# Learning from Human Gaze: Human-like Robot Social Navigation in Dense Crowds

**Zhecheng Yu**[*1], **Yan Lyu**[*1], **Chen Yang**[1], **Tao Chen**[1], **Yishuang Zhang**[1], **Bo Ling**[1], **Peng Wang**[2], **Guanyu Gao**[3], **Weiwei Wu**[1], **Brian Y. Lim**[†4]

[1]Southeast University
[2] University of Surrey
[3]Nanjing University of Science and Technology
[4]National University of Singapore
{yuzhecheng, lvyanly, ch_yang, chentao, zhangyishuang, boling}@seu.edu.cn,
peng.wang@surrey.ac.uk, gygao@njust.edu.cn, weiweiwu@seu.edu.cn, brianlim@comp.nus.edu.sg

## Abstract

Robot navigation in dense crowds requires understanding social cues that humans naturally use, yet existing methods struggle with real-world complexity. We investigate two questions: (1) Where do pedestrians look when navigating crowds? and (2) Can eye tracking improve robot navigation? To answer, we introduce GazeNav, an egocentric dataset collected via wearable eye trackers, featuring synchronized video, gaze, and trajectories in crowded environments. Analysis reveals that the gaze of pedestrians is closely related to the semantic presence and movement of other individuals, exhibiting distinct attention patterns across navigation behaviors. Building on this, we propose Gaze2Nav, a modular framework that first predicts human gaze to infer socially salient pedestrians, then incorporates the semantic attention into motion planning alongside visual inputs. Our method achieves 87.6% salient pedestrian prediction accuracy and reduces trajectory error by 15.4% over state-of-the-art baselines. By aligning with human gaze, our framework improves both performance and interpretability, advancing toward human-like, socially intelligent robot navigation.

## Introduction

Mobile robots are increasingly used for interactive tasks like delivery and guidance in public spaces. A critical capability is robust social navigation through dynamic pedestrian crowds. This requires more than simple obstacle avoidance, demanding the perception of social cues, prediction of human motion, and execution of safe, explainable, and human-like paths (Kretzschmar et al. 2016).

Existing navigation methods struggle with real-world crowds. Traditional rule-based planners, while interpretable, are often too conservative and can cause robots to ”freeze” in dense situations (Van Den Berg et al. 2010). Learning-based methods face their own challenges, often relying on impractical inputs like unavailable global trajectories (Gupta et al. 2018) or complex sensor setups (Chen et al. 2023). While reinforcement learning (RL) struggles with realistic crowd simulation (Everett et al. 2021), behavior cloning (BC) from egocentric video has emerged as a practical alternative. However, even these BC models lack explicit modeling of human social behavior and attention (Li et al. 2025). This highlights the need for algorithms that can reason about the environment using human-like social and semantic cues.

A fundamental question remains: *Where do pedestrians look when navigating dense crowds?* Human gaze reveals what is semantically and socially important to their decision-making. However, raw gaze coordinates alone are insufficient. Humans use gaze not merely to look, but to think—to extract semantics, encode temporary memory, and suppress distractions. Replicating this high-level attentional reasoning is key for human-like robot navigation. This leads to our core question: *Can we improve robot navigation by engineering an understanding of human gaze into models?*

Answering these questions requires new data and models. Current navigation datasets are often limited by sparse crowds and a lack of synchronized gaze data. To address these gaps, we introduce **GazeNav**, an egocentric dataset collected from humans navigating dense, real-world crowds using wearable eye trackers (see example frames with gaze points in Fig 1). Our analysis reveals that pedestrians spend over 88% of their time looking at other people, with distinct gaze patterns that correspond to different navigational behaviors and actions.

Building on these insights, we propose **Gaze2Nav**, a navigation framework that integrates human-like attentional reasoning. Gaze2Nav first predicts where a human would look in a scene. By matching this predicted gaze with semantic information about pedestrians, it generates semantic attention that guides a Motion Planner to produce human-like navigation actions. This process mirrors the human cognitive function of first looking to identify what is important, then deciding where to go. In summary, our contributions are

- An egocentric navigation dataset GazeNav featuring synchronized egocentric video, gaze coordinates, and trajectories from humans navigating high-dense, real-world crowds, which we use to analyze human social attention.

[*]These authors contributed equally.
[†]Corresponding author.

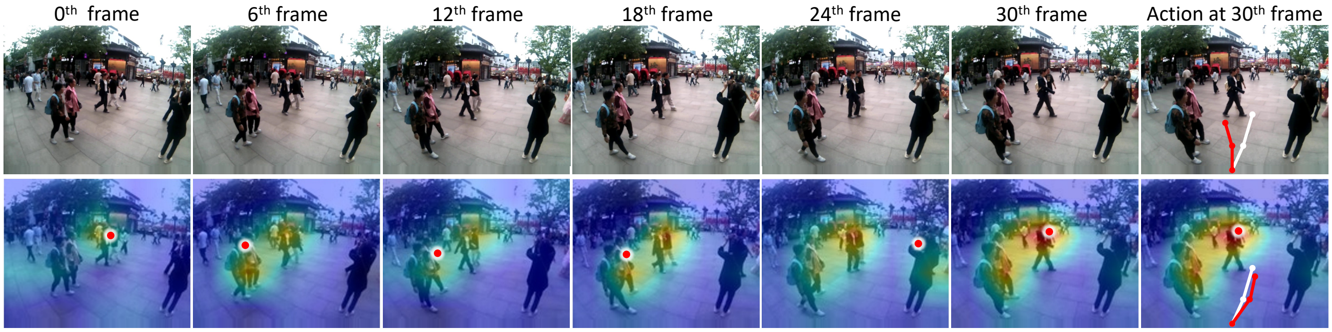


Figure 1: Example from our GazeNav dataset comparing a baseline method to our proposed Gaze2Nav. Top: A baseline model ViNT (Shah et al. 2023b) produces an unsafe trajectory (red) that deviates from the ground truth (white). Bottom: By using human gaze data, our Gaze2Nav model generates a safe and accurate trajectory. Its internal attention (heatmaps) correctly aligns with human gaze (red dots), showing it has learned to focus on critical social cues.

- A modular gaze-augmented navigation framework Gaze2Nav that mimics human cognition by using gaze prediction to identify socially important pedestrians and guide motion planning.
- Experiments show that Gaze2Nav achieves 87.6% accuracy in predicting salient pedestrians and reduces trajectory error by 15.4% compared to state-of-the-art methods, paving the way for more socially-aware robots.

## Related Work

This section surveys prior work, highlighting that current navigation methods lack the social awareness afforded by human gaze, a gap due to the absence of suitable datasets.

### Crowd Navigation

Conventional rule-based and optimization-based navigation methods (Helbing et al. 1995; Van Den Berg et al. 2010; Curtis et al. 2014) often perform poorly in dense crowds. Their conservative nature can lead to the "freezing" problem, where the robot halts facing with a crowd. Learning-based methods have been explored to overcome these limitations. However, approaches that learn from pre-extracted global trajectories (Everett et al. 2021; Mohamed et al. 2020) are impractical for real-time deployment as this data is not accessible in real-time robot deployment. Other methods that fuse multi-modal sensor inputs like RGB and LiDAR (Chen et al. 2023; Lai et al. 2025) also face significant deployment challenges due to complex sensor synchronization and calibration requirements.

We adopt a behavior cloning (BC) framework with egocentric vision, as it is more practical for learning socially-aware navigation from real-world demonstrations than reinforcement learning (RL), which struggles to simulate realistic crowds (Everett et al. 2021; Ling et al. 2024). While recent egocentric models like GNM (Shah et al. 2023a), ViNT (Shah et al. 2023b), and NoMaD (Sridhar et al. 2024) have shown strong performance, they lack explicit modeling of human social behaviors or attentional cues. Our Gaze2Nav framework addresses this gap by using predicted human gaze to enable more interpretable and human-aligned navigation in crowded scenes.

### Gaze-Augmented Learning

Human gaze offers a powerful, predictive signal of attention and intent (Rodin et al. 2021), yet this cue is largely missing from current navigation models. While prior work has used recorded gaze for video analysis (Huang et al. 2018; Tavakoli et al. 2019; Shen et al. 2018) or as expert supervision in imitation learning (Zhang et al. 2020; Guo et al. 2021; Saran et al. 2021), we propose a model that learns to *predict* human-like gaze in new scenes and, critically, grounds this prediction semantically to socially relevant objects. To our knowledge, this is the first work to integrate semantic gaze prediction into a real-world social navigation task, enabling an agent to perceive and respond to human-centric cues in dense crowds.

### Robot Navigation Datasets

Existing robot navigation datasets suffer from critical limitations, such as perspective mismatch in bird's-eye-view data (Zhou et al. 2011; Yi et al. 2015), unnatural pedestrian behavior elicited by teleoperated robots (Karnan et al. 2022; Bae et al. 2023), or a lack of dense crowds and synchronized gaze data in other wearable setups (Nguyen et al. 2023). To address these gaps, we introduce GazeNav, which captures naturalistic human navigation in dense, real-world crowds. GazeNav provides synchronized egocentric video, gaze coordinates, and trajectories. This unique combination of multimodal data across diverse scenarios enables new research into the joint analysis of visual attention and action for socially-aware navigation.

## The GazeNav Dataset

To investigate *where do pedestrians look to navigate dense crowds*, we introduce the **GazeNav** dataset. Collected in real-world environments using wearable sensors, it contains synchronized egocentric video, eye-tracking data, and 2D trajectories. The dataset captures naturalistic interactions for analyzing the link between social attention and navigation.

### Data Collection Procedure

We collected data from four participants (3 male, 1 female; mean_age=22.1, SD=2.5) using a wearable recording

| Scenario | Trajectory: Time (min) | Trajectory: Walk Dist (m) | Trajectory: Avg. Speed (m/s) | Avg. # people | Gaze: Dispersion (px) | Gaze: Saccade (Hz) | Gaze: On people (%) |
|---|---|---|---|---|---|---|---|
| Following | 35.44 | 2487.91 | 1.2±0.2 | 28±6 | 15.17 | 1.78 | 88.4 |
| Overtaking | 48.37 | 4614.83 | 1.6±0.1 | 24±11 | 16.30 | 1.90 | 95.2 |
| Crossing | 19.73 | 1243.11 | 1.0±0.1 | 12±4 | 19.97 | 2.23 | 98.6 |
| **Total** | **103.54** | **8345.85** | **1.3±0.2** | **23±7** | **16.56** | **1.91** | **93.5** |

Table 1: Trajectory and gaze statistics by scenario (mean ± std). Dispersion is the root-mean-square distance of gaze points from their centroid (in pixels); higher values indicate more scattered attention. Saccade captures how frequently gaze fixation shifts occur[1]. On people (%) denotes the percentage of fixations that fall on pedestrians.

(a) Pupil Neon Eye Tracking Glasses (b) UWB Positioning System

Figure 2: Data collection equipment. (a) Eye-tracking glasses capture gaze data and egocentric RGB video. (b) UWB positioning system records participant trajectories.

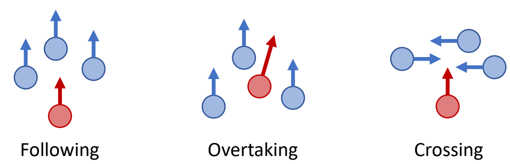


Figure 3: Illustration on Following, Overtaking & Crossing.

setup. This included Pupil Neon eye-tracking glasses for egocentric RGB video (1600x1200, 30Hz) and gaze data (200Hz) (Drews et al. 2024), and an ultra-wideband (UWB) system to track the participant's 2D position (200Hz, 10 cm accuracy) (see Figure 2). All data was temporally synchronized and downsampled to 5Hz.

To elicit diverse behaviors, we collected data across five scenes (a university campus and four busy tourist streets) using three defined scenarios. Participants were instructed to either: 1) *Follow* a random pedestrian leisurely, 2) *Overtake* others in a crowded area, or 3) *Cross* a perpendicular flow of pedestrian traffic (see Figure 3).

## Description Statistics

We collected approximately 104 minutes of recording from four participants, covering a total distance of 8.35 km. A key feature of GazeNav is its high crowd density. Using Mask R-CNN (He et al. 2017), we detected an average of 23 pedestrians per frame (with accuracy of 97%), ensuring complex social interactions. As shown in Table 1, the scenarios exhibit distinct behaviors: *Overtaking* has the highest average speed (1.6 m/s), reflecting assertive motion. *Following* maintains a moderate speed (1.2 m/s), while *Crossing* is the slowest (1.0 m/s), indicating cautious navigation across traffic.

## Gaze Analysis

Previous studies show that gaze guides human locomotion (Joshi et al. 2021). We analyzed its characteristics to understand how it supports crowd navigation.

*Spatial Temporal Characteristics*. As shown in Table 1, *Following* exhibits the lowest gaze dispersion and saccade frequency, reflecting stable attention on a lead pedestrian. In contrast, *Crossing* shows the highest dispersion and most frequent saccades, indicating rapid environmental scanning. *Overtaking* has moderate values, balancing forward tracking and peripheral monitoring.

*Semantic Characteristics*. To understand what people are looking at, we used Mask R-CNN (He et al. 2017) to identify objects at gaze locations. Across all scenarios, participants spent 88% to 98% of their gaze time on other pedestrians (see last column Table 1). This confirms that pedestrians are the most critical semantic objects for navigation, proving that socially-aware models must prioritize them.

*Coupling Gaze with Motion*. To explore how attention influences action, we analyzed how gaze on different pedestrian types affects the navigator's subsequent motion, (i.e., heading angle change and acceleration). We classified pedestrians by proximity (near/far) and movement direction (same/opposite/lateral). Figure 4 plots the distributions for heading angle change (red) and acceleration (green), conditioned on the context of the pedestrian the participant was looking at.

In *Following*, gaze is used to maintain a safe distance. Fixating on a nearby, same-direction pedestrian leads to slowing down to maintain distance, while fixating on an oncoming person triggers strong avoidance maneuvers. In *Overtaking*, gaze is used to identify opportunities to advance. Fixating on any nearby pedestrian often precedes a change in direction as the navigator actively seeks open space to pass. Conversely, when not fixating on any nearby pedestrians, participants maintain a steady, straight trajectory. In *Crossing*, gaze is used for collision avoidance and social compliance. Participants tend to veer in the opposite direction of a fixated pedestrian's movement (e.g., turning left when looking at a rightward-moving person, and vise versa) with slower speed, allowing them pass behind moving pedestrians rather than intercepting their path.

[1]Fixations are defined as stabilized gaze lasting $\geq$ 70 ms, while saccades are fast shifts $\geq 1.0^\circ$ and $\geq$ 10 ms.

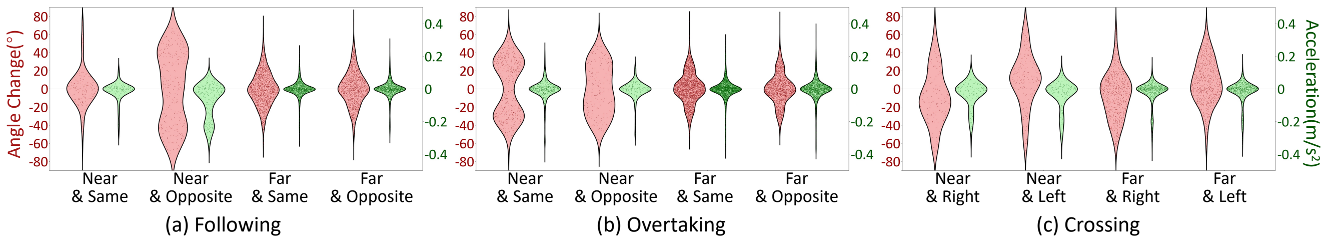


Figure 4: Violin plots on distribution of participant trajectory Angle Change (red) and Acceleration (green), conditioned by the attributes of the fixated pedestrian: proximity (Near/Far), relative direction (Same/Opposite), and lateral motion (Right/Left). Each dot represents a data point. (a) In Following, participants significantly changed their direction and slowed down after fixating on pedestrians walking toward them. (b) In Overtaking, fixating on any nearby pedestrian prompted significant changes in direction. (c) In Crossing, participants tended to veer in the opposite direction of a fixated pedestrian's movement (e.g., turning left (negative angle change) when looking at a rightward-moving person).

Our analysis is the first to investigate gaze in real-world crowd navigation, and uncover systematic, task-dependent links between gaze and motion. These findings provide the empirical foundation for our Gaze2Nav framework, which is designed to replicate this human-like process of using gaze to identify salient pedestrians and guide motion planning.

## Method

Inspired by our findings, we propose **Gaze2Nav**, a gaze-augmented egocentric social navigation framework that mimics human decision-making by coupling attention-guided perception with socially-aware planning.

### Method Overview

Our method, Gaze2Nav, generates human-like navigation actions by imitating human gaze and subsequent decision-making. As illustrated in Figure 5, the framework is trained via behavior cloning and consists of three sequential modules: a Gaze Predictor, a Semantic Saliency Matching module, and a Motion Planner. First, Gaze Predictor forecasts human social attention from historical video frames and gaze heatmaps. Next, the Semantic Saliency Matching module identifies socially salient pedestrians by spatially aligning the predicted gaze with instance-level pedestrian masks. Finally, Motion Planner uses the current frame and these salient semantic masks to generate the future trajectory. This modular design intentionally separates low-level perception (gaze prediction) from high-level planning (action generation), mirroring the human cognitive process of first looking to gather semantic saliency, then deciding where to go (Joshi et al. 2021; Hollands et al. 2002).

### Gaze Predictor

Though rich with social cues, raw human gaze data is inherently scene-specific and cannot directly guide a robot in novel environments where this information is unavailable. To address this, we propose a Gaze Predictor to learn the underlying patterns of human social attention from our dataset. By doing so, it can forecast gaze locations in real-time for any given scene, providing the downstream navigation planner with crucial social attention.

Gaze Predictor forecasts the gaze coordinate $\hat{g}_t$ at time step $t$, with input of a sequence of $T$ historical video RGB frames $\{I_i\}_{i\in\{t-T,\cdots,t-1\}}$, together with current frame $I_t$, and their corresponding historical gaze heatmaps $\{G_i\}_{i\in\{t-T,\cdots,t-1\}}$. The raw gaze data consists of 2D coordinates $g_i = (x, y)_i$, is rendered into a heatmap $G_i$ using a 2D Gaussian kernel ($\sigma = 10$ pixels) to align spatially with the image frames. The frames are processed by an EfficientNet-B0 backbone (Tan et al. 2019), while the gaze heatmaps are processed by a lightweight 3-layer CNN. Both visual and gaze features are passed into separate Transformer encoders. This step models the spatiotemporal dependencies within each modality. The last MLP layer outputs the predicted gaze coordinates $\hat{g}_t \in [0, 1]^2$ for time $t$.

**Loss**. Gaze Predictor optimizes for two goals: positional accuracy and behavioral alignment with the navigation task. First, we supervise with a standard gaze coordinate regression loss using the L2 distance: $\mathcal{L}_{\text{coord}} = \|\hat{g}_t - g_t\|_2$.

Second, we argue that a purely historical model is insufficient. Human gaze is not merely a continuation of past movements; it is also proactive and tightly coupled with future navigational intent. For instance, the attentional pattern required to aggressively overtake a group is fundamentally different from that of leisurely following them. A good gaze prediction must therefore be consistent with the intended future action. To enforce this behavioral consistency, we introduce an attention alignment loss as a regularizer, encouraging the model attention from Gaze Predictor's Transformer ($A_t^{\text{gaze}}$) to align with the that from Motion Planner's Transformer ($A_t^{\text{act}}$). We use Kullback-Leibler (KL) divergence to measure the similarity between the two attention maps: $\mathcal{L}_{\text{a2g-align}} = \text{KL}(A_t^{\text{act}} \| A_t^{\text{gaze}})$.

The final loss for Gaze Predictor is a weighted sum of these two components:

$$\mathcal{L}_{\text{gaze}} = (1-\alpha)\|\hat{g}_t - g_t\|_2 + \alpha \cdot \text{KL}(A_t^{\text{act}} \| A_t^{\text{gaze}}) \quad (1)$$

where $\alpha \in [0, 1]$ trades off between gaze prediction accuracy and learning motion-consistent attention.

### Semantic Saliency Matching

Raw gaze coordinates are often too low-level and ambiguous for direct use, as they can be noisy or misdirected. Our Se-

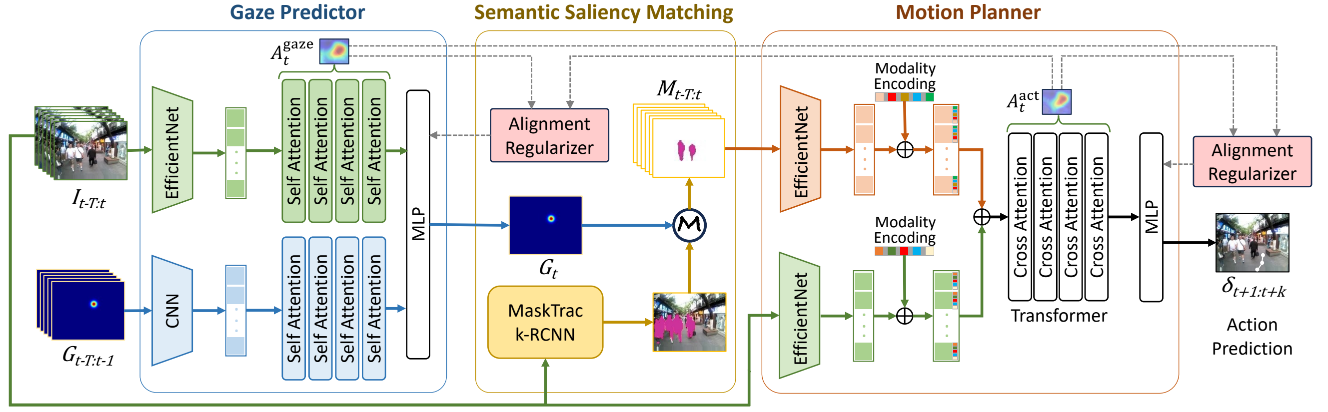


Figure 5: Architecture of Gaze2Nav. The framework consists of a Gaze Predictor to encode and forecast human gaze, a Semantic Saliency Matching module to identify salient pedestrians, and a Motion Planner to generate navigation actions. Attention alignment losses (gray dashed arrows) regularize consistency between the gaze and motion modules.

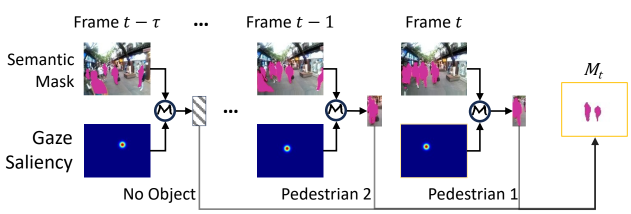


Figure 6: Semantic Saliency Matching. This module outputs a saliency mask $M_t$, by aggregating the masks of all pedestrians considered salient. A pedestrian is marked salient if predicted gaze point falls on their mask or if they were attended to within a recent time window $[t-\tau, t-1]$.

mantic Saliency Matching module translates them into high-level social understanding. By spatially aligning predicted gaze with detected pedestrians, the module identifies salient individuals, providing Motion Planner with a more robust and interpretable input than a simple gaze heatmap.

Furthermore, human motion relies on more than just momentary glances; it involves short-term memory of relevant individuals (Posner 1980). To emulate this, we introduce a temporal matching mechanism that aggregates salient pedestrian information over a short sliding window. This creates a stable and continuous attention signal, ensuring that key pedestrians remain "in focus" for the planner even if the gaze momentarily shifts elsewhere.

Our matching process begins by tracking pedestrian masks with MaskTrack R-CNN (Yang et al. 2019). As illustrated in Figure 6, a pedestrian is marked as "salient" at time step $t$ if either of two conditions is met: (1) the current predicted gaze point $\hat{p}_t$ falls within their segmentation mask, or (2) they were marked as salient within a recent temporal window $[t-\tau, t-1]$ and are still visible in the current frame. The masks of all salient pedestrians are then aggregated into a single binary saliency map, $M_t \in \{0,1\}^{H\times W}$, providing a robust, social-centric attention signal to Motion Planner.

## Motion Planner

Our Motion Planner learns navigation actions from human demonstrations via behavior cloning, using the Visual Navigation Transformer (ViNT) (Shah et al. 2023b) as its backbone. Unlike the original ViNT architecture, our Motion Planner takes a history of video frames $\{I_i\}_{i\in\{t-T,\cdots,t\}}$ and corresponding saliency masks $\{M_i\}_{i\in\{t-T,\cdots,t\}}$ as input to predict a future trajectory $\{\hat{l}_{t+1},\ldots,\hat{l}_{t+k}\}$. It enables ego-centric navigation without a global positioning system.

The network uses two EfficientNet encoders to extract features from the video frames and saliency masks. To distinguish these inputs, a unique modality encoding vector is added to each feature map. The features are then tokenized, concatenated, and processed by a Transformer that models complex interactions. An MLP head then decodes the Transformer's representation into the final predicted trajectory.

**Loss.** Motion Planner is trained with a dual-objective loss that balances accurate motion prediction with human-like visual attention. The first objective is a regression loss defined as the mean squared error (MSE) between the predicted and ground-truth trajectory over future $k$ time steps: $\mathcal{L}_{\text{traj}} = \frac{1}{k}\sum_{i=1}^{k}\|\hat{l}_{t+i} - l_{t+i}\|_2^2$.

While trajectory loss ensures geometric accuracy, it doesn't guarantee attention to social cues. To encourage more human-aligned behavior, we introduce an attention alignment term that regularizes Motion Planner's attention map ($A_t^{\text{act}}$) to match Gaze Predictor's $A_t^{\text{gaze}}$. This promotes perceptual consistency and encourages the planner to use human-like cues. The alignment loss is computed as: $\mathcal{L}_{\text{g2a-align}} = \text{KL}(A_t^{\text{gaze}}\|A_t^{\text{act}})$.

The final loss of motion planning at time $t$ is a weighted sum of both objectives:

$$\mathcal{L}_{\text{act}} = (1-\beta)\frac{1}{k}\sum_{i=1}^{k}\|\hat{l}_{t+i}-l_{t+i}\|_2^2+\beta\cdot\text{KL}(A_t^{\text{gaze}}\|A_t^{\text{act}}) \quad (2)$$

where $\beta \in [0,1]$ controls the trade-off between reproducing expert-like motion and learning gaze-consistent attention.

| Model | ADE ↓ | FDE ↓ | Fréchet ↓ | Cos Sim ↑ |
|---|---|---|---|---|
| GNM | 0.4139 | 0.7570 | 0.8244 | 0.8482 |
| GNM+CGL | 0.3775 | 0.7320 | 0.7965 | 0.8673 |
| GNM+Foveated | 0.3850 | 0.7241 | 0.8078 | 0.8792 |
| ViNT | 0.3160 | 0.7220 | 0.7932 | 0.8584 |
| ViNT+CGL | 0.3005 | 0.6994 | 0.7774 | 0.8731 |
| ViNT+Foveated | 0.3011 | 0.7034 | 0.7801 | 0.8799 |
| w/o Semantic | 0.2946 | 0.7213 | 0.7567 | 0.8946 |
| w/o Gaze | 0.2861 | 0.7059 | 0.7533 | 0.8846 |
| w/o GazeReg | 0.2768 | 0.6712 | 0.7349 | 0.9012 |
| w/o MotionReg | 0.2653 | 0.6704 | 0.7295 | 0.8971 |
| SalientMaskPred | 0.2915 | 0.6806 | 0.7630 | 0.8993 |
| **Gaze2Nav** | **0.2541** | **0.6536** | **0.7124** | **0.9150** |

Table 2: Trajectory performance across metrics. ↓ indicates lower is better; ↑ indicates higher is better.

| Model | Following | Overtaking | Crossing |
|---|---|---|---|
| GNM | 0.3296 | 0.5046 | 0.3151 |
| GNM+CGL | 0.3064 | 0.4512 | 0.3032 |
| GNM+Foveated | 0.2945 | 0.4633 | 0.3247 |
| ViNT | 0.2608 | 0.3750 | 0.2438 |
| ViNT+CGL | 0.2415 | 0.3601 | 0.2345 |
| ViNT+Foveated | 0.2398 | 0.3632 | 0.2322 |
| **Gaze2Nav** | **0.2289** | **0.2908** | **0.1910** |

Table 3: ADE↓ across three different navigation behaviors.

## Experiments

This section evaluates the Gaze2Nav framework on gaze and trajectory prediction. We present quantitative and qualitative results to highlight the benefits of gaze-guided saliency.

### Experimental Settings

**Baselines** We compare Gaze2Nav against two vision-only behavior cloning based navigation network backbones and two existing gaze-integration techniques.

- Network Backbones: *GNM* which is a lightweight CNN (Shah et al. 2023a) and *ViNT* which is a Vision Transformer (Shah et al. 2023b) we also adopt.
- Gaze-Integration Methods: *CGL* which adopts KL divergence loss to align feature maps with gaze heatmaps (Saran et al. 2021) and *Foveated* which simulates foveal vision by blurring periphery of input images around gaze point (Zhang et al. 2020). Both applied to GNM and ViNT for fair comparison.

To validate our design, we test several ablations:

- *w/o Gaze* lets Motion Planner receive masks for all detected pedestrians, with no gaze-based saliency filtering.
- *w/o Semantic* directly feeds predicted gaze heatmaps into Motion Planner without Semantic Saliency Matching.
- *w/o GazeReg* removes the gaze regularization loss $\mathcal{L}_{\text{g2a-align}}$ in Motion Planner.
- *w/o MotionReg* removes the motion regularization loss $\mathcal{L}_{\text{a2g-align}}$ in Gaze Predictor.

| Model | Gaze MAE ↓ | Saliency F1 Score ↑ |
|---|---|---|
| w/o MotionReg | 5.969 | 0.8556 |
| SalientMaskPred | – | 0.5739 |
| **Gaze2Nav** | **3.876** | **0.8755** |

Table 4: Accuracy of gaze and semantic saliency prediction.

- *SalientMaskPred* is trained to directly predict the salient pedestrian mask, which is then fed to Motion Planner.

**Metrics** We evaluate trajectory quality by:

- *Average Displacement Error (ADE):* The average L2 distance between the predicted future positions and the actual ground-truth positions over future $k$ time steps.
- *Final Displacement Error (FDE):* The L2 distance between the *final* predicted position and the final ground-truth position at time step $t + k$.
- *Fréchet Distance* captures the overall shape similarity between the predicted and ground-truth trajectories, accounting for spatiotemporal alignment.
- *Cosine Similarity* evaluates the average angular similarity between trajectory vectors of $k$ steps, assessing whether the motion direction aligns with humans.

We also evaluate the accuracy of gaze prediction by

- *Gaze Prediction MAE* measures the average L1 distance between predicted and ground-truth gaze coordinates.
- *Saliency F1 Score* calculates the F1 score for identifying the correct salient pedestrian mask.

**Implementation** We use a history of $T = 5$ frames to predict $k = 2$ future steps (a 1.2-second horizon for real-time robot control). The dataset is split 80/20 training/testing at the trajectory level, i.e., 80% of trajectories for training and 20% for testing. Both Gaze Predictor and Motion Planner are trained separately for 200 epochs using the AdamW optimizer with a batch size of 64, an initial learning rate of $1 \times 10^{-4}$, and gradient clipping with a max norm of 1.0. The learning rate follows a cosine annealing schedule with a 4-epoch warm-up. The hyperparameter $\alpha$ and $\beta$ is tuned using a grid search. All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU.

### Comparison Results

**Motion Planning Performance.** As shown in Table 2, Gaze2Nav outperforms all baselines across every metric, demonstrating superior trajectory accuracy, shape similarity, and directional alignment. Compared to the strongest gaze-augmented baseline (ViNT+CGL), Gaze2Nav reduces ADE by 15.4% and Fréchet Distance by 8.4%, proving a more effective use of gaze information. This strong performance holds across all navigation scenarios (Table 3), with Gaze2Nav achieving the lowest ADE in all three behaviors. The improvements are most significant in the challenging Overtaking (19% improvement) and Crossing (18% improvement) scenarios, where its ability to capture gaze cues enables proactive decision-making.

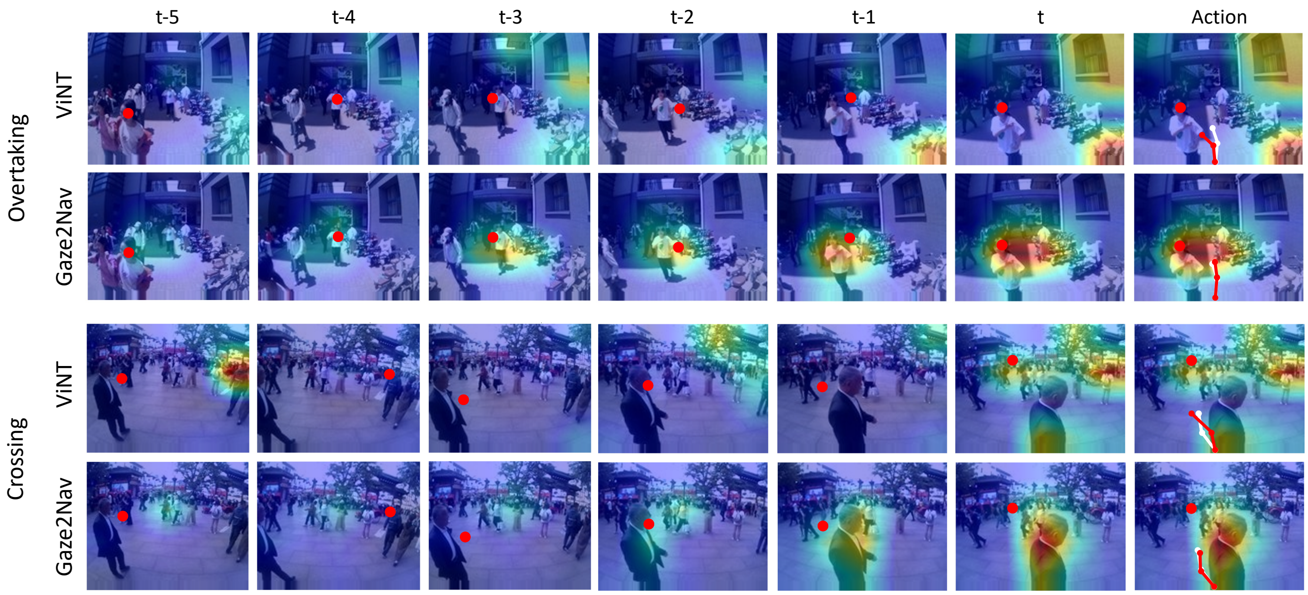


Figure 7: Case study of attention alignment in two challenging navigation scenarios Overtaking and Crossing. The attention maps from Motion Planner are visualized as heat maps, with red dots of human gaze points, and planned trajectories (red) comparing to the ground truth (white). Compared to the baseline ViNT, Gaze2Nav produces attention maps that more closely align with human gaze, avoid irrelevant image regions (e.g., borders), and lead to more accurate action predictions.

While existing gaze-enhanced baselines like GNM+CGL and ViNT+Foveated, improve upon their vanilla counterparts, they still underperform compared to Gaze2Nav. This suggests that merely providing gaze data is insufficient; Gaze2Nav's superior performance stems from its fusion of gaze with semantic information and the alignment of attention between its perception and planning modules, as validated in the following ablation study.

**Gaze Integration Ablation.** To isolate the contributions of our core mechanisms, we compare Gaze2Nav with several ablated variants. (a) Semantic and Gaze Input: Removing semantic matching (*w/o Semantic*) or the entire gaze-based filtering mechanism (*w/o Gaze*) degrades performance across all metrics (Table 2). This highlights that not all pedestrian information is equally useful; gaze is critical for focusing the model's attention on socially relevant pedestrians and preventing suboptimal planning. (b) Gaze Prediction Strategy: We also tested a model that directly predicts the salient pedestrian mask (*SalientMaskPred*) without first predicting a gaze point. This approach proved far less effective, achieving a much lower Saliency F1 score (Table 4) and worse trajectory accuracy. This result validates our cognitively-inspired strategy: first predict where to look, then identify who to attend to. (c) Attention Alignment Loss: Removing either the motion regularization from Gaze Predictor (*w/o MotionReg*) or the gaze regularization from Motion Planner (*w/o GazeReg*) hurts performance (Table 2). Aligning motion attention improves gaze prediction accuracy, which in turn benefits planning. Conversely, aligning gaze attention helps the planner focus on the most relevant cues, improving trajectory accuracy.

### Case Study

We further look at how gaze-informed semantic grounding enables the model to avoid distractions, focus on socially relevant agents, and produce more accurate, human-like navigation behavior on two challenging scenarios, Overtaking and Crossing. Figure 7 visualizes the attention maps from Motion Planner for Gaze2Nav and the baseline ViNT, comparing the model's focus and predicted path (red) with the human's gaze (red dot) and ground-truth trajectory (white). 1) Overtaking: When facing oncoming pedestrians, the ViNT baseline gets distracted by irrelevant image borders. In contrast, Gaze2Nav correctly focuses on the incoming people, identifies collision risks, and executes a smooth avoidance maneuver that matches the human's path. 2) Crossing: As a pedestrian crosses, ViNT's attention is too diffuse, causing a delayed and misaligned reaction. Gaze2Nav, however, sharply focuses on pedestrians and performs timely and precise evasive actions aligned with the ground truth.

## Conclusion

In this work, we present GazeNav, a novel egocentric dataset with synchronized gaze and trajectory data. Our analysis shows that human gaze is a critical social cue, which led us to develop Gaze2Nav, a framework that mimics human cognition by using predicted gaze to identify salient pedestrians and guide motion planning. Gaze2Nav achieves 87.6% accuracy in predicting salient pedestrians and reduces trajectory error by 15.4% over state-of-the-art methods, enhancing both performance and interpretability. Future work will expand the dataset, include user studies on human-likeness and trust, and deploy the system on physical robots.

## Acknowledgments

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (Award No: DTC-RGC-09), the Natural Science Foundation of China under Grant 62232004, 62572120, 62472093, 62572246, the Natural Science Foundation of Jiangsu Province under Grant No.BK20230024 and No.BE2023799, the Fundamental Research Funds for the Central Universities, the Royal Society Short Industry Fellowship (5199753/251064), and EPSRC/Henry Royce Institute (4939265). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

## References

Bae, J. W.; Kim, J.; Yun, J.; Kang, C.; Choi, J.; Kim, C.; Lee, J.; Choi, J.; and Choi, J. W. 2023. Sit dataset: socially interactive pedestrian trajectory dataset for social navigation robots. *Advances in neural information processing systems*, 36: 24552–24563.

Chen, H.-Y.; Huang, P.-H.; and Fu, L.-C. 2023. Social crowd navigation of a mobile robot based on human trajectory prediction and hybrid sensing. *Autonomous robots*, 47(4): 339–351.

Curtis, S.; and Manocha, D. 2014. Pedestrian simulation using geometric reasoning in velocity space. In *Pedestrian and evacuation dynamics 2012*, 875–890. Springer.

Drews, M.; and Dierkes, K. 2024. Strategies for enhancing automatic fixation detection in head-mounted eye tracking. *Behavior Research Methods*, 56(6): 6276–6298.

Everett, M.; Chen, Y. F.; and How, J. P. 2021. Collision avoidance in pedestrian-rich environments with deep reinforcement learning. *Ieee Access*, 9: 10357–10377.

Guo, S. S.; Zhang, R.; Liu, B.; Zhu, Y.; Ballard, D.; Hayhoe, M.; and Stone, P. 2021. Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 34: 25370–25385.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2255–2264.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Helbing, D.; and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E*, 51(5): 4282.

Hollands, M. A.; Patla, A. E.; and Vickers, J. N. 2002. "Look where you're going!": gaze behaviour associated with maintaining and changing the direction of locomotion. *Experimental brain research*, 143(2): 221–230.

Huang, Y.; Cai, M.; Li, Z.; and Sato, Y. 2018. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, 754–769.

Joshi, H. B.; Cybis, W.; Kehayia, E.; Archambault, P. S.; and Lamontagne, A. 2021. Gaze behavior during pedestrian interactions in a community environment: a real-world perspective. *Experimental brain research*, 239(7): 2317–2330.

Karnan, H.; Nair, A.; Xiao, X.; Warnell, G.; Pirk, S.; Toshev, A.; Hart, J.; Biswas, J.; and Stone, P. 2022. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4): 11807–11814.

Kretzschmar, H.; Spies, M.; Sprunk, C.; and Burgard, W. 2016. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11): 1289–1307.

Lai, D.; Zhang, Y.; Liu, Y.; Li, C.; and Mo, H. 2025. Deep Learning-Based Multi-Modal Fusion for Robust Robot Perception and Navigation. *arXiv preprint arXiv:2504.19002*.

Li, X.; Qiu, H.; Wang, L.; Zhang, H.; Qi, C.; Han, L.; Xiong, H.; and Li, H. 2025. Challenges and Trends in Egocentric Vision: A Survey. *arXiv preprint arXiv:2503.15275*.

Ling, B.; Lyu, Y.; Li, D.; Gao, G.; Shi, Y.; Xu, X.; and Wu, W. 2024. Socialgail: Faithful crowd simulation for social robot navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 16873–16880. IEEE.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14424–14432.

Nguyen, D. M.; Nazeri, M.; Payandeh, A.; Datar, A.; and Xiao, X. 2023. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7442–7447. IEEE.

Posner, M. I. 1980. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1): 3–25.

Rodin, I.; Furnari, A.; Mavroeidis, D.; and Farinella, G. M. 2021. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211: 103252.

Saran, A.; Zhang, R.; Short, E. S.; and Niekum, S. 2021. Efficiently Guiding Imitation Learning Agents with Human Gaze. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 1109–1117.

Shah, D.; Sridhar, A.; Bhorkar, A.; Hirose, N.; and Levine, S. 2023a. GNM: A General Navigation Model to Drive Any Robot. In *International Conference on Robotics and Automation (ICRA)*.

Shah, D.; Sridhar, A.; Dashora, N.; Stachowicz, K.; Black, K.; Hirose, N.; and Levine, S. 2023b. ViNT: A Foundation Model for Visual Navigation. In *7th Annual Conference on Robot Learning*.

Shen, Y.; Ni, B.; Li, Z.; and Zhuang, N. 2018. Egocentric activity prediction via event modulated attention. In *Proceedings of the European conference on computer vision (ECCV)*, 197–212.

Sridhar, A.; Shah, D.; Glossop, C.; and Levine, S. 2024. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 63–70. IEEE.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tavakoli, H. R.; Rahtu, E.; Kannala, J.; and Borji, A. 2019. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 273–282. IEEE.

Van Den Berg, J.; Guy, S. J.; Lin, M.; and Manocha, D. 2010. Optimal reciprocal collision avoidance for multi-agent navigation. In *Proc. of the IEEE International Conference on Robotics and Automation, Anchorage (AK), USA*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5188–5197.

Yi, S.; Li, H.; and Wang, X. 2015. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3488–3496.

Zhang, R.; Walshe, C.; Liu, Z.; Guan, L.; Muller, K.; Whritner, J.; Zhang, L.; Hayhoe, M.; and Ballard, D. 2020. Atari-head: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6811–6820.

Zhou, B.; Wang, X.; and Tang, X. 2011. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *CVPR 2011*, 3441–3448. IEEE.