Time-Aware Assistive Navigation

Masaki Kuribayashi*[1,2], Zhongkai Shangguan*[2], and Eshed Ohn-Bar[2]

***Abstract*— Can interactive vision-and-language agents learn not just what to say but also *when* to say it? Current language models rarely plan over whether and when to realize a real-time response to a user. However, providing accurate and timely support for human decision-making, such as when guiding visually impaired individuals through urban environments, requires careful real-time responsiveness–poorly timed responses can distract users or add unnecessary cognitive load. As a machine intelligence challenge for Multimodal Large Language Model (MLLM)-based agents, we introduce a large-scale multimodal benchmark for an egocentric, assistive navigation task in complex outdoor environments. Using this benchmark, we uncover a fundamental limitation of off-the-shelf MLLMs in delivering safe and time-sensitive navigation instructions, even with model fine-tuning on substantial amounts of data. We then demonstrate that a simple yet effective modification of the model, including direct supervision to predict the underlying reason for each instruction, yields significant performance gains across open-loop, closed-loop, and sim-to-real generalization settings. However, our analysis highlights persistent challenges in temporal reasoning, safety-critical object awareness, and relational and distance understanding. To advance the development of scalable assistive agents, we will release our simulation, benchmark, and code (available at project website: https://timeli-icra.github.io/).**

I. Introduction

Effective assistive vision-and-language models need to go beyond merely describing their surroundings——they must understand users' needs, anticipate risks, and determine when to intervene, or to remain silent. To realize their potential as real-world assistive agents [1]–[5], such as wearable devices and robotic platforms guiding visually impaired users through unfamiliar environments and chaotic city streets, they must be designed to provide proactive, safe, and timely support. Do current Multimodal Large Language Models (MLLMs), have the necessary spatial, temporal, and planning capabilities needed to maintain continual situational awareness and provide seamless assistance to diverse users?

Consider the recent Be My Eyes and OpenAI [6] demonstration of GPT-4o [7] as an interactive, real-time outdoor assistant for a blind user navigating urban environments (we show overall task example in Fig. 1). While promising in principle, even minor errors in the content and timing of an instruction in this context could have serious consequences, such as leading the user off course or into a dangerous situation. For example, the user might be misled when the model mixes up a single yet crucial word in its response, *e.g.*, *"left"* with *"right,"* which is a common error based on our planning-oriented analysis. Eventually realizing the error, our user may attempt to cross safely while listening to surrounding signals and traffic. However, the verbose model may continue to provide lengthy instructions (another common behavior [8], [9]) *while the user is crossing*, causing significant distraction and cognitive load [10], [11]. In this scenario, the user could attempt to prompt the model with more examples and specifications regarding instructions, *e.g.*, *"The intersection was on the left, not on the right, and don't talk at intersections when user is crossing, it's dangerous."* However, the model does not follow the request as it was not trained with the objective of acting as a mobility guide for a blind user [12]–[14], with very few of the training samples, if any, included blind individuals (*e.g.*, none in Webvid-2M [15], VALOR [16]). In this work, we develop a suitable benchmark and model for instilling usability knowledge into MLLM-based agents.

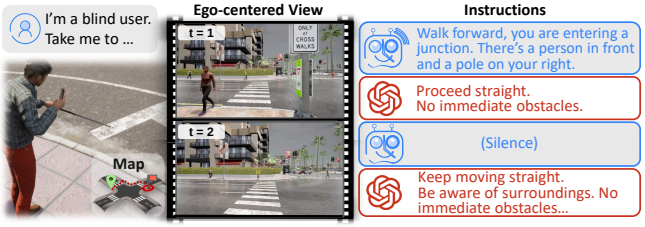

Fig. 1: **Step-by-Step Video Instruction Generation Task.** We introduce **TIMELI**, a novel machine intelligence task and model for providing temporally-aware, concise, and safe assistance to blind users navigating toward a goal. Based on our analysis, we demonstrate how current MLLMs, *e.g.*, ChatGPT (depicted in the figure), provide ineffective and overly verbose instructions for our task. In contrast, the TIMELI-based agent reasons about both what to say and when to say it, ensuring alignment with the needs of blind users, *e.g.*, remaining silent at intersections to reduce cognitive load and prevent hazardous situations, which is a standard strategy employed by mobility guides.

Despite recent progress in temporal video understanding, surprisingly little work has explored safe, real-time assistive navigation instruction. Instead, most approaches focus on video captioning and generic summarization, *i.e.*, producing broad, task-agnostic descriptions [1], [17]–[24] rather than goal-oriented, real-time assistance. For example, VizWiz [1], widely used in MLLM studies [25], primarily focuses on single-frame tasks, such as close-up object and text recognition. Similarly, prior work in Vision-and-Language Navigation (VLN) [3], [5], [26]–[32] generally studies static and simple environments, *i.e.*, without fine-grained relational understanding, reactive dynamic obstacles, complex outdoor layouts, weathers, or continual assistance for disabled users.

* indicates equal contribution. [1]Waseda University; [2]Boston University.

Indeed, even with fine-tuning on ample data, our analysis reveals that most off-the-shelf models demonstrate poor performance across key task metrics.

Contribution: Our goal is to advance user-aware, dynamic, and safe assistive agents. First, we introduce **TIMELI**, an extensive video-based, TIME-aware Language Instruction benchmark designed to accurately evaluate safety-critical temporal reasoning in multimodal AI systems for blind users. TIMELI includes both synthetic and real-world data, and was developed in collaboration with mobility guides and blind users to model expert-level guidance with context-aware instructions, including directing users to walk alongside buildings, avoid cane-length obstacles, and remain silent at intersections to ensure instructions are situationally appropriate, safe, and timely. Second, we evaluate several models on TIMELI and demonstrate their failure in both open-loop (video-based) and closed-loop (interactive) settings, even after fine-tuning on synthetic data to reduce the domain gap.

TABLE I: **Comparing Vision-and-Language Navigation Datasets.** Compared to prior datasets, our benchmark includes rich contextual instructions and temporal information, specifying when and what to say to guide a visually impaired user. Additionally, we emphasize the involvement of dynamic obstacles (*e.g.*, pedestrians) in outdoor navigation scenarios. For comparison with other non-temporal benchmarks, we also note the size of each dataset (total number of images in TIMELI is 1,233,616). While the synthetic benchmark is procedurally generated to incorporate diverse factors (e.g., weather, time of day, safety hazard), we further curate a real-world dataset sourced from an in-situ study and manually annotated YouTube videos, reserved for validation.

Dataset	Dynamic	Context	Temporal Information	For Blind End-User	Size	Collection
R2R [62]	✗	Indoor	✗	✗	21,567 Images	Real-World
CVDN [70]	✗	Indoor	✗	✗	7,000 Images	Real-World
REVERIE [71]	✗	Indoor	✗	✗	21,702 Images	Real-World
Touchdown [64]	✓	Outdoor	✗	✗	9326 Images	Real-World
Talk the Walk [72]	✓	Outdoor	✗	✗	10,000 Images	Real-World
RxR [51]	✗	Indoor	✗	✗	126,069 Images	Real-World
WAY [73]	✗	Indoor	✗	✗	6,154 Images	Real-World
UrbanWalk [5]	✓	Outdoor	✗	✓	399,126 Images	Synthetic
TIMELI - Sim	✓	Outdoor	✓	✓	67,101 Videos	Synthetic
TIMELI - Real	✓	Outdoor	✓	✓	2,000 Videos	Real-World

II. Related Work

Video and Temporal Understanding: Our assistive navigation task requires precise modeling of spatio-temporal relationships in video data. Recently, image encoders such as CLIP [33] and ViT [34] have been adapted to encode video data [18], [22], [24], [35]–[40], enabling advancements in tasks such as video summarization [17], [19], [23], [25], [40]–[42] and turn-by-turn question answering [20], [21], [43], [44]. However, visual understanding capabilities of state-of-the-art open-source and closed-source models [7], [45] remain limited [36], [46], [47], raising concerns about their usability as interactive contexts. Moreover, prior benchmarks for question answering, *e.g.*, VizWiz [48]–[50] (often used in MLLM evaluations [25]), do not generally consider the issue of planning for *when to cue* for a response, yet both how and when to realize a cue can be tightly integrated in our context. Similarly, while recent work in generalized time-aware instruction tuning [20] provides an initial step in reasoning over temporal observations, current models cannot capture interactions between fine-grained landmarks, 3D spatial planning, and safety and conciseness constraints within dynamic, real-time contexts. Our analysis reveals that existing models do not generalize to such instruction tasks.

User-Aware Navigation Instruction Generation: In contrast to the standard Vision-and-Language Navigation (VLN) task [27], [51]–[64], which involves interpreting instructions into low-level navigation actions, we are concerned with instruction generation to a user. Within our task, most prior work considers static (*e.g.*, on R2R [62]) and simplified settings [5], [26], [29]–[32], [65]–[67], without video reasoning, prior interaction history, instructional timing, or real-time user state. Similarly, step-by-step commercially-available services, *e.g.*, based on Google Maps, are limited as they tend to generate instructions based on various planning and location heuristics [68], [69], thus not considering issues with GPS error, user preferences, or accessibility needs. To re-iterate, prior approaches and models lack the ability to determine optimal timing for instructions, even though timing and content are inherently linked. Instead, an instruction may always be sampled from such models for any instantaneous state, even in cases where it may not be needed or could be dangerous, *e.g.*, during crossings [10], [11]. In this work, our task requires the model to sequentially decide what to say and when, which involves an intricate planning task over the high-level plan, environmental hazards, and overall task and path adherence.

Designing Assistive Systems: Our ultimate goal is to develop a system that optimally assists users by dynamically generating navigation guidance instructions. Notably, the design of accessible navigation systems remains an ongoing area of research [74]–[76]. Moreover, appropriate timing of instructions is as crucial as their content to avoid overwhelming the user's cognitive load, especially given that the required information can change drastically depending on time and situation in diverse real-world environments [11], [77]–[82]. While various practical systems have been developed, using modalities like vibration [77], [83], [84] and sonification [85]–[87], audio-based language remains a commonly preferred approach by users [69], [88]–[90]. However, these audio cues are typically manually crafted, limiting adaptability to new environments or situations. Remote sighted human assistance [11], [78], [91]–[93], in which a sighted assistant conveys instructions remotely, has gained adoption and offers flexible guidance, but its usage is costly. Several preliminary and small-scale approaches based on MLLMs have been recently developed [47], [94]–[96], while relying on rule-based timing for instructions [94] or focus on basic captions rather than active goal-driven navigation [47], [93], [95], [97]. We referred to user responses and failure scenarios in these prior studies, as well as semi-structured interviews with guides and blind users when developing our novel

simulation and benchmark. A key finding is that effective instruction can be nuanced and contextual, even around basic turns, *e.g.*, *"find the building, you are inside a recessed doorway, step back, now turn left, trace the building, careful there's an overhead branch."* Another common principle is to remain silent at intersections, allowing the navigator to rely on mobility skills learned through orientation and mobility training. Here, unnecessary instructions could pose a danger [11] (our benchmark directly quantifies performance in this task). Finally, users prefer avoiding excessive amounts of information (a known issue in LLMs [8], [9]) and instead prioritizing certain information at the right time, *e.g.*, obstacles in cane length [74]. Thus, while focusing on accessibility, our study provides a benchmark for social machine intelligence and a step toward generalized assistive agents that can holistically understand user needs.

III. Method

The goal of our benchmark is to evaluate how well MLLMs can generate time-sensitive and situationally appropriate navigation instructions for blind users, with an overview of model components shown in Fig. 2. In our task, we assume a scenario where a blind person navigates using an assistive system in an outdoor setting, equipped either on a smartphone [69], [98] or a wearable device [89], with an RGB camera to perceive the front field of view. The system sequentially observes the front view and issues guidance instructions to the user. Following common systems that utilize localization information (*e.g.*, GPS) and destination information, we also assume access to a noisy location and a high-level path planner to the destination [99]. We formalize our task in Sec. III-A, followed by a description of our benchmark creation (Sec. III-B) and network optimization process (Sec. III-C).

A. Task Formulation

Time-aware instruction generation is a sequential decision-making task, where the goal at each time step t is to generate an *output instruction* $\mathbf{w}_t^{\text{out}} = (w_t^1, \ldots, w_t^{N_w}) \in \mathcal{W}$, consisting of N_w word tokens, *i.e.*, $w_t \in [1, D]$, D being the vocabulary size. At inference time, we assume access to a stream of *observations* $\mathbf{o}_t = (\mathbf{V}_t, \mathbf{w}_t^{\text{in}}) \in \mathcal{O}$ corresponding to: (1) an ego-centered *video* $\mathbf{V_t} \in \mathbb{R}^{W \times H \times 3 \times T}$ of resolution $W \times H$, *i.e.*, including the current image and past $T-1$ frames ($T = 1$ for image-based models); and (2) a language-based input to the model $\mathbf{w}^{\text{in}} \in \mathcal{W}$, comprising *user and task specifications*, a high-level navigation *plan* toward a goal represented as a command and a sequence of 2D waypoints [100], [101]. We note that the high-level plan only provides coarse guidance for the final instructions, as it is based on a static, standard-definition map. We also leverage the *history of prior instructions*, as further detailed below.

Task Specification and History: Our task includes detailed specifications of the user and task:

> *"You are guiding a blind user. You will need to instruct the user to stay on the path to the goal. Notify them of immediate turns and obstacles within 1.5m to avoid. Keep instructions at junctions minimal, for example..."*

The task specification also includes additional relevant context, specifically the goal: *"The blind person needs to approach a relative goal: [x,y] =* $[x_{dest}, y_{dest}]$*,"* where $[x_{dest}, y_{dest}] \in \mathbb{R}^2$ is the next waypoint. Additionally, a text-based command is used to specify the high-level navigation plan, derived from the next goal waypoint, *e.g.*, *"turn left"*. The plan (out of seven possibilities) is computed based on the angle between the goal coordinate and the agent's heading direction of the user. Finally, the task specification input also incorporates a history of prior instructions over a time window of length T, along with their associated timestamps, *e.g.*, *"2 seconds ago, the instruction was: pedestrian on your left. 5 seconds ago, the instruction was: go forward..."* Crucially, the task specification includes an explicit instruction requiring the model to output *"None"* when no user feedback is necessary or contextually appropriate. This ensures that models learn to remain silent when redundant or distracting guidance would be unhelpful.

B. The TIMELI Benchmark

We introduce *TIMELI*, the first video-based time-aware benchmark designed for navigating pedestrians in dynamic urban environments. Despite a large number of vision-and-language navigation benchmarks, there is currently no *video-based* dataset or environment for training and evaluating models that deliver timely and safety-aware navigation instructions to humans. Moreover, in interactive domains where collection and annotation of real-world human data can be difficult, realistic synthetic environments offer a scalable, privacy-preserving, and reproducible solution.

Design Through Collaboration with Mobility Guides: In conjunction with guidance from existing literature, we conducted a pilot study with ten blind participants and local Orientation and Mobility guides to explore remote navigation guidance. Participants wore chest-mounted cameras [92], while guides gave directions via Zoom using Google Maps. With IRB approval, we recorded video and audio for analysis. From observations and feedback, we found that frequent, brief cues help maintain orientation, especially in complex environments. Guides confirmed obstacle-free paths, referenced nearby walls for alignment, and used simple directions like *"left"*, or *"right"* or clock-face terms. Numeric distances were avoided in favor of natural phrases like "keep walking straight." These insights informed the TIMELI benchmark's instruction timing and phrasing.

Multimodal Temporal Data Collection: As real-world data collection is inherently constrained by safety concerns, data scarcity, and privacy issues, synthetic data has emerged as a promising solution, enabling the creation of artificial data that mimics real-world patterns [102]–[104]. We leverage the

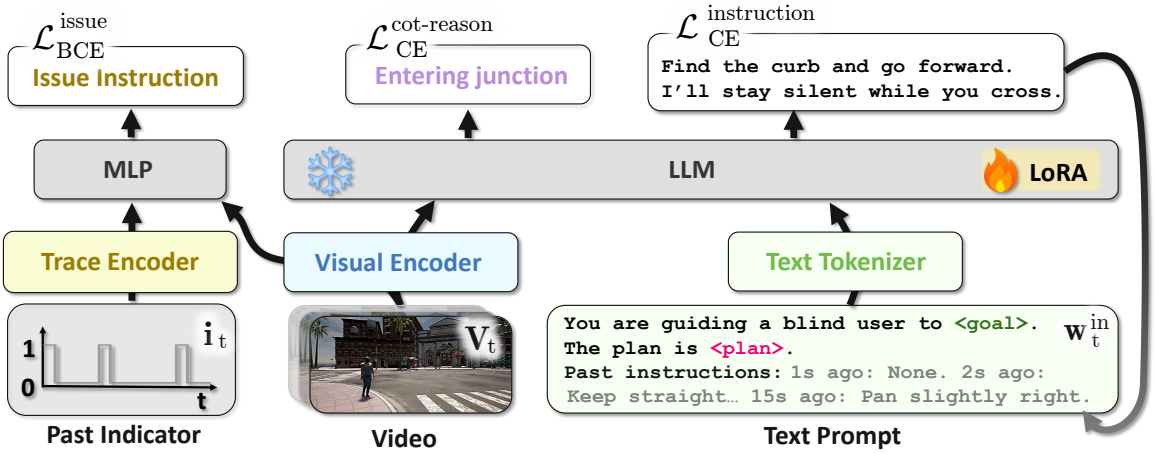

Fig. 2: **Model Structure and Task Overview.** The model takes input video, plan (defined from a set of map-based waypoints), and task specifications with historical instructions (Sec. III-A). We explicitly supervise for instructional reason (Sec III-C), which we find improves model guidance performance.

open-source simulation environment based on CARLA [99], [105] to collect multimodal data for the pedestrian navigation task. While CARLA is mainly used for autonomous driving, we adapt it to simulate first-person pedestrian data under varied conditions. Pedestrians are randomly spawned on sidewalks with random destinations, following A* paths that comply with traffic rules. As pedestrians move, we capture multimodal data (location, ego-centric images, depth, and semantic segmentation) at 1 FPS. Intersection locations are extracted due to their navigation complexity. Our data spans six weather conditions and five towns.

Time-Aware Navigation Instruction Generation: Our procedural navigation instruction generation process considers both the content and the timing of the instructions, ensuring clear and essential communication that supports effective navigation without overwhelming the pedestrian. Specifically, the content of instructions encompasses two aspects: directional cues (*e.g.*, *"Now, turn slightly right and continue walking"*) and environmental descriptions (*e.g.*, *"Be careful, there's a pedestrian in front of you"*). Directional cues are generated based on the relative displacement between the pedestrian's current position and the position one second earlier. For example, the system instructs the pedestrian to walk forward when the change is within 30°, uses the term "slightly" for turns under 75°, and provides explicit left or right turn instructions for direction changes exceeding 75°. Repeated alerts for the same obstacle are removed within a specified time frame, except when an obstacle is detected within one meter (cane length). Immediate instructions are also issued when a change of direction is necessary. In complex scenarios such as intersections, we deliver instructions with warnings for upcoming turns or obstacles prior to entering the intersections. Upon entering or exiting intersections, the instruction includes phrases like *"You are entering/crossing the junction"* or *"You have exited the junction."* To validate our simulation data's real-world relevance, we curated an additional test set from YouTube videos. We developed a custom annotation interface where annotators review frames in a video by editing a text box that shows alongside suggested instructions generated by GPT-4o. The interface displays four consecutive frames, with the current frame on the left and the next three seconds to the right or a GIF (1 FPS) for dynamic visualization. Human annotators manually matched scenarios to the user study, followed mobility guides' instructions, and annotated the navigator's direction.

Data Statistics: Overall, we collected 67K navigation videos, where each navigation video contains T frames. Out of the 67K samples, 79.1% of the data contained none (remained silent), and 20.9% had an instruction issued. Among the spoken instructions, 81.0% directed to move forward, 17.6% directed to turn left or turn right, and the rest directed to turn around or to stop. The average length of the spoken instruction was 11.56 words, with a standard deviation of 4.59 words, with a maximum instruction length of 32 words and a minimum of two words (*"Turn around."*). Additionally, 21.6% of the data were recorded at intersections. Examples of the instructions include: "*Turn slightly right and continue walking,*" "*Okay, keep walking straight. Be careful, pedestrian in your front,*" and "*Now, turn slightly right and keep walking. Fence on your right.*"

C. Model Architecture

Our task extends prior MLLM-based agents to real-time settings. Given the challenges associated with our task, we refine the instruction space and introduce efficient and targeted supervision for chain-of-thought (CoT) reasoning. Moreover, to further enhance the instruction generation in real-time scenarios, we incorporate auxiliary supervision for timing.

Reason Prediction: To improve the model's decision-making capabilities, we restructure the output format of $\mathbf{w}^{\text{out}}$ so that the model first predicts the rationale behind an instruction before generating the instruction itself. The reason prediction is explicitly supervised, representing the reasoning behind each guidance instruction and is categorized into one of eight distinct labels:*"remain silent,"*

"remain silent in junction," "enter junction," "exit junction," "obstacle in front," "constant instruction," "direction change," or *"stop."* This structure is designed not only to enhance the model's ability to reason over instruction realization but also to diagnose predictions. The reason for the ground truth can be computed using the simulation's state.

Timing Supervision: To efficiently issue instructions while facilitating temporal alignment, we introduce additional timing supervision-a binary classifier head determining whether an instruction should be issued. Specifically, we encode instructional cues as a 1D binary vector $\mathbf{i}_t \in \{0,1\}^T$, where each entry indicates whether an instruction was given at a past timestep. This trace vector is encoded via an MLP and combined with visual encoder features to predict whether an instruction should be generated. During inference, the timing prediction can also serve as a gating mechanism [106] for reduced latency: if no instruction is needed according to the timing classifier, the model performs an early exit, bypassing the instruction generation stage and thereby reducing both inference time and computational overhead.

Loss Function: We supervise the model via Cross-Entropy (CE) loss over the generated output:

$$\mathcal{L}_{\text{CE}} = \exp\left[-\frac{1}{N_w}\sum_{n=1}^{N_w}\log P(w^n \mid w^1,\ldots,w^{n-1})\right] \quad (1)$$

where N_w is the total output length. We note that the generated output contains both reason supervision (which is not announced to the user) and output instruction. The overall loss is then:

$$\mathcal{L}_{\text{out}} = \mathcal{L}_{\text{CE}}^{\text{cot-reason}} + \mathcal{L}_{\text{CE}}^{\text{instruction}} + \beta\mathcal{L}_{\text{BCE}}^{\text{issue}} \quad (2)$$

where we add the reason prediction and issuance classification (Binary Cross Entropy loss) as auxiliary tasks, and a constant hyperparameter β.

IV. Experiments

In this section, we first discuss our experimental setup, including the models and evaluation metrics. Next, we discuss our results with off-the-shelf and fine-tuned models in open and closed-loop settings.

A. Experimental Setup

Models: We build on several state-of-the-art MLLM, such as ones take an image [7], [27], [107], [108] or video [7], [20], [25], [60] as input. We split our dataset into training (Town05 and Town10 in CARLA) and testing (Town02, Town03, and Town04). Models were initiated from an open-source checkpoint with four A6000 GPUs. For MLLMs with an image input, we provide the current frame of the video.

Metrics: We evaluate our model's performance based on timing, navigation instruction, and navigation performance. For timing, we report the F1 score and AUC. Timing metrics were calculated with a one-second timing tolerance, where predictions within one second of the ground truth are considered correct. We then employ standard language metrics for instruction evaluation, including BLEU-4 [109] and ROUGE-L [110], computed only when the predicted timing falls within this tolerance as the model learns to omit instructions when not necessary. We also report conciseness by identifying the number of overlapping words between the prediction and the ground truth after removing stopwords from both, then dividing this value by the prediction's word count without stopwords. We leverage CARLA's interactive environment for closed-loop evaluation, measuring general VLN metrics [32], including success rate, route completion, navigation score, collisions per minute, and instructions per minute to understand the models' guidance performance. Finally, we define a Navigation Quality Score (NQS) metric:

$$\text{NQS} = (SR \cdot RC \cdot NS \cdot S_C \cdot S_I)^{1/5} \quad (3)$$

where SR, RC, and NS denote success rate, route completion, and navigation score. The collision penalty and instruction efficiency are defined as

$$S_C = \exp(-C/4), \quad S_I = \exp\left(-\frac{(I-I^*)^2}{2\sigma^2}\right) \quad (4)$$

where C is collisions per minute, I is instructions per minute, $I^* = 20$ is the optimal instruction rate, and $\sigma = 10$ controls the acceptable deviation.

TABLE II: **Open-Loop Assistive Instruction Generation Using Off-the-Shelf Models and Finetuned Models.** We sequentially provide models with images and videos in the TIMELI benchmark to evaluate their safety-critical navigation performance. The output of the off-the-shelf models is guided via in-context learning. While all models have access to goal coordinates, the finetuned model denoted by "+" incorporates instruction history, high-level plan, and reason, while those denoted by "++" additionally include timing supervision.

Model	Size	Modality	BLEU-4	ROUGE-L	Timing F1	Timing AUC	Conciseness
Off-the-Shelf Models							
VILA [25]	3B	Image	0.626	6.545	0.579	0.778	0.554
TinyLLaVA [107]	3B	Image	0.000	5.728	0.361	0.500	0.264
VILA [25]	3B	Video	0.000	4.377	0.465	0.651	0.552
LLaVA-v1.6 [108]	7B	Image	1.849	7.851	0.575	0.776	0.068
NaVid [60]	7B	Video	0.000	6.659	0.384	0.606	0.046
TimeChat [20]	7B	Video	0.000	7.423	0.379	0.536	0.026
VILA [25]	8B	Image	0.201	3.916	0.566	0. 788	0.010
LLaVA-v1.6 [108]	13B	Image	3.205	12.991	0.499	0.726	0.116
NavGPT [27]	-	Image	2.536	12.998	0.042	0.508	0.167
GPT-4o [7]	-	Image	7.395	20.059	0.451	0.654	0.220
GPT-4o [7]	-	Video	2.536	12.998	0.042	0.508	0.245
Finetuned Models on TIMELI Benchmark							
TinyLLaVA	3B	Image	0.000	0.000	0.000	0.500	0.000
TinyLLaVA+	3B	Image	13.294	27.730	0.254	0.573	0.325
VILA	3B	Video	8.618	20.063	0.540	0.726	0.048
VILA+	3B	Video	10.916	23.789	0.410	0.629	0.300
TimeChat	7B	Video	0.000	6.788	0.030	0.508	0.067
TimeChat+	7B	Video	11.891	29.127	0.101	0.525	0.316
LLaVA-v1.6	7B	Image	12.398	24.765	0.572	0.749	0.064
LLaVA-v1.6+	7B	Image	20.390	38.250	0.776	0.869	0.395
LLaVA-v1.6++	7B	Image	19.660	37.709	0.792	0.879	0.219

B. Results

Off-the-Shelf Model Failure and Role of Domain Finetuning: Table II presents the comparison between off-the-shelf models and TIMELI fine-tuned models. The off-the-shelf models were guided with an in-context learning method, which the prompt contains example inputs and outputs. We

TABLE III: **Zero-Shot Transfer from Simulation to Real-World.** We evaluate models trained on the TIMELI benchmark in open-loop settings on real-world videos. We find models transfer knowledge from our extensive simulation benchmark to real-world scenarios without any further fine-tuning.

Model	Size	Modality	BLEU-4	ROUGE-L	Timing F1	Timing AUC	Conciseness
TinyLLaVA+	3B	Image	0.065	1.592	0.454	0.500	0.019
VILA+	3B	Video	7.005	20.359	0.186	0.545	0.253
TimeChat+	7B	Video	0.001	22.995	0.025	0.506	0.383
LLaVA-v1.6+	7B	Image	8.478	22.587	0.732	0.815	0.278
LLaVA-v1.6++	7B	Image	9.479	23.970	0.672	0.771	0.266

TABLE IV: **Closed-Loop Performance Using Finetuned Models.** We perform a study where pedestrians are controlled based on instructions generated from models to walk along 12 routes in the simulation environment (see metric definition in Sec. IV-A). Note that some models may achieve low errors due to minimal progress along the route.

Model	Size	Modality	NQS	Success Rate	Route Completion	Navigation Score	Collision/Min	Instructions/Min
TinyLLaVA	3B	Image	0.00	0.00	0.01	0.01	0.00	0.00
TinyLLaVA+	3B	Image	0.81	0.47	0.93	0.92	0.51	18.91
VILA	3B	Video	0.29	0.02	0.55	0.44	0.68	8.66
VILA+	3B	Video	0.57	0.14	0.79	0.63	0.60	16.35
TimeChat	7B	Video	0.24	0.04	0.45	0.46	0.41	2.95
TimeChat+	7B	Video	0.38	0.17	0.68	0.54	0.51	3.66
LLaVA-v1.6	7B	Image	0.31	0.04	0.36	0.29	0.54	12.65
LLaVA-v1.6+	7B	Image	0.79	0.44	0.93	0.92	0.69	19.57
LLaVA-v1.6++	7B	Image	0.82	0.50	0.92	0.91	0.60	21.22

find current off-the-self models to perform poorly on our task, as indicated by the low BLEU-4 scores, and fail to provide instructions at the appropriate times, with a Timing F1 score near or below 0.50. In our prompt and output ablations for the fine-tuned models, we observe a general trend where instruction history, high-level plan, and reason prediction output complement to improve model performance. Also, the model trained along with timing prediction increased their performance on the timing metric. Interestingly, image-based models outperform their video-based counterparts. We explain this due to the high task complexity and input dimensionality; we find video models to struggle in compressing long-term information and learning to attend to correct video elements. This finding highlights a future direction in more effective video modeling techniques beyond basic video descriptions.

Real-World Transfer Evaluation: To evaluate sim-to-real transferability, we validated our models using annotated real-world videos (Table III). Interestingly, we find some models fine-tuned on purely synthetic data to transfer to real-world settings. While we observed performance degradation compared to the simulation environment, the degradation in image models was not severe, whereas video models showed a more significant drop in performance. We find that current video MLLMs inefficiently handle longer inputs (consistent with findings in other domains [111]).

Closed-Loop Evaluation in CARLA: Table IV shows the result of the interactive closed-loop evaluation in 12 routes in the CARLA environment. We control the pedestrian by mapping the model-generated instructions to actions using a rule-based function that determines pedestrian movement. To simulate realistic sensor measurements, we introduce Gaussian noise to the goal coordinates and pedestrian orientation. To simulate real-world uncertainty, we inject Gaussian localization noise (mean of 0 m and SD of 2.0 m) into the planner, and orientation noise (mean of 0° and SD of 5°). Our LLaVA-v1.6++ image-based model achieves a success rate of up to 50% in guiding users to the goal without repetitive instructions. We also observe that success in language modeling metrics in open-loop does not always lead to better performance in closed-loop evaluation for some of the models, which is consistent with prior findings in embodied navigation [67].

V. Conclusion

An overarching goal in interactive vision-based systems is to develop adaptable models that do not require extensive fine-tuning or annotations across different settings. However, our findings reveal that current MLLMs struggle with fundamental concepts of assistive navigation, often failing to determine both when and what to instruct. These models tend to prioritize content generation while neglecting crucial temporal and contextual factors necessary for safe and effective assistance. While our model shows potential in supporting visually impaired individuals by automating navigation tasks, we acknowledge the inherent risks of AI-generated instructions. Misleading or poorly timed guidance could lead to hazardous situations, making reliability and safety paramount. We aim to enhance MLLM performance through improved reasoning and rigorous testing in diverse real-world conditions. We hope that our framework and tools will contribute to developing intelligent guidance systems capable of reasoning over diverse user needs, tasks, and environments. While we focus on the smartphone platform due to its scalability and low cost, future evaluations of generalization on other robotic platforms, such as ground robots [112]–[114], will be crucial to validate and broaden the applicability of our approach.

Acknowledgments: We thank the National Science Foundation (IIS-2152077, IIS 2443719) for supporting this research.

References

[1] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *CVPR*, 2018.
[2] L. Fei-Fei and R. Krishna, "Searching for computer vision north stars," *Daedalus*, 2022.
[3] X. Puig, T. Shu, S. Li, Z. Wang, Y.-H. Liao, J. B. Tenenbaum, S. Fidler, and A. Torralba, "Watch-and-help: A challenge for social perception and human-AI collaboration," *ICLR*, 2021.
[4] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun *et al.*, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *CoRL*, 2023.
[5] Z. Huang, Z. Shangguan, J. Zhang, G. Bar, M. Boyd, and E. Ohn-Bar, "ASSISTER: Assistive navigation via conditional instruction generation," in *ECCV*, 2022.
[6] OpenAI, "Be My Eyes uses GPT-4 to transform visual accessibility," https://openai.com/index/be-my-eyes/, 2024.
[7] OpenAI, "Hello GPT-4o," https://openai.com/index/hello-gpt-4o/.
[8] K. Saito, A. Wachi, K. Wataoka, and Y. Akimoto, "Verbosity bias in preference labeling by large language models," *arXiv*, 2023.
[9] H. Zhao, M. Andriushchenko, F. Croce, and N. Flammarion, "Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning," *ICML*, 2024.
[10] N. Aging and D. T. Center, "Can I cross the street? considerations for a blind pedestrian," https://www.nadtc.org/news/blog/can-i-cross-the-street-considerations-for-a-blind-pedestrian/, 2022.

[11] S. Lee, M. Reddie, C.-H. Tsai, J. Beck, M. B. Rosson, and J. M. Carroll, "The emerging professional practice of remote sighted assistance for people with visual impairments," in *CHI*, 2020.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *NeurIPS*, 2022.

[13] K. A. Mack, R. Qadri, R. Denton, S. K. Kane, and C. L. Bennett, ""they only care to show us the wheelchair": Disability representation in text-to-image ai models," in *CHI*, 2024.

[14] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and Fairness in Large Language Models: A Survey," *JCL*, 2024.

[15] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021.

[16] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "VALOR: Vision-audio-language omni-perception pretraining model and dataset," *PAMI*, 2024.

[17] X. Shen, Y. Xiong, C. Zhao, L. Wu, J. Chen, C. Zhu, Z. Liu, F. Xiao, B. Varadarajan, F. Bordes *et al.*, "Longvu: Spatiotemporal adaptive compression for long video-language understanding," *arXiv preprint arXiv:2410.17434*, 2024.

[18] J. Vaishnavi and V. Narmatha, "Video captioning–a survey," *Multimed. Tools Appl.*, 2024.

[19] R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[20] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "TimeChat: A time-sensitive multimodal large language model for long video understanding," in *CVPR*, 2024.

[21] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang *et al.*, "Moviechat: From dense token to sparse memory for long video understanding," in *CVPR*, 2024.

[22] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," in *EMNLP: Demo*, 2023.

[23] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *EMNLP*, 2024.

[24] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Tubedetr: Spatio-temporal video grounding with transformers," in *CVPR*, 2022.

[25] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," in *CVPR*, 2024.

[26] X. Wang, W. Wang, J. Shao, and Y. Yang, "Learning to follow and generate instructions for language-capable navigation," *PAMI*, 2023.

[27] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," in *AAAI*, 2024.

[28] X. Wang, W. Wang, J. Shao, and Y. Yang, "Lana: A language-capable navigator for instruction following and generation," in *CVPR*, 2023.

[29] H. Zeng, X. Wang, W. Wang, and Y. Yang, "Kefa: A knowledge enhanced and fine-grained aligned speaker for navigation instruction generation," *arXiv preprint arXiv:2307.13368*, 2023.

[30] S. Fan, R. Liu, W. Wang, and Y. Yang, "Navigation instruction generation with bev perception and large language models," in *ECCV*, 2024.

[31] X. Kong, J. Chen, W. Wang, H. Su, X. Hu, Y. Yang, and S. Liu, "Controllable navigation instruction generation with chain of thought prompting," *arXiv*, 2024.

[32] S. Wang, C. Montgomery, J. Orbay, V. Birodkar, A. Faust, I. Gur, N. Jaques, A. Waters, J. Baldridge, and P. Anderson, "Less is more: Generating grounded navigation instructions from landmarks," in *CVPR*, 2022.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[35] M. U. Khattak, M. F. Naeem, J. Hassan, M. Naseer, F. Tombari, F. S. Khan, and S. Khan, "Complex video reasoning and robustness evaluation suite for video-lmms," *arXiv preprint arXiv:2405.03690*, 2024.

[36] M. Cai, R. Tan, J. Zhang, B. Zou, K. Zhang, F. Yao, F. Zhu, J. Gu, Y. Zhong, Y. Shang *et al.*, "Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models," *arXiv preprint arXiv:2410.10818*, 2024.

[37] H. Wu, D. Li, B. Chen, and J. Li, "Longvideobench: A benchmark for long-context interleaved video-language understanding," *NeurIPS*, 2024.

[38] D. Li, Y. Liu, H. Wu, Y. Wang, Z. Shen, B. Qu, X. Niu, G. Wang, B. Chen, and J. Li, "Aria: An open multimodal native mixture-of-experts model," *arXiv preprint arXiv:2410.05993*, 2024.

[39] Y. Zeng, H. Zhang, J. Zheng, J. Xia, G. Wei, Y. Wei, Y. Zhang, T. Kong, and R. Song, "What matters in training a gpt4-style language model with multimodal inputs?" in *ACL*, 2024.

[40] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv*, 2024.

[41] G. Chen, Y.-D. Zheng, J. Wang, J. Xu, Y. Huang, J. Pan, Y. Wang, Y. Wang, Y. Qiao, T. Lu *et al.*, "Videollm: Modeling video sequence with large language models," *arXiv preprint arXiv:2305.13292*, 2023.

[42] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *CVPR*, 2023.

[43] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[44] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[45] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[46] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in *CVPR*, 2024.

[47] Y. Kaniwa, M. Kuribayashi, S. Kayukawa, D. Sato, H. Takagi, C. Asakawa, and S. Morishima, "Chitchatguide: Conversational interaction using large language models for assisting people with visual impairments to explore a shopping mall," *PACMHCI*, 2024.

[48] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. P. Bigham, "Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people," in *CVPR*, 2019.

[49] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White *et al.*, "Vizwiz: nearly real-time answers to visual questions," in *UIST*, 2010.

[50] Y.-Y. Tseng, A. Bell, and D. Gurari, "Vizwiz-fewshot: Locating objects in images taken by people with visual impairments," in *ECCV*, 2022.

[51] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," *EMNLP*, 2020.

[52] Y. Hong, C. Rodriguez-Opazo, Q. Wu, and S. Gould, "Sub-instruction aware vision-and-language navigation," in *EMNLP*, 2020.

[53] Y. Qi, Z. Pan, S. Zhang, A. van den Hengel, and Q. Wu, "Object-and-action aware model for visual language navigation," in *ECCV*, 2020.

[54] R. Liu, X. Wang, W. Wang, and Y. Yang, "Bird's-eye-view scene graph for vision-language navigation," in *ICCV*, 2023.

[55] Y. Zhang, Z. Ma, J. Li, Y. Qiao, Z. Wang, J. Chai, Q. Wu, M. Bansal, and P. Kordjamshidi, "Vision-and-language navigation today and tomorrow: A survey in the era of foundation models," *arXiv*, 2024.

[56] A. Moudgil, A. Majumdar, H. Agrawal, S. Lee, and D. Batra, "Soat: A scene-and object-aware transformer for vision-and-language navigation," *NeurIPS*, 2021.

[57] Y. Xu, Y. Pan, Z. Liu, and H. Wang, "Flame: Learning to navigate with multimodal llm in urban environments," *arXiv*, 2024.

[58] J. Gu, E. Stefani, Q. Wu, J. Thomason, and X. E. Wang, "Vision-and-language navigation: A survey of tasks, methods, and future directions," *arXiv preprint arXiv:2203.12667*, 2022.

[59] W. Wu, T. Chang, X. Li, Q. Yin, and Y. Hu, "Vision-language navigation: A survey and taxonomy," *Neural Comput. Appl.*, 2023.

[60] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *arXiv*, 2024.

[61] E. Ohn-Bar, K. Kitani, and C. Asakawa, "Personalized dynamics models for adaptive assistive navigation systems," in *CoRL*, 2018.

[62] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018.

[63] P. Anderson, A. Shrivastava, J. Truong, A. Majumdar, D. Parikh, D. Batra, and S. Lee, "Sim-to-real transfer for vision-and-language navigation," in *CoRL*, 2021.

[64] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *CVPR*, 2019.

[65] A. F. Daniele, M. Bansal, and M. R. Walter, "Navigational instruction generation as inverse reinforcement learning with neural machine translation," in *HRI*, 2017.

[66] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," *NeurIPS*, 2018.

[67] M. Zhao, P. Anderson, V. Jain, S. Wang, A. Ku, J. Baldridge, and E. Ie, "On the evaluation of vision-and-language navigation instructions," *ACL*, 2021.

[68] D. Fogli, A. Arenghi, and F. Gentilin, "A universal design approach to wayfinding and navigation," *Multim. Tools Appl.*, 2020.

[69] D. Sato, U. Oh, J. Guerreiro, D. Ahmetovic, K. Naito, H. Takagi, K. M. Kitani, and C. Asakawa, "Navcog3 in the wild: Large-scale blind indoor navigation assistant with semantic features," *TACCESS*, 2019.

[70] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *CoRL*, 2020.

[71] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel, "Reverie: Remote embodied visual referring expression in real indoor environments," in *CVPR*, 2020.

[72] H. De Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, and D. Kiela, "Talk the walk: Navigating new york city through grounded dialogue," *arXiv preprint arXiv:1807.03367*, 2018.

[73] M. Hahn, J. Krantz, D. Batra, D. Parikh, J. Rehg, S. Lee, and P. Anderson, "Where are you? localization from embodied dialog," in *EMNLP*, 2020.

[74] K. M. Hoogsteen, S. Szpiro, G. Kreiman, and E. Peli, "Beyond the cane: Describing urban scenes to blind people for mobility tasks," *TACCESS*, 2022.

[75] K. Müller, C. Engel, C. Loitsch, R. Stiefelhagen, and G. Weber, "Traveling more independently: A study on the diverse needs and challenges of people with visual or mobility impairments in unfamiliar indoor environments," *TACCESS*, 2022.

[76] H. J. Kim, K. Sengupta, M. Kuribayashi, H. Kacorri, and E. Ohn-Bar, "Text to blind motion," *NeurIPS*, 2024.

[77] M. Martinez, A. Constantinescu, B. Schauerte, D. Koester, and R. Stiefelhagen, "Cognitive evaluation of haptic and audio feedback in short range navigation tasks," in *ICCHP*, 2014.

[78] S. Lee, R. Yu, J. Xie, S. M. Billah, and J. M. Carroll, "Opportunities for human-AI collaboration in remote sighted assistance," in *IUI*, 2022.

[79] J. Guerreiro, E. Ohn-Bar, D. Ahmetovic, K. Kitani, and C. Asakawa, "How context and user behavior affect indoor navigation assistance for blind people," in *W4A*, 2018.

[80] D. Ahmetovic, J. Guerreiro, E. Ohn-Bar, K. M. Kitani, and C. Asakawa, "Impact of expertise on interaction preferences for navigation assistance of visually impaired individuals," in *W4A*, 2019.

[81] E. Ohn-Bar, J. Guerreiro, D. Ahmetovic, K. Kitani, and C. Asakawa, "Modeling expertise in assistive navigation interfaces for blind people," in *IUI*, 2018.

[82] H. Kacorri, E. Ohn-Bar, K. Kitani, and C. Asakawa, "Environmental factors in indoor navigation based on real-world trajectories of blind users," in *CHI*, 2018.

[83] M. Kuribayashi, S. Kayukawa, H. Takagi, C. Asakawa, and S. Morishima, "Linechaser: A smartphone-based navigation system for blind people to stand in lines," in *CHI*, 2021.

[84] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *TNSRE*, 2018.

[85] G. Presti *et al.*, "Watchout: Obstacle sonification for people with visual impairment or blindness," in *ASSETS*, 2019.

[86] D. Ahmetovic, F. Avanzini, A. Baratè, C. Bernareggi, G. Galimberti, L. A. Ludovico, S. Mascetti, and G. Presti, "Sonification of rotation instructions to support navigation of people with visual impairment," in *PerCom*, 2019.

[87] C. Yoon *et al.*, "Leveraging augmented reality to create apps for people with visual disabilities: A case study in indoor navigation," in *ASSETS*, 2019.

[88] A. Arditi and Y. Tian, "User interface preferences in the design of a camera-based navigation and wayfinding aid," *JVIB*, 2013.

[89] B. Li, J. P. Munoz, X. Rong, J. Xiao, Y. Tian, and A. Arditi, "Isana: Wearable context-aware indoor assistive navigation with obstacle avoidance for the blind," in *ECCV*, 2016.

[90] S. Kayukawa, D. Sato, M. Murata, T. Ishihara, A. Kosugi, H. Takagi, S. Morishima, and C. Asakawa, "How users, facility managers, and bystanders perceive and accept a navigation robot for visually impaired people in public buildings," in *ROMAN*, 2022.

[91] E. Ohn-Bar, R. Zhu, J. Zhang, and L. Zhang, "Navigating the challenges of remotely supporting blind riders in ridesharing," *IJHCS*, 2025.

[92] R. Kamikubo, N. Kato, K. Higuchi, R. Yonetani, and Y. Sato, "Support strategies for remote guides in assisting people with visual impairments for effective indoor navigation," in *CHI*, 2020.

[93] R.-C. Chang, Y. Liu, and A. Guo, "Worldscribe: Towards context-aware live visual descriptions," in *UIST*, 2024.

[94] H. Wang, J. Qin, A. Bastola, X. Chen, J. Suchanek, Z. Gong, and A. Razi, "Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation," *arXiv preprint arXiv:2403.12415*, 2024.

[95] B. Yang, L. He, K. Liu, and Z. Yan, "Viassist: Adapting multi-modal large language models for users with visual impairments," *arXiv preprint arXiv:2404.02508*, 2024.

[96] S. Song, S. Kodagoda, A. Gunatilake, M. G. Carmichael, K. Thiyagarajan, and J. Martin, "Guide-LLM: An embodied llm agent and text-based topological map for robotic guidance of people with visual impairments," *arXiv preprint arXiv:2410.20666*, 2024.

[97] H. Zhang, N. J. Falletta, J. Xie, R. Yu, S. Lee, S. M. Billah, and J. M. Carroll, "Enhancing the travel experience for people with visual impairments through multimodal interaction: Navigpt, a real-time ai-driven mobile navigation system," *arXiv*, 2024.

[98] H. Kacorri, S. Mascetti, A. Gerino, D. Ahmetovic, H. Takagi, and C. Asakawa, "Supporting orientation of people with visual impairment: Analysis of large scale usage data," in *ASSETS*, 2016.

[99] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *CoRL*, 2017.

[100] J. Zhang, R. Zhu, and E. Ohn-Bar, "SelfD: Self-learning large-scale driving policies from the web," in *CVPR*, 2022.

[101] J. Zhang, Z. Huang, A. Ray, and E. Ohn-Bar, "Feedback-guided autonomous driving," in *CVPR*, 2024.

[102] Y.-T. Hu, J. Wang, R. A. Yeh, and A. G. Schwing, "Sail-vos 3D: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data," in *CVPR*, 2021.

[103] M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, and R. Cucchiara, "Motsynth: How can synthetic data help pedestrian detection and tracking?" in *ICCV*, 2021.

[104] Z. Cai, M. Zhang, J. Ren, C. Wei, D. Ren, Z. Lin, H. Zhao, L. Yang, C. C. Loy, and Z. Liu, "Playing for 3D human recovery," *PAMI*, 2024.

[105] J. Zhang, M. Zheng, M. Boyd, and E. Ohn-Bar, "X-world: Accessibility, vision, and autonomy meet," in *ICCV*, 2021.

[106] K. Sengupta, Z. Shangguan, S. Bharadwaj, S. Arora, E. Ohn-Bar, and R. Mancuso, "UniLCD: Unified local-cloud decision-making via reinforcement learning," *ECCV*, 2024.

[107] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, and L. Huang, "TinyLLaVA: A framework of small-scale large multimodal models," *arXiv preprint arXiv:2402.14289*, 2024.

[108] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024.

[109] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[110] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004.

[111] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *T-ACL*, 2024.

[112] D. DeFazio, E. Hirota, and S. Zhang, "Seeing-eye quadruped navigation with force responsive locomotion control," in *CoRL*, 2023.

[113] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *ICRA*, 2017.

[114] H. Zhang and C. Ye, "A visual positioning system for indoor blind navigation," in *ICRA*, 2020.