AI Guide Dog: Egocentric Path Prediction on Smartphone

Aishwarya Jadhav[1,*], Jeffery Cao[1], Abhishree Shetty[1], Urvashi Kumar[1], Aditi Sharma[1], Ben Sukboontip[1], Jayant Tamarapalli[1], Jingyi Zhang[1], Anirudh Koul[2]

[1]Carnegie Mellon University, Pittsburgh
[2]Pinterest

Abstract

This paper presents AI Guide Dog (AIGD), a lightweight egocentric (first-person) navigation system for visually impaired users, designed for real-time deployment on smartphones. AIGD employs a vision-only multi-label classification approach to predict directional commands, ensuring safe navigation across diverse environments. We introduce a novel technique for goal-based outdoor navigation by integrating GPS signals and high-level directions, while also handling uncertain multi-path predictions for destination-free indoor navigation. As the first navigation assistance system to handle both goal-oriented and exploratory navigation across indoor and outdoor settings, AIGD establishes a new benchmark in blind navigation. We present methods, datasets, evaluations, and deployment insights to encourage further innovations in assistive navigation systems.

Introduction

Navigation assistance systems for visually impaired individuals have been studied for several decades (Dakopoulos and Bourbakis 2010), yet their widespread adoption remains limited due to (1) reliance on expensive, custom-built devices, (2) the lack of efficient, robust models that ensure user safety, and (3) limited user trust and convenience.

Existing systems often rely on expensive devices with built-in sensors like LiDAR, or laser scanners (Wang et al. 2017; Wachaja et al. 2015; Hesch and Roumeliotis 2010), for 3D mapping, or IMUs, gyroscopes, and pedometers (Lee and Medioni 2014; Hesch and Roumeliotis 2010) for localizing user position and orientation. While accurate, these systems are bulky and prohibitively expensive, limiting their accessibility. To address these challenges, we propose a lightweight system leveraging only video feed from a smartphone camera. An on-device model generates navigation instructions in real-time, that are translated into audio cues for the blind user. This facilitates accessibility and ease of adoption while maintaining robust performance.

Most prior work (Wang, Liu, and Kennedy 2024; Qiu et al. 2022; Soo Park et al. 2016; Yagi et al. 2018) on blind navigation formulates it as a trajectory forecasting problem, predicting precise 3D positions of the user. However, this approach typically reflects the behavior of sighted individuals, who navigate by dynamically avoiding obstacles and other pedestrians. Blind navigation is fundamentally different: they typically prefer retaining canes (Ohn-Bar, Kitani, and Asakawa 2018), even when assisted by robotic systems. Canes help detect obstacles and signal others to yield space, facilitating smoother navigation. Thus, the user-environment dynamics of blind individuals differ significantly from those of sighted users.

This insight allows us to simplify the navigation task into an egocentric path prediction problem, where the system predicts all possible future user navigation actions—such as continuing straight, turning left, or turning right. This abstraction avoids the uncertainties of precise trajectory prediction and instead focuses on the user's heading direction and actions relative to their environment. We adopt a multi-label classification approach to accommodate multiple possible navigation directions and enable easy translation of the model's outputs to actionable commands for users. While limited prior work (Wang et al. 2017; Singh, Fatahalian, and Efros 2016) explored similar ideas, they were restricted to single-class predictions in constrained environments.

Existing blind navigation models lack the robustness needed for reliable real-world use, which requires adaptability across diverse scenarios—avoiding collisions in cluttered indoor spaces, as well as facilitating outdoor navigation with goal-based guidance from GPS-enabled apps like Google Maps. However, instructions from such apps (e.g., "Turn left at W. 4th St." (GoogleMaps 2024)) are often impractical for blind users. Consequently, most prior work focuses on no-goal (Wang, Liu, and Kennedy 2024; Qiu et al. 2022) or fixed-path (Ohn-Bar, Kitani, and Asakawa 2018) navigation, typically limited to either indoor or outdoor settings. AIGD bridges this gap by enabling both exploratory and goal-based navigation, allowing users to navigate freely or follow specific destinations. AIGD is the first system to handle scenarios across indoor and outdoor environments, with and without the *intent* of reaching a final destination, while also accounting for multiple possible directions in the absence of a goal.

First-person camera inputs face challenges like jitter, blur, limited fields of view, and variations in smartphone quality, camera positioning, and walking speeds. Although the slower, deliberate movements typical of visually impaired

*Corresponding author: anjadhav@alumni.cmu.edu

users (average walking speed: 0.72 m/s (Liu et al. 2019)) reduce extreme ego-motion effects, these challenges persist. Furthermore, real-world navigation data exhibits significant imbalance, with far more straight motions than turning actions. These observations inform our data collection, modeling and deployment processes.

Our key contributions include:

1. **A robust, lightweight multi-label classification model** addressing turn class imbalance, and effective across scenarios with or without destination intent.
2. **A methodology** for integrating destination and high-level direction signals into vision-only prediction models, validated by extensive experiments.
3. **An open-source dataset** comprising egocentric videos and associated mobile sensor data collected across diverse scenes and participants, facilitating future research.
4. **A low-latency smartphone app** deploying the model for real-world navigation assistance.

Related Work

Blind Navigation Assistance Systems: Previous systems primarily use either wearable devices or robotic helpers. Wearable systems incorporate body-mounted sensors (Abidi et al. 2024) (e.g., on feet, knees, or waist) and rely on standalone devices, like laptops in backpacks (Lee and Medioni 2016), smartphones (Sato et al. 2017; Pawar et al. 2022) or tablets (Li et al. 2019) for computations. For instance, Lee and Medioni (2016) developed a head-mounted RGB-D camera system for ego-motion estimation and obstacle-aware path planning, providing tactile feedback through a haptic vest. Wang et al. (2017) introduced a wearable structured light camera-based system providing feedback via vibrations and Braille.

Robotic helpers, such as smart canes (Hesch and Roumeliotis 2010; Gupta et al. 2015; Yang, Gao, and Choi 2024) or suitcase-mounted devices (Manglik et al. 2019), act as robotic guide dogs. For instance, Wachaja et al. (2015) proposed a robotic walker with laser rangefinders and servo motors for egomotion estimation and obstacle detection, while ISANA (Li et al. 2019) integrates an RGB-D camera with a Google Tango tablet providing multimodal feedback through speech, audio, and haptics.

Egocentric Navigation: Egocentric navigation comprises trajectory forecasting, which predicts future 2D/3D positions, and path prediction, which identifies discrete directional actions (e.g., left, right, forward), with related research extending into goal-based navigation. Trajectory forecasting, while extensively studied for vehicles, has seen limited attention for human navigation. Yagi et al. (2018) proposed a convolution-deconvolution network using pedestrian poses and ego-motion, while others have used GRU-CNN (Styles, Sanchez, and Guha 2020) and LSTM encoder-decoder models (Qiu et al. 2021) to predict human trajectories in indoor environments. More recent methods leverage multimodal transformers (Qiu et al. 2022) and diffusion models (Wang, Liu, and Kennedy 2024) to incorporate scene semantics and past trajectories for future prediction.

Egocentric path prediction methods include Lee and Medioni (2014), which generates 3D occupancy maps and utilizes D*-Lite planning for four directional cues (straight, left, right, stop); Singh, Fatahalian, and Efros (2016), which uses a fine-tuned CNN to predict discrete motion classes from single camera images; and Ohn-Bar, Kitani, and Asakawa (2018), which developed a dynamic path planning policy personalized to user reaction times, providing localized guidance based on global pre-planned layouts.

Most goal-based navigation research focuses on robotics (Sánchez-Ibáñez, Pérez-del Pulgar, and García-Cerezo 2021) and autonomous vehicles (Aradi 2022), relying on dynamic path planning and reinforcement learning. However, these approaches are unsuitable for modeling human behavior, particularly for blind users due to unique social and reaction constraints.

To the best of our knowledge, AIGD is the first system to generalize navigation for blind users across diverse scenarios. Our approach is motivated by the unique requirements of this use-case, allowing us to scope the problem to a limited set of instruction classes while incorporating goal-based navigation and directional uncertainty, without relying on complex dynamic planning or reinforcement learning. This ensures a practical and user-centric solution.

Method

System Overview

Our system features a smartphone app running a lightweight, real-time model on-device, using video input from the device's camera, and optionally GPS and Google Maps data for destination-based navigation. Sensor data (e.g., accelerometer, gyrometer) is used only during data collection for prediction label generation, not for model inference in the deployed user app. Navigation predictions are post-processed into audio instructions for the user.

Problem Definition

We model this task as a multi-label classification problem, where, for each frame sampled from the smartphone camera stream, the system predicts the user's heading direction one second into the future. Specifically, based on the current scene and, optionally, past frames, the model outputs classification scores for three possible directions (FRONT, LEFT, RIGHT) the user could take in the next second. The one-second future horizon is informed by studies on walking speeds (Liu et al. 2019) and average reaction times (Bhirud and Chandan 2017) specific to our blind user base.

Multiple turn labels are generated only in scenarios without destination intent, typically at intersections or when pathways diverge. In such cases, the model must predict all possible directions one second before the turn begins. During the turn, it must predict the active turn direction (LEFT-/RIGHT), and finally transition to predicting FRONT, with high confidence, as the turn concludes. Fig. 1 illustrates this with frames sampled at 1 FPS. For free-space navigation without obstacles, our labeling scheme conditions the model to output only the FRONT direction.

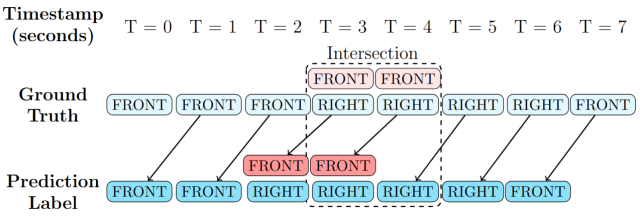

Figure 1: Labeling Scheme for frames sampled at 1 FPS. Red blocks denote other walkable directions at intersection.

Dataset

At the time of our study, existing egocentric walking datasets lacked the diversity of scenarios and the sensor or GPS data needed for our use case. Besides, most were confined to specific environments (e.g., labs, offices), and relied on specialized cameras or hardware.

To better reflect the real-world conditions of our app's usage, we collected a custom dataset using smartphone cameras, aiming to capture the walking speeds and styles of visually impaired individuals. To constrain the study, we used iPhone 13 with the AIGD data collection app installed. Eight participants, primarily graduate students and tech interns, collected data in semi-crowded spaces with stationary obstacles and people. To simulate social navigation interactions similar to those experienced by our target users, participants wore black glasses and carried the smartphone on a lanyard near their chest, walking slowly and cautiously. This setup also captured variations in first-person camera movements, camera positionings, and fields of view.

Data was collected for diverse indoor and outdoor scenes across Pittsburgh, Seattle, and the Bay Area, as described in Tab. 1, focusing on everyday venues outside users' homes or offices. Indoor environments included well-lit spaces with numerous aisles, such as grocery stores, to increase the frequency of left and right turns. Since there is no destination for these, participants were instructed to turn at every available opportunity. Outdoor data was collected in parks with winding pathways and city streets during daytime. All walking paths were unscripted and unplanned, with each video capturing a single scene and lasting 2 to 10 minutes.

In total, the dataset includes 57 hours of walking data at 30 fps comprising videos from smartphone cameras and sensor data (accelerometer, gyrometer, magnetometer, pedometer) captured at 0.1-second intervals. For outdoor scenes, GPS locations and directional data from the Google Maps API were also logged.

Video frames were down-sampled to 2 fps, and converted to 128×128 gray-scale. Sensor data underwent denoising, reference transformations, and windowing to generate ground truth labels, which were then timestamp-aligned with the video frames. GPS and high-level destination directions were aligned with these records, where available. Each data example consists of a frame, its label, the preceding 5 seconds (10 frames) as context, and associated GPS and direction features for all frames. Past context frames help inform future predictions by implicitly capturing the user's walking speed and reaction time. The dataset consists of 392,580 examples, split 60:20:20 for training, validation, and testing, ensuring no scene overlap across splits, as summarized in Table 1.

		Scene	Data Split
Indoor	32 hours; 220k examples	Library 1	Train
		CMU Hall	Train
		Library 2	Validation
		Grocery Store	Test
Outdoor	25 hours; 172k examples	Pittsburgh Street 1	Train
		Park (70% videos)	Train
		Pittsburgh Street 2	Validation
		Seattle Street 1	Test
		Park (30% videos)	Test

Table 1: Data Splits

We open-source[†] a subset of the collected egocentric videos and associated mobile sensor data, where release consent has been obtained from relevant authorities, ensuring compliance with ethical and privacy regulations.

Ground Truth Labels Labels for each sampled frame were derived from sensor data. Various methods, including accelerometer, compass, GPS, and optical flow-based approaches were evaluated for turn label calculation. Among these, the compass-based method consistently yielded the most accurate results, particularly at slower walking speeds, proving less noisy than accelerometers and more precise than GPS for heading estimation. Details of these approaches, evaluations and parameter tuning, follow (Markevych et al. 2021). Below is a brief overview.

For each data point, the turning angle is computed by comparing the agent's facing direction over a 0.5-second interval. Angles above 5° indicate a RIGHT turn, below -5° a LEFT turn, and within $\pm 5^\circ$ represent FRONT movement. The 5° threshold, an adjustable hyperparameter, controls sensitivity to minor orientation changes.

For indoor scenarios without destination intent, auto-calculated turn labels were manually re-labeled to ensure all possible turn directions at intersections were captured. Multiple labels were assigned to the initial 2 seconds of each turn.

Models

This section outlines the models used to validate our proposed problem formulation and intent integration methodology. For no-destination (no-intent) navigation, we implement multi-label classification models, including simple baselines such as CNN and ConvLSTM, commonly employed in prior work (Styles, Sanchez, and Guha 2020; Qiu et al. 2021; Singh, Fatahalian, and Efros 2016). We then extend these no-intent models by incorporating destination intent features and GPS signals to enable goal-conditioned predictions.

CNN: Following Singh, Fatahalian, and Efros (2016), we finetuned a ResNet34 model with a linear classification head

[†]Online open-sourced dataset: http://bit.ly/41h7jJn

to encode individual frames. This baseline model only considers per-frame information, disregarding the temporal context provided by preceding frames.

ConvLSTM: The ConvLSTM (SHI et al. 2015) architecture, described in Fig. 2a, designed for spatiotemporal prediction, serves as our temporal baseline to leverage the visual information in the preceding context frames. However, it is computationally intensive and susceptible to overfitting, particularly with limited fine-tuning data.

PredRNN: PredRNN (Wang et al. 2023) utilizes spatiotemporal LSTM units to model sequential dependencies in video data, and is widely used for future frame prediction tasks (Jadhav 2020; Ma, Zhang, and Liu 2022). We explore PredRNN's ability to model complex short-term dynamics for our future direction prediction use-case. However, like ConvLSTM, PredRNN is computationally demanding for both training and inference, with even higher latency due to its increased architectural complexity.

Intent-based Navigation

For outdoor navigation, directions from Google Maps provide high-level guidance by localizing the user via GPS. However, GPS accuracy ($\sim$ 4.9 meters (GPS.gov 2024)) is insufficient for precise local navigation. To address this, the model must predict local directions, and the corresponding actions to take in the next second, that are aligned with the high-level Maps directions and the user's GPS history.

In this work, we use the *Google Maps Directions API* (GoogleMaps 2024), which provides step-by-step walking instructions for specified start and destination addresses. For the "walking" mode, the API returns an array of steps, each containing a `start_location` (latitude-longitude), `end_location`, and a `maneuver` or action to take at the `end_location`. Each step corresponds to a travel segment. At each timestep, we pick the appropriate segment to retrieve the maneuver from based on the user's GPS position and the segment's `start_location` and `end_location`. Relevant maneuvers for walking include `turn-slight-left`, `turn-sharp-left`, `turn-left`, `turn-slight-right`, `turn-sharp-right`, `turn-right`, and `straight`. We one-hot-encode the maneuver values and append them with the latitude and longitude from the `start_location` and `end_location` fields to create an ***Intent Feature*** for each step. The Intent Feature is then combined with the user's current GPS coordinates and passed through a linear embedding layer to generate the ***Intent Embedding*** vector, which serves as a high-level intent conditioning input for the model.

The following sections detail the modifications made to the baseline CNN and ConvLSTM architectures to incorporate Intent Embeddings. As discussed in the results section , the performance gains offered by PredRNN are outweighed by its high computational requirements and latency, which are critical factors for smartphone deployment. Hence, it is excluded from the intent-based experiments. Instead, we implement a more efficient, hybrid CNN+LSTM architecture to capture both temporal and intent signals. **CNN with Intent**: Intent embeddings are concatenated with

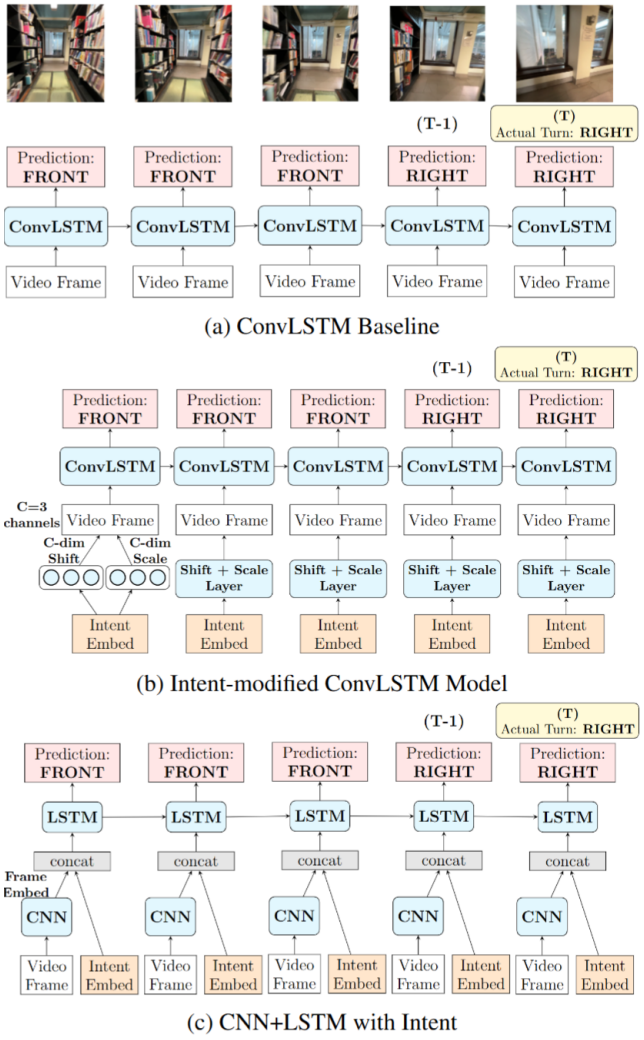

Figure 2: Model Architectures.

CNN-extracted frame embeddings, which are then passed through a 2-layer MLP for prediction.

ConvLSTM with Intent (Fig. 2b): Intent embeddings are projected down to two C-dimensional vectors, where C=3 corresponds to the number of video frame channels. These vectors are added as shift and scale factors to the frame input channels before passing through the ConvLSTM layers. While we explored other early fusion strategies, this method demonstrated the best performance vs complexity tradeoff.

CNN + LSTM with Intent (Fig. 2c): This architecture enhances the CNN+Intent model by replacing the MLP in the final classifier with a 2-layer LSTM. It combines past frame embeddings, extracted via the powerful ResNet feature extractor, with the corresponding timestep's intent embeddings, and uses LSTMs to model past context temporal relationships. Compared to ConvLSTM, CNN+LSTM is more computationally efficient as the LSTM processes lower-dimensional embeddings instead of full image data.

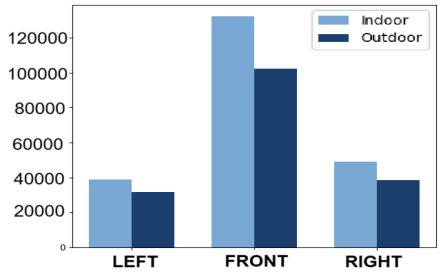

Figure 3: Distribution of class labels in the dataset.

Augmentation	Specs	Purpose
Translation	5–25 pixels vertical and horizontal	Varying camera heights
Color Jitter	0-20% HSV	Varying lighting
Random Crop	5-20 pixel squares	Occlusions
Rotation	-20 to +20 degrees	Camera rotations during walking
Noise	Gaussian or salt-pepper	Differing camera qualities

Table 2: Augmentations Settings

Experiments

Label Imbalance

Despite efforts to collect more turn-based data, the dataset remains significantly skewed toward the FRONT label, as seen in Fig. 3. However, predicting turns (LEFT and RIGHT) is more critical for navigation. To address this imbalance and improve turn prediction, we implemented the following during training:

1. **Minority Oversampling:** Doubled LEFT/RIGHT class examples.
2. **Data Augmentation:** Applied transforms described in Tab. 2 with a 20% probability.
3. **Loss Function:** Oversampling reduces class imbalance but does not fully address the more challenging, yet rarer, turn prediction cases near the start and end of a turn. To mitigate this, we employed Focal Loss (Lin et al. 2020), which emphasizes harder-to-classify samples by dynamically scaling their loss contribution. We also experimented with Weighted BCE Loss using class weights of 2:2:1 (LEFT:RIGHT:FRONT), which provided minor performance gains. These weights were integrated into our focal loss formulation.

Sampling, augmentation, and loss settings were determined through experiments on CNN/ConvLSTM baselines.

Experimental Setup

All models were fine-tuned using the label balancing settings described above. Best available public checkpoints were used to initialize the CNN and ConvLSTM components for both no-intent and intent-based models, as well as PredRNN. The final MLP in the CNN-based models, LSTM in the CNN+LSTM+Intent model and the Intent Embedding layers were trained from scratch. Training was conducted for 30 epochs with a batch size of 64, using the Adam optimizer with a weight decay of 1e-3. The CosineAnnealingLR scheduler was used, with learning rates of 1e-4 for layers trained from scratch and 1e-5 for fine-tuned layers.

We conducted ablation studies to assess the impact of different training data configurations. No-intent models are well-suited for indoor scenarios, where GPS and high-level directions are unavailable, but multi-label supervision is provided. In contrast, intent-based models effectively leverage GPS and directional features in outdoor datasets. Hence, we trained the no-intent models exclusively on the indoor dataset and the intent-based models on the outdoor dataset. Performance was evaluated on corresponding test sets and benchmarked against our generalized intent models trained on combined indoor and outdoor datasets.

Evaluation Metrics: We evaluate the models using accuracy, AUC, Precision, Recall, and F1 score.

Results

Tab. 3 details the performance of models trained on combined indoor and outdoor datasets, evaluated on a test set containing both indoor and outdoor video segments. Tab. 4 breaks down the AUC evaluations of these models for indoor and outdoor test videos separately. Tab. 5 summarizes the results of the training dataset ablations.

For all experiments, the performance of the FRONT label remains largely unchanged by the modifications. Since our use case emphasizes turn (LEFT/RIGHT) classes, the following sections focus only on their performance.

Performance of No-Intent Models

Without intent features, ConvLSTM and PredRNN outperform CNN by leveraging temporal context, which is particularly beneficial for path disambiguation in the absence of high-level intent cues. Temporal modeling enhances scene understanding, especially during turns, where past frames provide context about the ongoing action (e.g., turning), and the current frame helps decide whether to continue or conclude the turn.

Among no-intent models, PredRNN achieves the best performance on the benchmark test dataset, surpassing ConvLSTM due to its advanced future frame prediction architecture. Despite its complexity, PredRNN generalizes better than ConvLSTM, exhibiting less overfitting. However, its computational demands outweigh its performance gains, making it an unsuitable candidate for on-device deployment.

No-intent models perform better on indoor video segments than outdoor scenes. This is expected, as indoor datasets provide multi-label supervision for ambiguous scenarios like intersections, while outdoor datasets use single-label annotations for each turn. Consequently, we observe many false positives at outdoor intersections, because the no-intent models cannot leverage the disambiguation provided by the Intent Features.

Effects of Adding Intent Features

Incorporating intent features and GPS signals enhances the performance of CNN and ConvLSTM models over their no-

Model	LEFT				RIGHT				FRONT			
	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1
CNN	0.571	0.610	0.543	0.5746	0.608	0.702	0.567	0.6273	0.900	0.903	0.803	0.8501
ConvLSTM	0.622	0.689	0.544	0.608	0.645	0.725	0.572	0.6395	0.908	0.900	0.810	0.8526
PredRNN	**0.636**	0.708	0.549	**0.6184**	**0.657**	0.752	0.570	**0.6485**	0.912	0.910	0.840	0.8736
CNN + Intent	0.588	0.619	0.548	0.5813	0.622	0.711	0.571	0.6334	0.911	0.910	0.833	0.8698
ConvLSTM + Intent	0.638	0.706	0.556	0.6221	0.659	0.742	0.571	0.6454	0.912	0.918	0.830	0.8718
CNN + LSTM + Intent	**0.664**	0.728	0.559	**0.6324**	**0.700**	0.766	0.583	**0.6621**	0.920	0.920	0.846	0.8814

Table 3: Performance for models trained on combined Indoor + Outdoor training datasets. AUC PR, Precision, Recall and F1 scores are reported on entire the test set (Indoor + Outdoor).

Model	Indoor			Outdoor		
	LEFT	RIGHT	FRONT	LEFT	RIGHT	FRONT
CNN	0.579	0.614	0.905	0.550	0.592	0.900
ConvLSTM	0.628	0.649	0.909	0.609	0.634	0.905
PredRNN	0.641	0.662	0.918	0.621	0.645	0.910
CNN with Intent	0.577	0.613	0.900	0.607	0.638	0.914
ConvLSTM with Intent	0.632	0.649	0.911	0.651	0.672	0.914
CNN + LSTM with Intent	**0.660**	**0.695**	0.917	**0.671**	**0.707**	0.920

Table 4: Performance of models trained on combined Indoor + Outdoor training datasets, with AUCs reported separately for Indoor and Outdoor test data.

Model	Train / Test	LEFT	RIGHT	FRONT
CNN	Indoor	0.590	0.626	0.913
ConvLSTM	Indoor	0.643	0.655	0.918
PredRNN	Indoor	**0.667**	**0.687**	0.925
CNN + Intent	Outdoor	0.600	0.629	0.906
ConvLSTM + Intent	Outdoor	0.640	0.659	0.908
CNN + LSTM + Intent	Outdoor	0.66	0.682	0.912

Table 5: Ablations with different train and test data mixes.

intent counterparts, as shown in Tab. 3. The gains are particularly significant for outdoor test videos (Tab. 4), where intent and GPS signals provide explicit directional cues and help disambiguate turns.

For the indoor test set, the performance difference between no-intent and intent models is negligible. Given the substantial gains observed for outdoor scenarios, the intent models depict a net positive improvement while supporting both scenarios.

Finally, the CNN+LSTM+Intent model outperforms ConvLSTM+Intent in both evaluation metrics and computational efficiency. This mid-fusion approach surpasses the early fusion strategy in ConvLSTM+Intent by independently extracting frame features while jointly modeling temporal information from frame and intent embeddings. Notably, CNN+LSTM+Intent achieves greater gains on indoor videos compared to CNN+Intent and ConvLSTM+Intent models, reducing the performance gap between indoor and outdoor datasets. This is likely due to the later fusion of modalities in the classification head, which better isolates the contributions of video, and Intent/GPS features.

Training Data Ablations

In tab. 5, the no-intent models trained exclusively on indoor data outperform those trained on a mix of indoor and outdoor data when evaluated on the indoor test videos. In the absence of ambiguous turn supervision in the outdoor training set, these models learn from the much cleaner indoor training set, effectively capturing the aisle and turn patterns.

In contrast, intent models trained solely on outdoor data perform worse than those trained on a mix of indoor and outdoor datasets, overfitting the smaller outdoor training set. While indoor test metrics are not significantly enhanced by the intent modifications, including indoor videos in the training data benefits the overall performance of intent models.

Qualitative Analysis

Fig. 4 presents GradCAM (Selvaraju et al. 2017) visualizations from the CNN model. Even the simple CNN baseline effectively learns path features and curves, enabling it to detect turns in the near future.

Deployment

We deployed our best generalized model, CNN+LSTM+Intent, to an iPhone 13 using CoreML (Apple 2024), as shown in Fig. 5, optimizing for minimal inference latency and on-device resource usage, including memory, GPU, and battery. We tuned the video frame

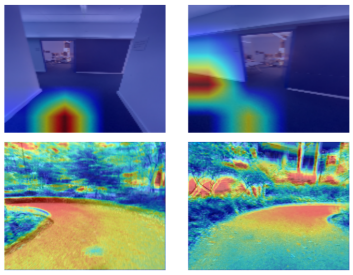

Figure 4: Grad-CAM heatmaps for indoor/outdoor scenes.

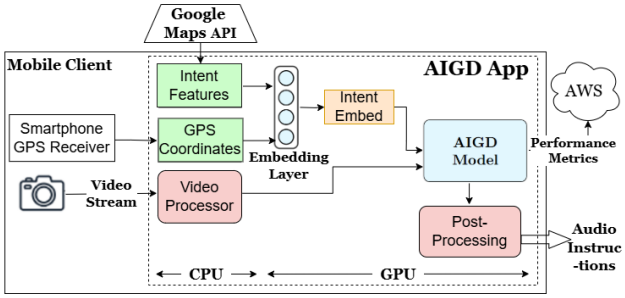

Figure 5: AIGD Deployment Architecture

rate and conducted quantization experiments, monitoring resource consumption metrics. Fig. 6 summarizes the results. The final setup used 16-bit model quantization and 2 FPS videos, balancing performance and resource efficiency.

Inference frequency is 2Hz, allowing the system to assess the state of the environment approximately every 0.5 seconds. Given our users' typical walking speeds and environments, this is sufficient for reliable real-time navigation while maintaining accuracy, adaptability and efficiency.

User Privacy: All processing occurs locally on the device, with only anonymized performance metrics sent to the server. No raw camera or sensor data is stored or shared.

Discussion

To ensure reliable and robust navigation, we assessed the model across varied settings. It demonstrates strong generalization from training environments (e.g., libraries, university halls) to unseen test locations like grocery stores. Future work will expand the dataset to further test performance in more unfamiliar environments. Addressing this is critical for user safety, especially in ambiguous situations.

Our novel intent-conditioning technique to extend ego-path prediction models utilizes simple architectures, commonly used in literature, due to latency constraints. Future research could explore the potential of this technique with advanced architectures and improved latency optimization.

The exploration of larger, more complex models could also be facilitated by expanding the dataset to include additional scenarios. Incorporating videos from diverse smartphone cameras would also improve the model's invariance and generalization across different hardware configurations.

Error Analysis and Future Improvements: An analysis of the most common error patterns in the model's predictions revealed challenges in: 1) Dynamic environments, such as mis-predictions around moving objects (e.g., pedestrians) or stopping for obstacles and traffic lights, and 2) Ambiguous path structures, including blocked or forked paths and nuanced turns that are not strictly LEFT or RIGHT.

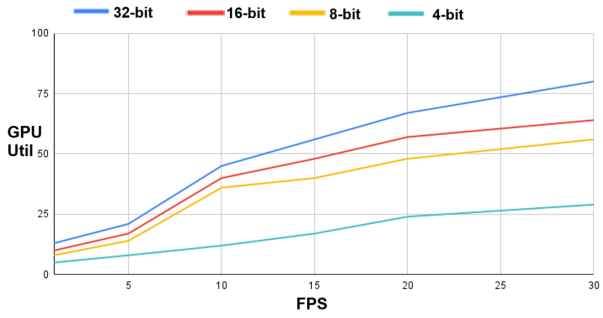

Figure 6: GPU utilization across quantization levels at varying frame rates.

Currently, our model supports three directional classes, but future work could easily introduce more granular classes, like finer-grained turning angles and start/stop walking commands, by collecting and labeling more data.

Our research into cane-walking patterns and social dynamics of blind navigation guided our approach based on the limited scene information from a smartphone camera, relying on the assumption that canes help detect obstacles and navigate blocked paths. However, explicitly modeling the behavior of pedestrians, vehicles, and other environmental entities could enable more nuanced navigation paths.

Implications of the Egocentric Dataset: Our released indoor and outdoor egocentric video dataset not only advances research in first-person view analysis, ego-motion estimation and forecasting but also has broader applications in mobility patterns, pedestrian behavior, and accessibility design. The dataset provides valuable insights into how visually impaired individuals navigate urban spaces, potentially aiding urban planning and infrastructure development by: 1) Identifying challenging navigation areas that need infrastructure improvements, 2) Optimizing pedestrian pathways for better accessibility, 3) Enhancing public transit by providing real-time accessibility information.

Conclusion

This paper presents AI Guide Dog (AIGD), an egocentric navigation system for visually impaired users. We introduce a novel intent-conditioning technique and multi-label classification framework to tackle goal-based navigation and directional uncertainty in no-destination scenarios. Our simple smartphone-based deployment lowers financial and technical barriers to adoption, without compromising performance. Additionally, our released egocentric dataset provides valuable insights to help further research in assistive technologies and accessibility-focused urban planning. Given the limited work in this domain, we hope to inspire future advancements in assistive navigation technologies.

References

Abidi, M. H.; Siddiquee, A. N.; Alkhalefah, H.; and Srivastava, V. 2024. A comprehensive review of navigation systems for visually impaired individuals. *Heliyon*, 10(11): e31825.

Apple. 2024. Core ML Documentation. Accessed: 2024-12-27.

Aradi, S. 2022. Survey of Deep Reinforcement Learning for Motion Planning of Autonomous Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 740–759.

Bhirud, B. G.; and Chandan, L. M. 2017. Comparative study of simple auditory reaction time in blind and blindfolded sighted individuals. *Natl J Physiol Pharm Pharmacol*, 7(1): 64–67. Online Published: 06 Aug 2016.

Dakopoulos, D.; and Bourbakis, N. G. 2010. Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1): 25–35.

GoogleMaps. 2024. Google Maps Directions API Documentation. Accessed: 2024-12-27.

GPS.gov. 2024. GPS Accuracy: Official U.S. Government Information. Accessed: 2024-12-27.

Gupta, S.; Sharma, I.; Tiwari, A.; and Chitranshi, G. 2015. Advanced guide cane for the visually impaired people. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 452–455.

Hesch, J. A.; and Roumeliotis, S. I. 2010. Design and Analysis of a Portable Indoor Localization Aid for the Visually Impaired. *Int. J. Rob. Res.*, 29(11): 1400–1415.

Jadhav, A. 2020. Variable Rate Video Compression using a Hybrid Recurrent Convolutional Learning Framework. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6.

Lee, Y. H.; and Medioni, G. 2014. Wearable RGBD Indoor Navigation System for the Blind. In *Computer Vision - ECCV 2014 Workshops*, 493–508. Springer.

Lee, Y. H.; and Medioni, G. 2016. RGB-D Camera Based Wearable Navigation System for the Visually Impaired. *Comput. Vis. Image Understand.*, 149(C): 3–20.

Li, B.; Muñoz, J. P.; Rong, X.; Chen, Q.; Xiao, J.; Tian, Y.; Arditi, A.; and Yousuf, M. 2019. Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People. *IEEE Transactions on Mobile Computing*, 18(3): 702–714.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Liu, X.; Zhang, S.; Zeng, J.; and Fan, F. 2019. Analysis and Optimization Strategy of Travel System for Urban Visually Impaired People. *Sustainability*, 11(6).

Ma, Z.; Zhang, H.; and Liu, J. 2022. MS-RNN: A Flexible Multi-Scale Framework for Spatiotemporal Predictive Learning. *ArXiv*, abs/2206.03010.

Manglik, A.; Weng, X.; Ohn-Bar, E.; and Kitanil, K. M. 2019. Forecasting Time-to-Collision from Monocular Video: Feasibility, Dataset, and Challenges. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8081–8088. IEEE Press.

Markevych, I.; Jing, A.; Wang, Y.; Katatkar, A.; Khadilkar, K.; Ganju, S.; Zitouni, T.; and Koul, A. 2021. Labeling by Moving: An End-to-End Data Pipeline for Egocentric Trajectory Prediction.

Ohn-Bar, E.; Kitani, K.; and Asakawa, C. 2018. Personalized Dynamics Models for Adaptive Assistive Navigation Systems. In *Conference on Robot Learning*.

Pawar, A.; et al. 2022. Smartphone based tactile feedback system providing navigation and obstacle avoidance to the blind and visually impaired. In *Proceedings of the 2022 5th International Conference on Advances in Science and Technology (ICAST)*. IEEE.

Qiu, J.; Chen, L.; Gu, X.; Lo, F. P.-W.; Tsai, Y.-Y.; Sun, J.; Liu, J.; and Lo, B. 2022. Egocentric Human Trajectory Forecasting With a Wearable Camera and Multi-Modal Fusion. *IEEE Robotics and Automation Letters*, 7(4): 8799–8806.

Qiu, J.; Lo, F. P.-W.; Gu, X.; Sun, Y.; Jiang, S.; and Lo, B. 2021. Indoor Future Person Localization from an Egocentric Wearable Camera. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Sato, D.; Oh, U.; Naito, K.; Takagi, H.; Kitani, K.; and Asakawa, C. 2017. NavCog3: An Evaluation of a Smartphone-Based Blind Indoor Navigation Assistant with Semantic Features in a Large-Scale Environment. *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

SHI, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-k.; and WOO, W.-c. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Singh, K. K.; Fatahalian, K.; and Efros, A. A. 2016. KrishnaCam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–9.

Soo Park, H.; Hwang, J.-J.; Niu, Y.; and Shi, J. 2016. Egocentric Future Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Styles, O.; Sanchez, V.; and Guha, T. 2020. Multiple Object Forecasting: Predicting Future Object Locations in Diverse Environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 690–699.

Sánchez-Ibáñez, J. R.; Pérez-del Pulgar, C. J.; and García-Cerezo, A. 2021. Path Planning for Autonomous Mobile Robots: A Review. *Sensors*, 21(23).

Wachaja, A.; Agarwal, P.; Zink, M.; Adame, M. R.; Möller, K.; and Burgard, W. 2015. Navigating blind people with a smart walker. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6014–6019.

Wang, H.-C.; Katzschmann, R. K.; Teng, S.; Araki, B.; Giarré, L.; and Rus, D. 2017. Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 6533–6540.

Wang, W.; Liu, C. K.; and Kennedy, M. 2024. EgoNav: Egocentric Scene-aware Human Trajectory Prediction. arXiv:2403.19026.

Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Yu, P. S.; and Long, M. 2023. PredRNN: A Recurrent Neural Network for Spatiotemporal Predictive Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2208–2225.

Yagi, T.; Mangalam, K.; Yonetani, R.; and Sato, Y. 2018. Future Person Localization in First-Person Videos. In *CVPR*, 7593–7602.

Yang, Z.; Gao, M.; and Choi, J. 2024. Smart walking cane based on triboelectric nanogenerators for assisting the visually impaired. *Nano Energy*, 124: 109485.