A Multi-Modal Dataset for Urban Navigation by Blind Individuals

Hee Jae Kim[1] **Kathakoli Sengupta**[1] **Masaki Kuribayashi**[2]
Hernisa Kacorri[3] **Eshed Ohn-Bar**[1]
[1]Boston University [2]Waseda University [3]University of Maryland, College Park

Abstract

People who are blind perceive the world differently than those who are sighted, which can result in distinct motion characteristics. For instance, when crossing at an intersection, blind individuals may have different patterns of movement, such as veering more from a straight path or using touch-based exploration around curbs and obstacles. These behaviors may appear less predictable to motion models embedded in autonomous vehicles. Yet, the ability of 3D motion models to capture such behavior has not been previously studied, as existing datasets for 3D human motion currently lack diversity and are biased toward people who are sighted. In this work, we introduce BlindWays, the first multimodal motion benchmark for pedestrians who are blind. We collect 3D motion data using wearable sensors with 11 blind participants navigating eight different routes in a real-world urban setting. Additionally, we provide rich textual descriptions that capture the distinctive movement characteristics of blind pedestrians and their interactions with both the navigation aid (*e.g.*, a white cane or a guide dog) and the environment. We benchmark state-of-the-art 3D human prediction models, finding poor performance with off-the-shelf and pre-training-based methods for our novel task. To contribute toward safer and more reliable autonomous systems that reason over diverse human movements in their environments, we will publicly release our novel text-and-motion benchmark.

1 Introduction

A blind pedestrian may not look forward to signal intent to cross before stepping into the road, and may take longer to explore tactile cues when crossing in various intersections [2, 1, 5, 6, 4]. Blind pedestrians may also significantly veer in open spaces, and unexpectedly step into the road due to a truck parked in obstructed intersections with damaged or ambiguous curbs. In such scenarios, reasoning over subtle 3D behaviors, *e.g.*, hand-aid coordination gestures, could improve future prediction in autonomous vehicles and avoid potential safety-critical outcomes. Yet, as far as we are aware, no prior work has investigated predicting pedestrian motion in such edge cases and their inherently distinct, subtle, and uncertain nature. This is a critical issue in autonomous driving and urban accessibility which we hope to discuss in UrbanAccess24, which includes in its scope AI-based mobility tools and techniques for autonomous vehicles, as well as "examine the emerging role of AI in the design of equitable and accessible cities." In this work, we are interested in understanding the capabilities of state-of-the-art 3D motion models for modeling and predicting future blind motion – ultimately, to ensure that autonomous systems and vehicles in urban environments operate safely around disabled pedestrians. Our overarching goal is to enable more robust, accurate, and needs-aware

[1]Software for Motion Capture: Xsens MVN Animate
[2]Navigation Guidance Map: Google Maps

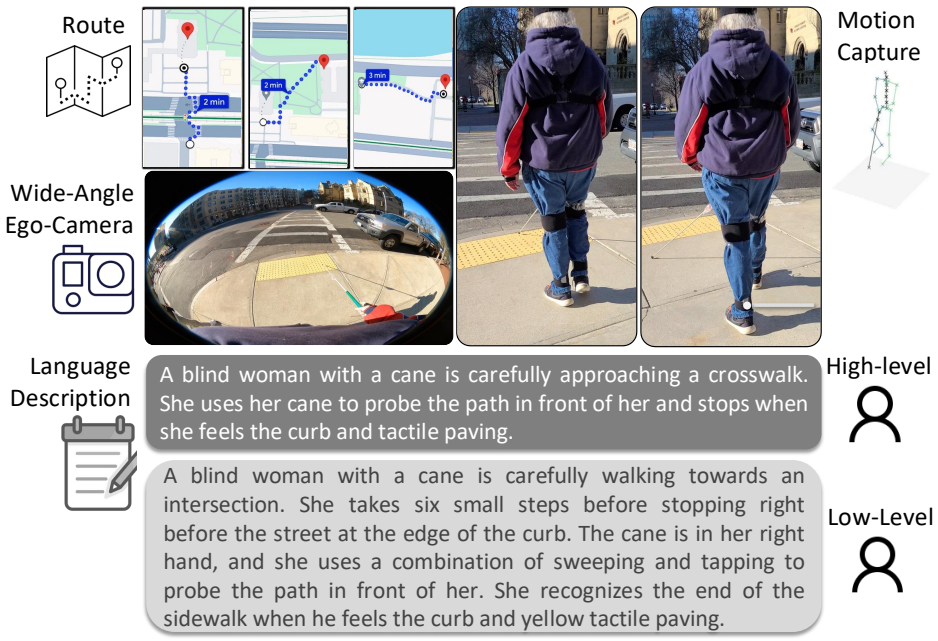

Figure 1: **Data Collection with Wearable IMU-based Sensors.** Depicting a frame from the study with diverse route stimuli, intersections, a motion capture, and a wide-angle egocentric camera view.

pedestrian behavior prediction models that effectively account for disability-related scenarios and behaviors.

2 The BlindWays Dataset

2.1 Overview

We collect BlindWays, a comprehensive blind motion dataset comprising 1,046 clips and 0.6 million frames, along with 2,092 detailed and paired high- and low-level text descriptions. We capture natural motion data from 11 blind and low-vision individuals navigating dynamic, previously unseen outdoor environments along carefully engineered paths exhibiting various challenges. Notably, this is the first work to propose blind motion data enriched with text descriptions, an exceptionally challenging and labor-intensive process. BlindWays's text descriptions are informed by third-person and egocentric videos, each totaling 0.3 million frames. Specifically, captured contextual videos play a critical role in the annotation process by providing an overall scene of blind motion, allowing annotators to sufficiently leverage scene and video context to accurately, precisely, and expressively describe the motion. To synchronize between motion data and videos, we asked participants to clap at the beginning of each route. To ensure high quality, the MoCap system is calibrated in each route, and text descriptions are annotated in-house by human annotators, including motion experts, and carefully checked. We employ a wearable system based on Inertial Measurement Units (IMUs) and filter noisy sequences to maintain accuracy and reliability.

2.2 Data Collection Procedure

We conducted a user study involving 11 participants, consisting of three women and eight men, all of whom are either blind (N=10) or have low vision (N=1). Each participant utilizes their own mobility aid, which includes either a cane or a guide dog, to record natural behavior. Our participants represent a diverse range of ages, levels of visual impairment, and mobility aids, ensuring a rich data collection of navigation behaviors. Participants are equipped with bone conduction headsets to receive real-time auditory instructions from Google Maps. Our data collection was approved by our Institutional Review Board (IRB#6052E). Each participant provided informed consent before

participating in the study and was compensated $50/hour for up to three hours including travel time; data collection sessions typically lasted two hours or less. We note that two researchers always followed the participants during dataset collection to ensure their safety.

Scenarios: In collaboration with local blind advisors and sighted certified orientation and mobility instructors, we engineered eight distinct routes to encompass a variety of real-world scenarios that blind people commonly encounter. These scenarios include walking on the curb, crossing streets, navigating open spaces, and ascending and descending staircases. For example, while crossing streets, participants faced a challenge when encountering a subway track midway, requiring them to stop, reassess, and then continue, which enabled us to capture their behavior while handling sudden stops and changes in terrain. Navigating open spaces presented another challenge due to the lack of obstacles providing environmental cues, forcing participants to rely heavily on auditory instructions. Walking on curbs involved dealing with intermittent obstacles like parked bicycles, trash cans, and overhanging branches. Ascending and descending staircases further added to the complexity, requiring careful coordination and heightened awareness of their immediate environment. Diverse and realistic scenarios enable BlindWays to capture rich and nuanced motion data, reflecting daily real-world challenges and strategies of blind individuals. Each route is carefully mapped and pre-tested to ensure both feasibility and participant safety. At the start of each route, we provided high-level instructions, including specific objectives and expected challenges. For example, we guided participants by informing them of their current location (e.g., surrounding street names) and the direction they were heading to help them better contextualize the audio navigation aid, which usually guides pedestrians by providing street names and directions. We also briefly explained potential obstacles they might encounter, such as a train/tram track in the middle of the route or stairs, to prepare them for critical challenges ahead.

Recording: We employ the Xsens motion capture system, consisting of 18 IMU sensors for body joints and a mobility aid, enabling realistic motion capture in various settings. To comprehensively capture the navigation process, we record third-person video of blind pedestrians and egocentric views, as well as motion data. For egocentric views, participants wear a GoPro HERO10 Black on their chest using a comfortable strap, allowing for hands-free and immersive (GoPro Max Lens Mod) recording. The camera is set to face around the participant's feet to meticulously capture cane movements. For third-person views, the accompanying researchers wear a Samsung Galaxy smartphone around the chest and follow the participants without interrupting their natural movements. All data are synchronized, allowing for an in-depth analysis and annotations of navigation strategies and challenges.

To gain further insights into participants' navigation experiences, upon completion of each route, participants are asked to rate their confidence on a scale of 1-7 in (i) their ability to navigate the route and (ii) the guidance that they received from the Google Maps app.

2.3 Data Annotation Pipeline

To achieve a nuanced understanding of the navigation behaviors of blind individuals, we employ a meticulous annotation pipeline build in-house that leverages the synchronized third-person view RGB videos along the motion data. To ensure privacy, we mosaic the faces of all people appearing in the videos, both the blind participants and passersby. The annotation process involves 15 human annotators, comprising three motion experts (human biomechanics, sensorimotor, and mobility researchers) and 12 novices, who are provided with detailed instructions, exemplars, and feedback.

High-Level Descriptions: For high-level annotations, annotators are requested to focus on describing the overall action of the motion, the purpose behind it, and how the participants were holding their mobility aids (*e.g.*, a cane and a guide dog). Annotators are instructed to provide clear and concise descriptions that convey the intent and broader context of the actions. For example, a high-level description might be: *"A blind man with a cane in his right-hand searches for a street post to press the button. He then orients himself in the direction he wants to cross the street."*

Low-Level Descriptions: Low-level annotations require more detailed descriptions of the motion behavior, such as the number of steps taken and the precise use of mobility aids. For instance, a low-level description might be: *"A blind man with a cane searches and locates a street post. He*

Table 1: **Embedding-Based Evaluation Metrics.** We show embedding-based metrics (based on Jiang *et al.* [7]). For the Diversity metrics, closer to results with Real data (first line) is better. Each experiment is repeated 20 times and a statistical interval with 95% confidence is reported.

Method	Training Set	R Top1 ↑	FID ↓	Diversity →	MModality ↑
Real	-	$0.106^{\pm.0.008}$	$0.257^{\pm.0.018}$	$6.232^{\pm.0.258}$	-
HumanML3D [3]	Motion-X [9]	$0.041^{\pm.0.007}$	$11.203^{\pm.0.109}$	$5.113^{\pm.0.258}$	$3.680^{\pm.0.026}$
MotionGPT [7]	Motion-X [9]	$0.046^{\pm.0.006}$	$15.002^{\pm.0.504}$	$5.871^{\pm.0.234}$	$4.646^{\pm.0.171}$
HumanML3D [3]	BlindWays	$\mathbf{0.060}^{\pm.0.012}$	$\mathbf{3.340}^{\pm.0.257}$	$5.861^{\pm.0.266}$	$1.896^{\pm.0.037}$
MotionGPT [7]	BlindWays	$0.054^{\pm.0.008}$	$5.101^{\pm.0.116}$	$5.098^{\pm.0.148}$	$3.993^{\pm.0.134}$
HumanML3D [3]	Motion-X [9] + BlindWays	$0.054^{\pm.0.009}$	$8.612^{\pm.0.480}$	$\mathbf{6.260}^{\pm.0.301}$	$\mathbf{4.921}^{\pm.0.051}$
MotionGPT [7]	Motion-X [9] + BlindWays	$0.036^{\pm.0.003}$	$10.313^{\pm.0.183}$	$3.874^{\pm.0.164}$	$2.759^{\pm.0.100}$

moves forward three steps to orient himself in the direction he wants to cross the street, using his cane in his right hand and positioned in front of him." The detailed information helps in capturing exact motion dynamics and interactions between the participant and the surrounding dynamic environment. Use of subjective adjectives (*e.g.*, confidently, hesitantly, or meticulously) is encouraged to capture observed behaviors in a more expressive way.

3 Experiments

Text-to-Motion: We provide a comprehensive comparison of text-to-motion baselines using embedding-based analysis [7]. Table 1 includes R Top1, FID, Diversity, and Multi-Modality metrics. R Top1 measures retrieval accuracy, FID assesses the realism of generated motions, Diversity evaluates the variance of generated motions, and Multi-Modality examines how generated motions vary within each text description. For evaluation, we train a feature embedding model proposed by HumanML3D [3]. Notably, we observe high FID when baselines are trained only with Motion-X. Due to the lack of blind motion in Motion-X, models trained solely on Motion-X tend to generate diverse but unrealistic blind motions, subsequently increasing the distance between the feature space of generated motions and real blind motions. We demonstrate training on BlindWays improves model performance here as well; for instance, models trained on BlindWays achieve an FID of 3.340, significantly closer to the real data's FID of 0.257, indicating higher realism.

Motion Prediction: Finally, we measure the capabilities of motion-conditioned models, where both past motion and text context are provided as input to the model. Unlike text-driven motion generation, this approach focuses on predicting diverse and plausible future motions given a history of motion. The models are trained to predict the next 9.5 seconds of future motion given 0.5 seconds of past motion. We further incorporate text embeddings into the stochastic modeling approaches, including CVAE [8] and DLow [10], allowing the model to be conditioned on text for controllable motion generation. We further ablate DLow leveraging a transformer module, referred to as DLow+. This ablation significantly increases sample diversity, without hindering accuracy. We evaluate MotionGPT [7] on motion-to-motion for a fair comparison. Table 2 shows CVAE-based models predict diverse future motion without compromising on realism. While CVAE achieves an APD of 7.68, DLow and its variant successfully enhance sample diversity by 52% and 97%, respectively. These results demonstrate the effectiveness of incorporating text embeddings through a late fusion technique in improving the diversity and accuracy of motion-driven motion generation models. CVAE-based methods show better accuracy compared to MotionGPT due to their focus on capturing fine-grained

Table 2: **Evaluating Motion Prediction on BlindWays.** Given text description and a motion history window as inputs, we predict future 9.5-second 3D poses and compute diversity (APD, higher is better) and accuracy (ADE and FDE, lower is better) pose metrics.

Method	APD ↑	ADE ↓	FDE ↓
MotionGPT [7]	-	3.40	3.44
CVAE [8]	7.68	0.47	**0.56**
DLow [10]	11.65	0.46	0.59
DLow+ [10]	**15.14**	**0.45**	**0.56**

motion details through conditional variational approaches, whereas MotionGPT prioritizes averaged motion patterns.

4 Conclusion

In this study, we introduce BlindWays, a novel benchmark focusing on the unique motion behaviors of blind and low-vision pedestrians navigating dynamic outdoor urban environments. Our dataset includes 3D motion data enriched with high- and low-level text descriptions informed by corresponding third-person and ego-centric RGB videos that meticulously capture the actions, purposes, and environmental contexts of blind motion, particularly how they utilize canes to interact with their surroundings. Our experiments demonstrate that existing state-of-the-art motion-language models struggle to generalize to blind motion despite their advancements, highlighting the unique challenges posed by this domain. This underscores the necessity of a blind motion benchmark to ensure safe and effective urban planning, such as autonomous driving. Furthermore, we emphasize the significant labor and time, an effort that took more than 2 years, required to capture and annotate a comprehensive, high-quality dataset. The BlindWays provides a rich, contextually detailed resource enabling models to accurately and diversely model blind motion, advancing the field of motion-language modeling and enhancing the safety and reliability of autonomous systems.

5 Limitations

Our work addresses a prevalent bias in motion modeling datasets, specifically the focus on sighted and simplified pedestrian motion. Our study underscores the complexity of diverse motion modeling, particularly in cases where pre-training may be non-beneficial or even detrimental to model predictions, such as with blind motion. To tackle this bias, we collected realistically complex data within an important but under-discussed use case. However, our study has several limitations. The sample size of 11 participants, providing a dataset of $1{,}005$ motion samples after filtering pose tracking failure cases, is representative of in-situ accessibility studies. Nonetheless, additional real-world data from a more diverse participant pool could help identify further biases and model issues (e.g., various physical characteristics such as different heights and backgrounds). Another limitation is the expensive ($\$6{,}500$) motion-capture suit, which may hinder larger-scale studies. While we chose higher-cost, higher-quality tracking technology, lower-cost solutions (e.g., inertial, vision-based) are continuously being developed and can facilitate easier and more scalable capture, leading to more robust and practical motion models across many underrepresented use cases in current human motion benchmarks.

6 Authors' Bios

Hee Jae Kim is a third-year Ph.D. student in the ECE department at Boston University. Hee Jae received the BS and MS degrees from Ewha Womans University. Hee Jae's research interests lie in human-interactive systems and 3D motion modeling,

Kathakoli Sengupta is a second-year MS student at Boston University. Kathakoli received the BS degree from Vellore Institute of Technology, India. Kathakoli's research interests lie in vision, navigation, intelligent assistive systems, and robotics.

Masaki Kuribayashi is a second-year Ph.D. student at Waseda University, Japan. Masaki received an BS and MS degrees from Waseda University, and was a visiting researcher at Boston University during the Spring of 2024. Masaki's research interests lie in Accessibility and Machine Learning, particularly navigation for blind people and visual language navigation. Masaki has won several awards, including the Research Fellowship for Young Scientists (DC1) JSPS.

Hernisa Kacorri is an Associate Professor with a joint appointment in the College of Information and the University of Maryland Institute for Advanced Computer Studies (UMIACS). She also holds an affiliate appointment in the Department of Computer Science and serves as a core faculty at the Trace RERC. She received her Ph.D. in Computer Science from The Graduate Center at City University of New York, and has conducted research at the University of Athens, IBM Research-Tokyo, Lawrence Berkeley National Lab, and Carnegie Mellon University. Her research focuses on accessibility and human-centered artificial intelligence, with a mix of participatory methods and rigorous, experimental

approaches for assessing impact. Her most recent work on teachable machines does not see end-users as passive consumers but as active directors of AI-infused technology. Her collaborations with students, colleagues, and advisors have received honorable mention and best paper awards at ACM ASSETS, ACM CHI, IEEE WACV, IEEE VL/HCC.

Eshed Ohn-Bar is an Assistant Professor in the ECE department at Boston University. Prior to joining BU he was a Humboldt Fellow at the Max Planck Institute. Eshed's research lies at the intersection of machine intelligence, systems, and accessibility. His work received several awards, including the best paper award at the workshop on Analysis and Modeling of Faces and Gestures (2013), best industry related paper award runner up at ICPR (2014), best student paper award runner up at ICPR (2016), and the 2017 best PhD dissertation award from the IEEE Intelligent Transportation Systems Society. He was also part of a team which won the semi-finalist for the 2022 Department of Transportation's Inclusive Design Grand Challenge. Eshed received the BS degree in Mathematics from UC Los Angeles in 2010, MEd from UC Los Angeles in 2011, and the PhD degree in Electrical Engineering from UC San Diego in 2017.

7 Acknowledgments

We thank the Rafik B. Hariri Institute for Computing and Computational Science and Engineering (Focused Research Program award #2023-07-001) and the National Science Foundation (IIS-2152077) for supporting this research. Hernisa Kacorri was supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS (grant 90REGE0024) and NSF (grant 2229885).

References

[1] D. Ashmead. Street crossing by sighted and blind pedestrians at a modern roundabout. *JTE*, 2005.

[2] D. Geruschat. Driver behavior in yielding to sighted and blind pedestrians at roundabouts. 2005.

[3] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022.

[4] D. Guth et al. Blind and sighted pedestrians' judgments of gaps in traffic at roundabouts. 2005.

[5] A. Harrell. Driver response to a disabled ped. using a dangerous crosswalk. In *JEP*, 1992.

[6] A. Harrell. Effects of blind pedestrians on motorists. In *JSP*, 1994.

[7] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2024.

[8] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv.org, 1312.6114*, 2013.

[9] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang. Motion-X: A large-scale 3d expressive whole-body human motion dataset. In *NeurIPS*, 2024.

[10] Y. Yuan and K. Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020.