,mcboyd,eohnbar}@bu.edu

Abstract. We introduce a novel vision-and-language navigation (VLN) task of learning to provide real-time guidance to a blind follower situated in complex dynamic navigation scenarios. Towards exploring real-time information needs and fundamental challenges in our novel modeling task, we first collect a multi-modal real-world benchmark with in-situ Orientation and Mobility (O&M) instructional guidance. Subsequently, we leverage the real-world study to inform the design of a larger-scale simulation benchmark, thus enabling comprehensive analysis of limitations in current VLN models. Motivated by how sighted O&M guides seamlessly and safely support the awareness of individuals with visual impairments when collaborating on navigation tasks, we present ASSISTER, an imitation-learned agent that can embody such effective guidance. The proposed assistive VLN agent is conditioned on navigational goals and commands for generating instructional sentences that are coherent with the surrounding visual scene, while also carefully accounting for the immediate assistive navigation task. Altogether, our introduced evaluation and training framework takes a step towards scalable development of the next generation of seamless, human-like assistive agents.

Keywords: goal-driven instruction synthesis, vision-and-language navigation, assistive technologies, visual impairment.

1 Introduction

Embodied Vision-and-Language Navigation (VLN) tasks [5,105,6,94,85,64,84] generally *assume a sighted following agent*, i.e., a situated robot [54,78,57,40] or human [16,91,25,13] that is visually perceiving their immediate surroundings while interpreting instructions. As a result, the utility of current VLN systems in assisting *blind navigators* during complex and dynamic navigation is rarely explored, despite immense societal potential for improving the quality-of-life of blind individuals [96,75,55,58]. How well can current visually-grounded language generation methods, which are often studied in static indoor scenes with generic instructions [57,40,13], learn to consider the intricate task-driven

*Equally contributed.

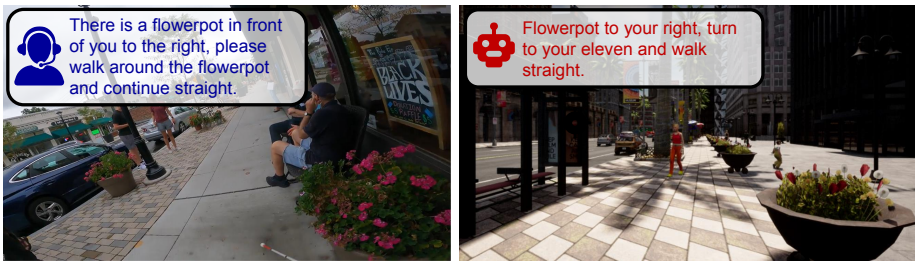


Fig. 1. Assistive Vision-and-Language Navigation (VLN) With a Situated Blind Walker. Our goal is to develop VLN agents that can consider the abilities of a blind walker when seamlessly providing task and safety-based contextual cues. Left: Real-world ego-centric image from the perspective of a blind participant in our dataset, with overlaid navigational instructions provided by an Orientation and Mobility (O&M) expert. Right: First-person view of a simulated pedestrian navigating an urban sidewalk with procedurally generated instructions overlaid.

and potentially dangerous process [81,74,28,79,51,12,95] of non-visual perception, decision-making, and exploration? Towards advancing the state-of-the-art of assistive VLN-based systems, we introduce diverse benchmarks and tools for training task and safety-critical agents that can collaborate with blind individuals.

Currently, there are two key challenges hindering the scope and development of learning-based assistive navigation systems. First, the ***difficulty and cost in obtaining sufficiently diverse data*** for training robust assistive agents, i.e., through IRB-approved user-studies, is prohibitive. Consequently, current computer vision tasks related to navigation (e.g., human motion modeling [72,34]) provide limited insights in our context as they do not incorporate blind navigators. Constrained by practical considerations, assistive technology researchers have mostly pursued studies within constrained navigational settings [43,28,92,24,1,65] (e.g., basic navigational layouts, no dynamic pedestrians, minimal acoustic noise, etc.). Second, compared to current VLN tasks, the addition of factors related to non-visual reasoning and safety requires more ***elaborate modeling of the information needs of the blind navigator*** [9,3,32,42]. For example, Orientation and Mobility (O&M) experts undergo specialized training to go beyond generic instruction and effectively accommodate various needs across diverse settings [80,52]. Due to these inherent challenges, developing learning-based assistive systems for maintaining the real-time awareness of a blind agent to visual and tactile context across diverse settings remains a grand challenge [22,95,96,9,102].

Based on our survey of prior work in Sec. 2, we realized how the instructional guidance properties of assistive systems are also often manually set and hand-tuned in a somewhat cumbersome and setting-specific manner [9,32,3,28,31,42]. Consequently, most aforementioned systems have been both developed and deployed within the same singular setting and fixed environment. In contrast, an

O&M expert can flexibly provide seamless and safe guidance under arbitrary conditions, i.e., through comprehensive understanding of the needs of a blind navigator. In this work, we sought to develop a paradigm for endowing machines with similar capabilities, as described next.

Contributions: Towards facilitating robust, safe, and scalable assistive VLN systems, we make three key contributions: (1) We collect a real-world multi-modal benchmark with diverse in-situ interactions between O&M experts and blind navigators during navigation in dynamic urban settings. (2) We develop and analyze a corresponding simulation environment based on CARLA [19] that is informed by the real-world task. (3) We leverage the two benchmarks to uncover new insights regarding the instructional design space and the extent to which a state-of-the-art VLN model can learn to imitate expert sighted guides. Our benchmark and models are publicly available at https://github.com/h2xlab/ASSISTER. While we envision our findings to benefit individuals with visual impairments, our results translate towards developing expressive, safe, and less biased VLN agents that can robustly model what, when, and how guidance should be given to diverse end-users in real-time.

2 Related Work

Our goal is to understand and model the assistive VLN task in the context of navigation with blind walkers. Our work builds on recent advancements in visually-grounded language generation and assistive navigation, as described next.

Vision-and-Language Navigation: While most prior work has focused on the VLN task of instruction understanding and execution [56,11,57,76,82,83,45,20,63,78,6,61,21,60,104,47], generic instruction generation [62,16,25,91] has recently received more attention with the introduction of suitable benchmarks (see Table 1). Recent advancements in this space aim to create more realistic instructional models, mostly set in static indoors setting [47,6,25,85,62,70,33], e.g., to find an item, or localization in outdoor environments [13,91]. Related to our work is the speaker-follower model of Fried et al. [25] where a speaker model is used for data augmentation and pragmatic selection of the most effective instructions. While relevant (our model can be interpreted as a speaker model), we learn our speaker model via imitation learning [64,67]. Moreover, our language space also includes more fine-grained obstacles and orientation directions. This enables us to empirically explore the optimal instructions to guide a blind navigator under safety-critical constraints and complex dynamic settings, i.e., beyond instruction following on the indoor R2R task [6]. In general, instructions in the aforementioned studies are also centered around visual cues, making this task inaccessible to individuals with vision impairments who may rely on spatial or tactile cues.

Visual Question Answering and Dialog: In Visual Question Answering (VQA), the inputted data may be a static image with a goal of understanding

Table 1. Comparison With Related Benchmarks for VLN. Compared to other photo-realistic navigation-centered datasets, our datasets (UrbanWalk-Sim and UrbanWalk-Real) analyze contextual cues and guidance instructions for navigators who are blind (Blind). We also emphasize navigation involving dynamic obstacles (Dynamic) during outdoor scenarios (table marks In/Outdoor). We also note the number of samples in each dataset (Size) and source of the language annotations (Collection).

Dataset	Dynamic	In/Out	Real-World	Blind	Size	Collection
R2R [6]	✗	I	✓	✗	21,567	Crowdsourced
CVDN [85]	✗	I	✓	✗	7,000	Crowdsourced
REVERIE [70]	✗	I	✓	✗	21,702	Crowdsourced
Touchdown [13]	✓	O	✓	✗	9,326	Crowdsourced
Talk the Walk [91]	✓	O	✓	✗	10,000	Crowdsourced
RxR [48]	✗	I	✓	✗	126,069	Crowdsourced
WAY [37]	✗	I	✓	✗	6,154	Crowdsourced
UrbanWalk-Sim (Ours)	✓	O	✗	✓	399,126	Generated
UrbanWalk-Real (Ours)	✓	O	✓	✓	2,395	In-Situ

what is being asked through Natural Language Processing (NLP) and gathering information from visual cues to answer a specified question [7,100,41,71,99]. Recent studies have also developed two-sided dialog as an extension of VQA [46,17,87,86]. VQA tasks have been recognized for its potential in assistive research as an aid for blind individuals [10,36,35]. While motivating to our study, our task focuses on how to navigate an individual from a current location to a target destination safely. Thus, we extend the concept of VQA to gather the visual information and communicate it via effective dialog to an individual who otherwise cannot utilize visual cues.

Orientation and Mobility Studies: In order to effectively guide blind individuals, accessibility researchers have long studied best navigational practices for people with visual impairments [96,97]. How to best support self-reliance, i.e., for everyday travel, is still an open research question [44,95,81,43]. While O&M guides can support the learning and memorization of a route [81], this is often a slow and lengthy process. Moreover, optimal real-time support in unfamiliar settings is challenging, due to factors such as cognitive load, dynamic obstacles, and ambient noise. This may explain some variability we find among the guides in our study. While there may not be universally accepted preferences among blind walkers due to various orientation and mobility skills [2,43], it is known that clock-based orientation descriptors are generally preferred [44]. We leverage such prior work when designing our instructions in simulation (Sec. 4.1) to ensure our models learn to support users' own mobility and orientation while collaborating effectively.

Assistive Navigation Technologies: There is extensive related research in designing non-AI assistive navigation technologies [77,30,68,92,28,24,65]. A relevant study is the work of Arditi et al. [8], which demonstrates speech to be a preferred assistance modality due to the minimal initial training requirements.

However, studies considering blind individuals are often performed in indoor environments with simplified route stimuli (e.g., narrow corridors, minimal obstacles, clear acoustics). Moreover, many current assistive technologies do not learn to generate instructions at all, but instead rely on extensive rule-based hand-engineered instruction, which may not generalize beyond simplified indoor environments with perfect perception and sparse route stimuli. In contrast, we aim to provide the foundations for future development of more human-like systems, i.e., systems that can seamlessly scale to operate in dynamic and real-world conditions. This goal motivates us to go beyond many prior assistive AI research tasks performed in simulation [69,23], and analyze naturalistic conditions with a real-world benchmark in Sec. 4.1 in addition to our simulation. Our study is motivated by the success of commercial real-time smartphone-based assistance apps that are based on an assistive remote human [59,4]. While very costly to use, the usability of such systems guides our real-world study design in Sec. 4.1. In particular, we outline a preliminary but highly scalable study design based on remote guidance. This design choice also supports model training from the limited perspective of a wearable assistive system.

3 Method

We introduce a task of learning to synthesize contextual and task-relevant natural language for guiding a blind follower. We emphasize that our VLN settings are inherently more complex compared to prior tasks (Table 1) which do not generally include real-time interaction with dynamic scenes. In this section, we first present our learning framework (Sec. 3.1) and novel ASSISTER model architecture (Sec. 3.2) for generating intuitive goal-conditional instructions. Subsequently, Sec. 4 introduces a novel benchmark with natural language from real-world O&M guides (Sec. 4.1) as well as procedurally generated instructions based on known information needs of blind walkers (Sec. 4.2).

3.1 Conditional Instruction Generation for Assistive Navigation

Problem Setting: We consider the task of learning instruction synthesis from observations $\mathbf{o} = [\mathbf{I}, \mathbf{p}_0, \mathbf{P}] \in \mathcal{O}$ comprising a front-view camera image $\mathbf{I}$, the current position and heading of the instruction follower $\mathbf{p}_0 \in \mathbb{R}^3$ (location and heading in map view), and a planned route $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]$ specified in terms of positional waypoints towards a goal state. Our learning goal is to obtain a mapping function $f_{\boldsymbol{\Theta}} \colon \mathcal{O} \rightarrow \mathcal{W}$, parameterized by $\boldsymbol{\Theta} \in \mathbb{R}^l$, for generating a sequence of instructional tokens $\mathbf{w} = \{w_1, \ldots, w_M\} \in \mathcal{W}$ for guiding a follower along a planned route. In our study, we leverage ubiquitous GPS- and IMU-based localization to obtain location and heading estimates as well as employ A* planning [38] to plan the high-level route. While our trained model should learn to account for inherent location noise in localization methods, more elaborate planning and localization schemes (e.g., SLAM [26]) are orthogonal to our study and are left for future work.

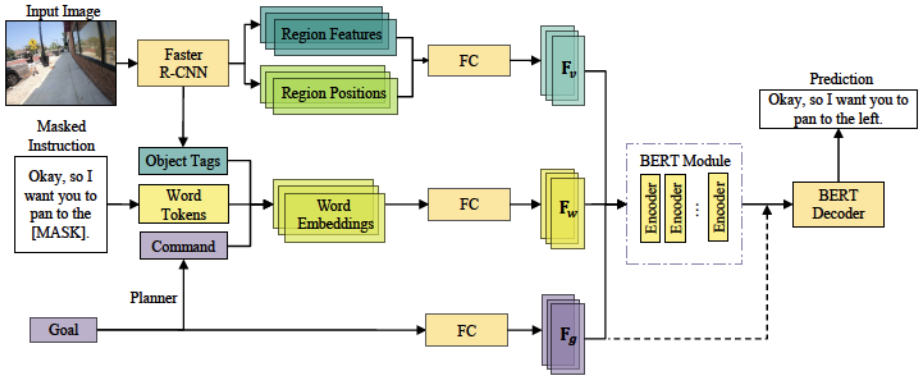


Fig. 2. ASSISTER Overview. The proposed model interlaces visual semantic, language, and goal-based features to carefully account for the immediate navigation task while maintaining situational awareness to surrounding context.

Imitating Expert Guides: As manually designing generalized assistance in dynamic and intricate real-world settings is challenging, our key insight is to leverage human-human interactions to design and train assistive VLN models. Thus, we assume access to a dataset with expert instructional guidance $\mathcal{D} = \{(\mathbf{o}_i, \mathbf{w}_i)\}_{i=1}^{N}$ in order to optimize $f_{\boldsymbol{\Theta}}$ and generate intuitive instructions. Consequently, an instruction generation model can be trained by optimizing a behavior cloning [14,15,103,101] objective

$$\underset{\boldsymbol{\Theta}}{\text{minimize}}\, \mathbb{E}_{(\mathbf{w},\mathbf{o})\sim\mathcal{D}} \left[\mathcal{L}(\mathbf{w}, f_{\boldsymbol{\Theta}}(\mathbf{o}))\right] \tag{1}$$

where $\mathcal{L}$ is a sequence prediction cross-entropy loss [49]. As Eqn. 1 involves aligning high-dimensional vision and language semantics for task-driven navigation, it involves a challenging optimization task. In addition to leveraging supervision from expert guidance annotations, we alleviate training issues through a suitable model structure, discussed next.

3.2 Network Architecture

We introduce strong computer vision and language priors into our model architecture. The priors enable more efficient learning of integrating visual scene context with navigational and language reasoning when assisting a blind person. Our model comprises three main components: (i) a visual semantics feature extractor for obtaining an object-based embedding from an input image, (ii) a self-attention-based [89] language generation module that semantically aligns instructional language with visual context, and (iii) a goal-conditional module which integrates navigational task reasoning. The overall architecture is illustrated in Fig. 2.

Object-based Visual Semantics Feature Extractor: Instead of optimizing $f_{\boldsymbol{\Theta}}$ from raw images, we first extract rich object-based context using a pre-trained

object detector. A Faster R-CNN [73] object detector with a ResNet-101 [39] backbone that is pre-trained on COCO [50] is used to extract and embed region features $\mathbf{F}_v \in \mathbb{R}^{50\times768}$ from the 50 highest scored regions in the image.

Language Generation Module: We integrate the visual embedding $\mathbf{F}_v$ with a state-of-the-art BERT language generation module [49,33,53,18]. The language module is pre-trained on a large language corpus following Li et al. [49] to facilitate natural human-like language synthesis. To further enable semantic image-language alignment, we follow [49] and extract word-based features $\mathbf{F}_w \in \mathbb{R}^{90\times768}$ from 30 explicit object tags, 20 commands and 40 instructional tokens for the current image. Based on BERT, the language module leverages a masked language objective where the model learns to recover masked instructional words from image and sentence context. Note that during inference the model sequentially infers instructional words, i.e., without access to ground truth instructional words.

Goal-Conditional Module: In our domain, safe and seamless navigation pivots on the ability of the model to perform extensive goal-based reasoning. To guide a blind follower in diverse settings, our instruction synthesis model must carefully consider the navigation goal to convey to blind followers only task-relevant surrounding information at any given moment. We therefore interlace goal-based features throughout the entire visually-grounded language generation process. We incorporate a goal embedding $\mathbf{F}_g \in \mathbb{R}^{768}$ computed based on a relative goal vector $\mathbf{g} \in \mathbb{R}^2$ to a near-range (five meters) waypoint along the planned route. We note that our assumption of knowledge of relative position and heading to a goal is standard when learning real-world vision-based navigation, e.g., [14,15,66].

Command-Conditional Module: In addition to the goal embedding, we propose to also leverage navigation commands generated from a *future planned path* to ease the learning of alignment among modalities. Specifically, we directly input the model with conditional navigational commands obtained via a path planner (e.g., 'turn left,' 'forward'). We input the commands as word tokens prior to computing the aligned word-based features $\mathbf{F}_w$. In this manner, the model can learn to generate natural goal-driven instructional sentences that are not only coherent with a visual scene but also account for the immediate navigation task.

4 The UrbanWalk Benchmark

Despite ample publicly available language benchmarks, there are no current datasets suitable for model training and evaluation of timely, safety-critical, and ability-aware navigation guidance to blind followers. Moreover, prior VLN tasks tend to leverage human-written instructions in simulations and not relevant instructions for providing in-situ navigation cues. Towards exploring real-time information needs and fundamental challenges in our novel modeling task, we collect the first multi-modal real-world benchmark with recorded O&M instructional guidance in dynamic urban walking navigation settings (Sec. 4.1). Subse-

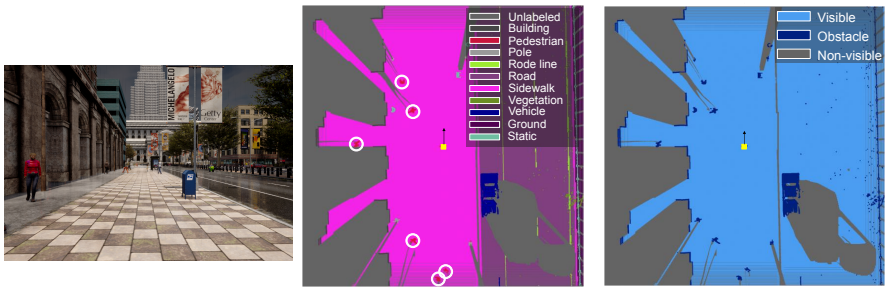
quently, we leverage the real-world study to inform the design of a larger-scale simulation-based benchmark (Sec. 4.2) and comprehensively analyze limitations in current VLN models across diverse scenarios (e.g., harsh weathers, geographical locations, etc.). Altogether, the two datasets are used to produce complementary analysis while tackling inherent issues in safety, cost, and scalability of real-world data collection with blind participants.

4.1 Real-World Benchmark

Although VLN tasks are often studied in simulated settings, realistically simulating interactions between a blind walker, their surroundings (e.g., acoustics, objects), and an expert guide is not trivial. Hence, we pursued a real-world study to ensure our models and findings are relevant to practical navigation scenarios.

Study Design: Our IRB-approved study was kept close to others in the field in terms of participant pool and mobility aids [24]. However, we are the first to collect synchronized multi-modal camera and sensor data together with their *corresponding in-situ expert instructional guidance*. We recruited 13 participants through the mailing list of a local blind individuals services center, including 10 blind and three O&M guides (to analyze expert diversity). To train our imitation learning-based assistive agent in Sec. 3.1, we sought to collect video and sensor measurements during blind navigation in real-world urban scenes with expert guidance *from the perspective of an assistive system*, i.e., a first-person camera. Therefore, in order to capture naturalistic navigation behavior and real-world challenges associated with assistive technologies, we opted for a remote guidance solution. While the limited perspective incorporates a practical challenge, this study design choice also lends to scalability due to minimal mount configuration, ease of data collection, and ultimate large-scale deployments on commodity devices, e.g., smartphones.

Navigation Task: We asked the blind participants to navigate an unfamiliar 110m planned route through a busy business district with typical weekday traffic, including pedestrians, vehicles, and shops. We ensured control for confounding factors: participants were called on different days and on varying hours. The equipment included a 5G smartphone, an additional GoPro camera mounted to a chest harness, and a Bluetooth bone-conducting headset to provide instructions without hindering acoustic reasoning. GPS, IMU, audio, and camera data were all captured synchronously. We note that the restricted forward view provided by a chest-mounted camera rarely provided a complete view of the surroundings and potential obstacles. This necessitated crucial collaboration between the blind navigator and the guide, an interactive functionality that we wish to embody in our assistive agent. For instance, in order to gather sufficient visual information for safe navigation the expert may ask the navigator to stop and scan the environment by rotating their torso to pan and tilt the camera. Audio transcription was performed in a semi-automatic manner, initially with Google's Speech-to-Text [29] followed by manual verification and error correction.



(a) Front Image (b) Semantic BEV (c) Walkable Space

Fig. 3. Simulation Visualization. We visualize (a) first-person view of a virtual walker, (b) corresponding BEV with semantics overlay, with a yellow square indicating the current position and heading of the walker (circles indicate surrounding pedestrians), and (c) walkable space computed from the semantic BEV. We then sample goals randomly, plan a path in walkable space, and generate contextualized instructional guidance from the path and semantic BEV objects.

Dataset: We extract a total of 2,395 interactions (on average, there are 21.7 words per instruction) from the continuous data stream. Example conversational language from the dataset for supporting guidance and situational awareness include:

> *"Okay, you're going to walk directly to the street and there's going to be a detectable curb. This is a cross-walk."*

> *"Good job, You're passing some bushes on the right. You might contact those with your body."*

> *"I'm going to have you turn to the right, so I can see that area."*

In addition to route-based instructions, we find the naturalistic instructions to regularly employ cues related the spatial layout, obstacles, and information gathering.

4.2 Simulation Benchmark

In our analysis, we sought to fully capture the complexity of naturalistic real-world in-situ interaction. However, despite our attempt towards a more scalable study design, real-world data collection is inherently limited with issues of safety, cost, and data diversity. We therefore supplement our analysis by leveraging a simulation (based on the CARLA environment [19]) which emulates our task without such constraints. While CARLA is typically used for development of autonomous driving policies, we modify the environment to collect instructional guidance and a sidewalk pedestrian perspective in various weathers and towns. In particular, we use the large synthetic dataset to rigorously analyze model

limitations across ample standardized data and more diverse visual conditions (e.g., new towns, harsh weathers). Nonetheless, we use the real-world navigational data (Sec. 4.1) to guide our simulation design as well as draw general conclusions among both benchmarks.

Procedural Instruction Generation: To collect a large simulation benchmark, we procedurally generate instructional guidance. We spawn navigating pedestrians and capture a first-person image perspective together with complete ground-truth information of surrounding landmarks and obstacles (i.e., 3D location of buildings, pedestrians, sidewalks, trees, etc.). Given a current walker position, a sampled goal, and a constructed Bird's-Eye-View (BEV) image, we extract walkable space and obtain a path using A* planning [38] a visualization of this process is shown in Fig. 3). We then employ the planned path to construct instructional sentences. We contextualize the instructions by extracting surrounding obstacle information from the BEV along the path and inform regarding obstacles in proximity (e.g., pedestrians, building). While this process can be used to generate standardized instructions, we leverage insights from our real-world study together with prior literature in orientation and mobility strategies [44,95,81,9,81] to consider relevant navigation strategies and immediate information needs. For instance, we leverage clock orientation to indicate turning which has been found to be more intuitive for blind users [44].

5 Experiments

Our goal is to facilitate assistive systems at scale, we emphasize model generalization across various settings and instructional guidance. In this section, we comprehensively analyze our assistive VLN task through the introduced benchmarks and task-conditional ASSISTER model.

5.1 Experimental Setup

CARLA [19] is not generally used to study instruction generation with navigators along the sidewalk. Next, we detail our data split strategy, including weather and ambient factors.

We use Town 5 of the simulation for collecting training data and Town 10 for testing. We randomly spawn pedestrians and goals in dense settings [15]. While we avoided harsh rain conditions due to safety concerns in the real-world data, it does contain natural variations in weather (including sunny weather and two sessions in slight rain conditions). While the real-world data is smaller in size, it also contains significant variability and diversity in the naturalistic instructions. We thus analyze a participant-based split. The overall event distribution in the datasets is depicted in Fig. 4. To facilitate meaningful analysis over the conversational nature of the real-world instructions, Fig. 4 plots a distribution of clustered instructions by types.

Language-based Metrics: We follow standard language evaluation using BLEU-4, CIDEr, and SPICE [90,91,85,104] metrics. We note that BLEU-4 is

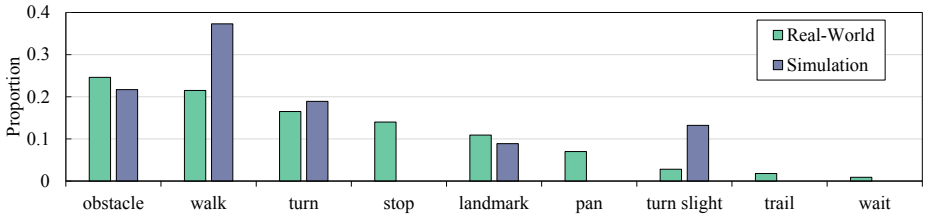


Fig. 4. Data Statistics. We cluster natural language instructions into types to present a high-level analysis of events among the two introduced datasets.

an n-gram based metric that puts equal weights to all words in a given sentences (including pronouns and connective words). Hence, it may be less relevant to our goal-driven navigation task, and is kept as reference. The recent work of Zhao et al. [104] suggests SPICE may be correlated with human wayfinding performance. However our overall navigation settings significantly differ from [104]. For instance, SPICE may be limited for our novel task due to its semantic graph which fails to properly generalize to our instructional context. Specifically, SPICE may fail to account for small but task-relevant changes in guidance, such as changing 'two,' in 'turn to two,' to 'ten.' In contrast, CIDEr identifies informative n-grams in the data from computed term-frequencies. This gives lower weight to more common n-grams, since they are likely to contain less information. Among the standard evaluation metrics we used, we qualitatively found CIDEr to produce slightly better results for our task, as it puts less emphasis on non-key words that occurs frequently appears across instructions such as 'to' and 'your,' while giving more weight to less frequent informative directional words such as 'left' and 'right.' Nonetheless, the two metrics are generally correlated based on our results.

Task-based Interactive Evaluation: Offline language-based evaluation metrics (e.g., CIDEr, SPICE) may not fully account for our sequentially interactive assistive navigation task [104]. For instance, errors in small but critical components of an instruction (e.g., confusion of 'left' with 'right') could have large impact on the ultimate success of the navigation task. While we take a first step towards learned guidance models, several key challenges in safety and model robustness must still be tackled before real-world usability testing with ASSISTER can be performed. To provide further insights into interactive task-based model evaluation, we instead turn to our simulation environment and perform a user study with seven participants navigating routes in a blind simulation. To simulate blind navigation, only coarse orientation with noise up to 15 degrees and collision information is presented as participants following audio instructions. We emphasize that **no image of the scene is shown** as human controllers navigate in our simulation in real-time. The instructions are either generated by our procedural process (i.e., employing ground truth information about the surroundings and route) or sampled from the proposed model. In this manner, we

can directly evaluate the ability of the model to guide to a goal location successfully in complex, dynamic, and previously unseen test settings. Our model runs at about ten frames per second on a desktop with GeForce RTX 2080 Ti, which is sufficient given the time it takes to produce and follow instructions. Moreover, participants can also press a keyboard key in order to query the model as needed. We first familiarize the participants with the walker physics through several short training episodes where front-view image is available. We then evaluate participant route following behavior in our New Town and Weather test settings without image information. Following standard evaluation of navigation agents [19], we also timeout episodes beyond five minutes. To better understand model performance and limitations, we leverage our interactive simulation in a preliminary study which enables us to safely obtain metrics related to Success Rate (SR), Route Completion (RC), and Navigation Error (NE) [93]. We also report the average number of model queries by the human controller (per minute) and collision counts.

5.2 Results

To uncover challenges in our novel task and benchmark, we perform three main experiments. First, we evaluate the role of various inputs to the model on instruction generation and generalization in simulation. Second, we analyze model performance in the real-world data. Third, we analyze task-based performance of humans following instructions in a blind simulation.

Instruction Generation in Simulation: Table 2 analyzes model generalization across seen and unseen settings of new town and harsh weather conditions. Standard deviation results are shown over training runs. We also compare to two main baselines, OSCAR [49] and the Hard-Attention LSTM model of Xu et al. [98]. The results demonstrate the benefits of incorporating command-conditional input to the model when generating task-relevant navigational guidance. Specifically, we find our proposed conditional module to significantly outperform a goal-only ASSISTER model (only the goal vector fused into the model before and after the BERT decoder) across evaluation settings. While the trends are generally consistent across the language-based metrics, our findings demonstrate the overall challenging nature of our task. We also find weather and geographical perturbations to degrade performance of the model, in particular to unseen weathers. Given these insights, we now turn to study the models in our real-world dataset.

Instruction Generation in the Real-World: The real-world contains significantly more walker and guidance diversity due to the complex scenes and freeform instructional guidance. As shown in Table 3, this results in a significantly a challenging modeling task. While simulated data lacks realism, the resulting models suggest that our designed instructions in simulation are realistic. For instance, Table 3 shows generally similar trends to Table 2. Nonetheless, both CIDEr and SPICE are shown to be degraded, with the best performing

Table 2. Simulation Instruction Generation Results. Ablative results over ASSISTER model inputs, language metrics, and test conditions.

	Training Conditions			New Town			New Town and Weather		
Model	**BLEU-4**	**CIDEr**	**SPICE**	**BLEU-4**	**CIDEr**	**SPICE**	**BLEU-4**	**CIDEr**	**SPICE**
Xu et al. [98]	10.18 ± 0.03	15.46 ± 0.05	7.66 ± 0.02	6.81 ± 0.04	18.96 ± 0.05	6.39 ± 0.12	8.36 ± 0.01	15.19 ± 0.04	8.72 ± 0.06
OSCAR [49]	10.68 ± 0.04	17.63 ± 0.10	24.77 ± 0.03	9.25 ± 0.06	14.89 ± 0.07	**21.30 ± 0.11**	9.96 ± 0.01	11.94 ± 0.01	**22.02 ± 0.16**
ASSISTER (Goal Module)	10.58 ± 0.02	16.97 ± 0.03	23.98 ± 0.02	9.45 ± 0.01	14.49 ± 0.01	18.69 ± 0.07	9.97 ± 0.01	11.16 ± 0.01	19.56 ± 0.04
ASSISTER (Goal+Command Module)	**15.49 ± 0.03**	**23.58 ± 0.04**	**26.31 ± 0.01**	**12.81 ± 1.09**	**19.93 ± 0.51**	18.63 ± 2.24	**13.88 ± 0.61**	**19.61 ± 2.51**	21.88 ± 0.17

Table 3. Real-World Instruction Generation Results. Ablative results over ASSISTER model variants, language metrics, and evaluation settings.

	Cross-Subject		
Model	**BLEU-4**	**CIDEr**	**SPICE**
Xu et al. [98]	0.00 ± 0.00	1.17 ± 0.54	1.31 ± 0.55
OSCAR [49]	2.40 ± 0.65	10.47 ± 2.09	8.89 ± 1.58
ASSISTER (Goal Module)	2.43 ± 0.36	10.42 ± 2.36	8.98 ± 1.54
ASSISTER (Goal+Command Module)	**2.50 ± 0.36**	**10.74 ± 3.47**	**9.14 ± 1.47**

model resulting in a 10.74 and 9.14 accuracy, respectively. There are several reasons that explain the low overall performance. First, there are natural variations among the guides when providing instructional context. Current language-based evaluation metrics cannot properly account for such variations. This challenge also motivated us to pursue a task-driven evaluation as a final experiment. Second, guides are able to accurately reason over scene acoustics and walker behavior. Integrating such information requires further study in the future. As safely generating instructions in the real-world is still beyond reach, we now turn to evaluation of the instruction generation model in simulation.

Task-based Evaluation in Simulation: We do not deploy our models in real-world settings with blind users to generate on-policy task-based evaluation. While task-based evaluation is the most informative, current state-of-the-art VLN models cannot be safely evaluated in closed-loop real-time scenarios with blind followers. Instead, we design an interactive blind simulation experiment by removing all visual display. Such closed-loop evaluation is critical in term of assistive navigation and highlights the benefits of the introduced simulation environment. Compared with the baseline model, our ASSISTER achieves high improvement in terms of success rate, route completion and navigation errors, indicating the effectiveness of the proposed method 4. We note that live navigation in the simulation without any visual feedback results in a highly challenging task. We therefore also benchmark our ground truth procedural generation process. The high route completion score (98.6%) further validates our instruction generation process in simulation. We also note that the baseline model of Xu et al. exhibits lower collision rates. However, this is partly due to frequent veering from the planned path to open spaces with less obstacles. Moreover, despite the low success rate for ASSISTER (38.1%), the high route completion results (74.1%) suggest generally suitable instructions are provided to the participants. However, timeouts can occur due to veering off the path as well, which partly contributes to the low success rate. Another main limitation is in the lack of realism of the walker, which can sense acoustic, motion and spatial proper-

Table 4. Task-based Evaluation in Simulation. Navigation following performance of humans in an interactive blind simulation. We show results using ASSISTER-based and ground-truth (using our procedural BEV-based instruction generation process), highlighting the challenging nature of our task. We show the average number of queries from the human walker, per minute, as well as collision events frequency with 'D' (dynamic obstacles, pedestrians) and 'S' (static obstacles).

Model	SR↑	RC↑	NE↓	Queries/min↓	Collision-D↓	Collision-S↓
Xu et al. [98]	9.52	46.3	3.53	11.31	**1.67**	**6.0**
ASSISTER	**38.1**	**74.1**	**2.72**	**8.65**	2.14	14.0
Ground Truth	90.5	98.6	1.02	7.92	1.71	3.10

ties in the real-world. While the process in which a blind person interprets and reacts to surrounding environmental properties and guidance cognitive load is complex [27,88], realizing such reasoning in simulation is still a current open problem and a potential future direction. While our simulation study take a first step towards robust and scalable instruction generation, future improvements can result in additional real-world validation.

6 Conclusion

Our goal is to enable scalable assistive VLN models that can seamlessly and safely guide across diverse walkers and environments. In our study, we tackle learning-based assistive navigation systems through a novel data-driven framework, tools, and analysis. We demonstrate our novel spoken guidance task to provide a challenging setting for VLN models, both in real-world and simulated environments. As future work, transferring models trained in simulation to the real-world could further alleviate issues in cumbersome, costly, and potentially safety-critical real-world studies performed with participants who are blind. Given the potential impact of acoustic properties of the scene on navigation, a next step could explore generalization of the proposed ASSISTER model to include such inputs. While current VLN models and assistive systems do not yet consider acoustic properties, our data can facilitate such models as it was collected in busy urban settings with ambient noise. We kept such data in our experiments in order to ensure our analysis extend to real-world scenarios and usability. Finally, while the participant size is representative of the upper limit of previous studies in accessibility, future studies can replicate our scalable study design to collect data from additional locations and environments. While data can be scarce in our application context, this can facilitate further exploration into the intricate interdependence between a vision-based system and a situated blind navigator.

Acknowledgments: We thank our study participants and the support of the Department of Transportation Inclusive Design Challenge, NSF (IIS-2152077), and a Boston University CISE grant.

References

1. Ahmetovic, D., Gleason, C., Ruan, C., Kitani, K., Takagi, H., Asakawa, C.: Navcog: a navigational cognitive assistant for the blind. In: MobileHCI (2016)
2. Ahmetovic, D., Guerreiro, J., Ohn-Bar, E., Kitani, K.M., Asakawa, C.: Impact of expertise on interaction preferences for navigation assistance of visually impaired individuals. In: W4A (2019)
3. Ahmetovic, D., Murata, M., Gleason, C., Brady, E., Takagi, H., Kitani, K., Asakawa, C.: Achieving practical and accurate indoor navigation for people with visual impairments. In: W4A (2017)
4. Aira: Aira app, https://aira.io/
5. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv (2018)
6. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018)
7. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: ICCV (2015)
8. Arditi, A., Tian, Y.: User interface preferences in the design of a camera-based navigation and wayfinding aid. Journal of Visual Impairment & Blindness (2013)
9. Banovic, N., Franz, R.L., Truong, K.N., Mankoff, J., Dey, A.K.: Uncovering information needs for independent spatial learning for users who are visually impaired. In: ASSETS (2013)
10. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., et al.: Vizwiz: nearly real-time answers to visual questions. In: UIST (2010)
11. Blukis, V., Paxton, C., Fox, D., Garg, A., Artzi, Y.: A persistent spatial semantic representation for high-level natural language instruction execution. arXiv (2021)
12. Brady, E.L., Sato, D., Ruan, C., Takagi, H., Asakawa, C.: Exploring interface design for independent navigation by people with visual impairments. In: ASSETS (2015)
13. Chen, H., Shur, A., Misra, D., Snavely, N., Artzi, Yoav, I., Gould, S., van den Hengel, A.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: CVPR (2019)
14. Codevilla, F., Müller, M., López, A., Koltun, V., Dosovitskiy, A.: End-to-end driving via conditional imitation learning. In: ICRA (2018)
15. Codevilla, F., Santana, E., López, A.M., Gaidon, A.: Exploring the limitations of behavior cloning for autonomous driving. In: ICCV (2019)
16. Daniele, A.F., Bansal, M., Walter, M.R.: Navigational instruction generation as inverse reinforcement learning with neural machine translation. In: HRI (2017)
17. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M.F., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2018)
19. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: CoRL (2017)
20. Duvallet, F., Kollar, T., Stentz, A.: Imitation learning for natural language direction following through unknown environments. In: ICRA (2013)

21. Duvallet, F., Walter, M.R., Howard, T., Hemachandra, S., Oh, J., Teller, S., Roy, N., Stentz, A.: Inferring maps and behaviors from natural language instructions. In: Experimental Robotics (2016)
22. Easley, W., Williams, M.A., Abdolrahmani, A., Galbraith, C., Branham, S.M., Hurst, A., Kane, S.K.: Let's get lost: Exploring social norms in predominately blind environments. In: CHI (2016)
23. Erickson, Z., Gangaram, V., Kapusta, A., Liu, C.K., Kemp, C.C.: Assistive Gym: A physics simulation framework for assistive robotics. ICRA (2020)
24. Fallah, N., Apostolopoulos, I., Bekris, K., Folmer, E.: Indoor human navigation systems: A survey. Interacting with Computers (2013)
25. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: NeurIPS (2018)
26. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artificial intelligence review (2015)
27. Geruschat, D.R., Turano, K.A., Stahl, J.W.: Traditional measures of mobility performance and retinitis pigmentosa. Optometry & Vision Science (7) (1998)
28. Giudice, N.A., Legge, G.E.: Blind navigation and the role of technology. The Engineering Handbook of Smart Technology for Aging, Disability, and Independence (2008)
29. Google: Google speech-to-text, https://cloud.google.com/speech-to-text
30. Granquist, C., Sun, S.Y., Montezuma, S.R., Tran, T.M., Gage, R., Legge, G.E.: Evaluation and comparison of artificial intelligence vision aids: Orcam myeye 1 and seeing ai. Journal of Visual Impairment & Blindness (2021)
31. Guerreiro, J., Ahmetovic, D., Sato, D., Kitani, K., Asakawa, C.: Airport accessibility and navigation assistance for people with visual impairments. In: CHI (2019)
32. Guerreiro, J., Ohn-Bar, E., Ahmetovic, D., Kitani, K., Asakawa, C.: How context and user behavior affect indoor navigation assistance for blind people. In: W4A (2018)
33. Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I., Schmid, C.: Airbert: In-domain pretraining for vision-and-language navigation. In: ICCV (2021)
34. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: CVPR (2018)
35. Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., Bigham, J.P.: Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In: CVPR (2019)
36. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018)
37. Hahn, M., Krantz, J., Batra, D., Parikh, D., Rehg, J.M., Lee, S., Anderson, P.: Where are you? localization from embodied dialog (2020)
38. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics (1968)
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
40. Hu, Z., Pan, J., Fan, T., Yang, R., Manocha, D.: Safe navigation with human instructions in complex scenes. IEEE Robotics and Automation Letters (2019)
41. Hudson, D.A., Manning, C.D.: Gqa: a new dataset for compositional question answering over real-world images. In: CVPR (2019)

42. Kacorri, H., Kitani, K.M., Bigham, J.P., Asakawa, C.: People with visual impairment training personal object recognizers: Feasibility and challenges. In: CHI (2017)
43. Kacorri, H., Mascetti, S., Gerino, A., Ahmetovic, D., Takagi, H., Asakawa, C.: Supporting orientation of people with visual impairment: Analysis of large scale usage data. In: ASSETS (2016)
44. Kamikubo, R., Kato, N., Higuchi, K., Yonetani, R., Sato, Y.: Support strategies for remote guides in assisting people with visual impairments for effective indoor navigation. In: CHI (2020)
45. Kollar, T., Tellex, S., Roy, D., Roy, N.: Toward understanding natural language directions. In: HRI (2010)
46. Kottur, S., Moura, J.M.F., Parikh, D., Batra, D., Rohrbach, M.: CLEV-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In: NAACL (2019)
47. Krantz, J., Gokaslan, A., Batra, D., Lee, S., Maksymets, O.: Waypoint models for instruction-guided navigation in continuous environments. In: ICCV (2021)
48. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding (2020)
49. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (2020)
50. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
51. Liu, G., Yu, T., Yu, C., Xu, H., Xu, S., Yang, C., Wang, F., Mi, H., Shi, Y.: Tactile compass: Enabling visually impaired people to follow a path with continuous directional feedback. In: CHI (2021)
52. Long, R.G., Hill, E.: Establishing and maintaining orientation for mobility. Foundations of orientation and mobility (1997)
53. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for Vision-and-Language Tasks. In: NeurIPS (2019)
54. Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A platform for embodied ai research. arXiv (2019)
55. Marston, J.R., Golledge, R.G.: The hidden demand for participation in activities and travel by persons who are visually impaired. Journal of Visual Impairment & Blindness (2003)
56. Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., Fox, D.: A joint model of language and perception for grounded attribute learning. In: ICML (2012)
57. Matuszek, C., Herbst, E., Zettlemoyer, L., Fox, D.: Learning to parse natural language commands to a robot control system. In: Experimental Robotics (2013)
58. Maunder, D., Venter, C., Rickert, T., Sentinella, J.: Improving transport access and mobility for people with disabilities. In: CILT (2004)
59. Microsoft: Seeing ai app from microsoft, https://www.microsoft.com/en-us/ai/seeing-ai
60. Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., Artzi, Y.: Mapping instructions to actions in 3d environments with visual goal prediction. In: EMNLP (2018)
61. Misra, D.K., Sung, J., Lee, K., Saxena, A.: Tell me dave: Context sensitive grounding of natural language to mobile manipulation instructions. In: RSS (2014)
62. Moudgil, A., Majumdar, A., Agrawal, H., Lee, S., Batra, D.: SOAT: A scene- and object-aware transformer for vision-and-language navigation. In: NeurIPS (2021)

63. Narasimhan, K., Kulkarni, T.D., Barzilay, R.: Language understanding for textbased games using deep reinforcement learning. In: EMNLP (2015)
64. Nguyen, K., Dey, D., Brockett, C., Dolan, B.: Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In: CVPR (2019)
65. Ohn-Bar, E., Kitani, K., Asakawa, C.: Personalized dynamics models for adaptive assistive navigation systems. In: CoRL (2018)
66. Ohn-Bar, E., Prakash, A., Behl, A., Chitta, K., Geiger, A.: Learning situational driving. In: CVPR (2020)
67. Osa, T., Pajarinen, J., Neumann, G., Bagnell, J.A., Abbeel, P., Peters, J.: An algorithmic perspective on imitation learning. arXiv (2018)
68. Peng, H., Song, G., You, J., Zhang, Y., Lian, J.: An indoor navigation service robot system based on vibration tactile feedback (2017)
69. Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.H., Tenenbaum, J.B., Fidler, S., Torralba, A.: Watch-and-help: A challenge for social perception and human-ai collaboration. In: ICLR (2021)
70. Qi, Y., Wu, Q., Anderson, P., Liu, M., Shen, C., van den Hengel, A.: Reverie: Remote embodied referring expressions in real indoor environments. In: CVPR (2020)
71. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. In: NeurIPS (2018)
72. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: ICCV (2019)
73. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. NeurIPS (2015)
74. Rieser, J.J., Guth, D., Hill, E.: Mental processes mediating independent travel: Implications for orientation and mobility. Journal of Visual Impairment and Blindness (1982)
75. Roberts, P.W., Babinard, J.: Transport strategy to improve accessibility in developing countries (2004)
76. Roh, J., Paxton, C., Pronobis, A., Farhadi, A., Fox, D.: Conditional driving from natural language instructions. In: CoRL (2020)
77. Sato, D., Oh, U., Naito, K., Takagi, H., Kitani, K., Asakawa, C.: Navcog3: An evaluation of a smartphone-based blind indoor navigation assistant with semantic features in a large-scale environment. In: ASSETS (2017)
78. Scheutz, M., Krause, E.A., Oosterveld, B., Frasca, T.M., Platt, R.W.: Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In: AAMAS (2017)
79. Schinazi, V.R., Thrash, T., Chebat, D.R.: Spatial navigation by congenitally blind individuals. Wiley Interdisciplinary Reviews: Cognitive Science (2016)
80. Soong, G.P., Lovie-Kitchin, J.E., Brown, B.: Does mobility performance of visually impaired adults improve immediately after orientation and mobility training? Optometry & Vision Science (2001)
81. Strelow, E.R.: What is needed for a theory of mobility: Direct perceptions and cognitive maps—lessons from the blind. Psychological review (1985)
82. Tellex, S., Knepper, R.A., Li, A., Rus, D., Roy, N.: Asking for help using inverse semantics. In: RSS (2014)
83. Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., Roy, N.: Understanding natural language commands for robotic navigation and mobile manipulation. In: AAAI (2011)

84. Thomason, J., Gordan, D., Bisk, Y.: Shifting the baseline: Single modality performance on visual navigation & qa. In: NAACL (2019)
85. Thomason, J., Murray, M., Cakmak, M., Zettlemoyer, L.: Vision-and-dialog navigation. In: CoRL (2019)
86. Thomason, J., Padmakumar, A., Sinapov, J., Walker, N., Jiang, Y., Yedidsion, H., Hart, J.W., Stone, P., Mooney, R.J.: Improving grounded natural language understanding through human-robot dialog. In: ICRA (2019)
87. Thomason, J., Zhang, S., Mooney, R., Stone, P.: Learning to interpret natural language commands through human-robot dialog. In: IJCAI (2015)
88. Turano, K., Geruschat, D., Stahl, J.W.: Mental effort required for walking: effects of retinitis pigmentosa. Optometry & Vision Science (1998)
89. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
90. Vedantam, R., Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
91. de Vries, H., Shuster, K., Batra, D., Parikh, D., Weston, J., Kiela, D.: Talk the walk: Navigating new york city through grounded dialogue (2018)
92. Wang, H.C., Katzschmann, R.K., Teng, S., Araki, B., Giarré, L., Rus, D.: Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In: ICRA (2017)
93. Wang, S., Montgomery, C., Orbay, J., Birodkar, V., Faust, A., Gur, I., Jaques, N., Waters, A., Baldridge, J., Anderson, P.: Less is more: Generating grounded navigation instructions from landmarks. arXiv (2021)
94. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: CVPR (2019)
95. Williams, M.A., Galbraith, C., Kane, S.K., Hurst, A.: "just let the cane hit it" how the blind and sighted see navigation differently. In: ASSETS (2014)
96. Williams, M.A., Hurst, A., Kane, S.K.: " pray before you step out" describing personal and situational blind navigation behaviors. In: ASSETS (2013)
97. Wong, S.: Traveling with blindness: A qualitative space-time approach to understanding visual impairment and urban mobility. Health & Place (2018)
98. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057. PMLR (2015)
99. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.B.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In: NeurIPS (2018)
100. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019)
101. Zhang, J., Ohn-Bar, E.: Learning by watching. In: CVPR (2021)
102. Zhang, J., Zheng, M., Boyd, M., Ohn-Bar, E.: X-world: Accessibility, vision, and autonomy meet. In: ICCV (2021)
103. Zhang, J., Zhu, R., Ohn-Bar, E.: SelfD: Self-learning large-scale driving policies from the web. In: CVPR (2022)
104. Zhao, M., Anderson, P., Jain, V., Wang, S., Ku, A., Baldridge, J., Ie, E.: On the evaluation of vision-and-language navigation instructions. ArXiv (2021)
105. Zhu, F., Zhu, Y., Lee, V., Liang, X., Chang, X.: Deep learning for embodied vision navigation: A survey. arXiv (2021)