EgoCogNav: Cognition-aware Human Egocentric Navigation

Zhiwen Qiu[1], Ziang Liu[1], Wenqian Niu[2], Tapomayukh Bhattacharjee[1], and Saleh Kalantari[1]

[1] Cornell University, Ithaca, NY, USA
[2] Georgia Institute of Technology, Atlanta, GA, USA

Abstract. Modeling the cognitive and experiential factors of human navigation is central to deepening our understanding of human–environment interaction and to enabling safe social navigation and effective assistive wayfinding. Most existing methods focus on forecasting motions in fully observed scenes and often neglect human factors that capture how people feel and respond to space. To address this gap, we propose EgoCogNav, a multimodal egocentric navigation framework that jointly forecasts perceived path uncertainty, trajectories and head motion from egocentric video, gaze, and motion history. To facilitate research in the field, we introduce the Cognition-aware Egocentric Navigation (CEN) dataset consisting of 6 hours real-world egocentric recordings capturing diverse navigation behaviors in real-world scenarios. Experiments show that EgoCogNav learns the perceived uncertainty that strongly correlates with human-like behaviors such as scanning, hesitation, and backtracking while improving trajectory and head-motion forecasting on held-out navigation recordings. Project page: `https://calvinzqiu.github.io/egocognav-project/`

Keywords: Egocentric Navigation · Trajectory Prediction · Perceived Uncertainty · Multimodal Learning

1 Introduction

Understanding human navigation processes is fundamental to safe, reliable human-environment and human-machine interactions [17, 52]. Humans perceive their environments from egocentric perspectives to inform decision-making for subsequent motions. Learning such cues is crucial for applications including autonomous driving [24,61], social robot navigation [41,54,62], and assistive wayfinding systems [13, 47, 67].

Modeling human navigation requires not only accurate scene representation and motion forecasting, but also human factors that capture how people experience and emotionally respond to built environments, which are closely related to productivity, satisfaction, and well-being [12, 26, 37, 51]. Incorporating these cognitive states into behavioral models can deepen our understanding of interacting with environments and substantially enhance their utility for salient

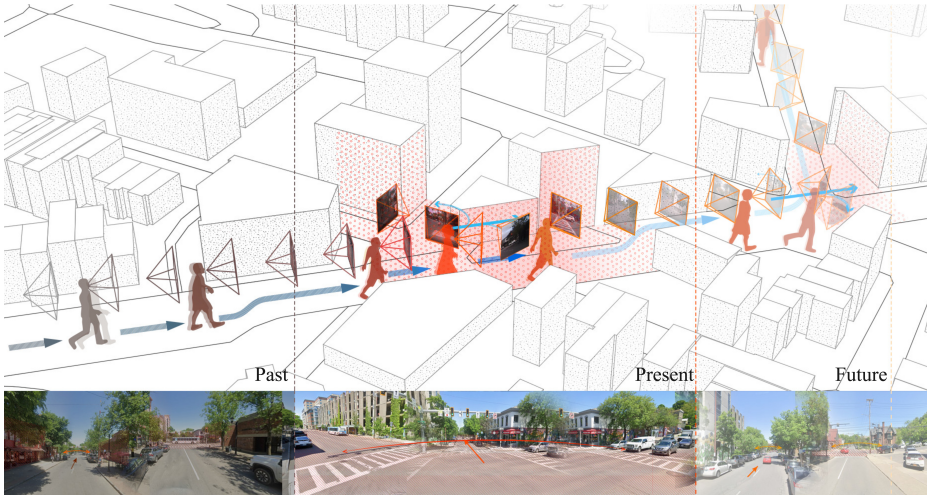

Fig. 1: Given a past window of motion, head rotations, gaze, and navigational goal, our model jointly predicts a future body-frame trajectory, head poses, and the current state of perceived uncertainty. This setting reflects real-world navigation challenges in environments where the model must anticipate internal cognitive state and making decisions for subsequent head motion and movement.

environmental design and personalized navigation assistance in complex settings [11, 26]. Perceived uncertainty, defined as a state in which an individual is trying to decide between alternative courses of action, is a driver of wayfinding behaviors tied to discomfort and negative emotions [11, 22, 49]. Predicted uncertainty levels can inform probabilistic wayfinding behaviors as well as a valuable metric for understanding users' experience of a space. However, most prior work in human trajectory prediction does not account for cognitive and psychological processes and relies instead on motion history and context for path forecasting [19, 25, 39], use rule- or agent-based methods [10, 40, 73], or develop socially compliant policies for navigation [21, 29, 43, 58]. Although some incorporate cognitive factors such as emotions, panic, and stress, these frameworks typically assume fully observed third-person or bird's-eye-view (BEV) scenes and often planar worlds [6,13,47,67]. Furthermore, there are limited multimodal egocentric navigation datasets, especially those with cognitive annotations, to study these effects at scale.

To address the challenges, we propose EgoCogNav, a multimodal egocentric navigation framework that jointly predicts perceived uncertainty, body-centered trajectories and head motion. The framework fuses first-person scene evidence from a pre-trained DINOv2 vision backbone [8,45] with recent motion where cognition and behaviors are learned in a single perception–decision–action loop. The design is modular and extensible for additional modality input with shared time-series forecasting module. Finally, we introduce the Cognition-Aware Egocentric Navigation (CEN) dataset that features rich multimodal streams from 17 partic-

ipants across diverse indoor and outdoor scenes for human navigation research. Our contribution can be summarized as follows: (1) We formalize cognition-aware egocentric forecasting task to jointly predict trajectory, head motion, and moment-to-moment perceived path uncertainty; (2) We propose EgoCogNav, which fuses multiple sensory inputs with human-grounded uncertainty to produce behaviorally realistic behavior forecasts useful for assistive navigation; (3) We introduce a dataset consisting of 6 hours of real-world human navigation sessions that covers diverse sites and environmental conditions.

2 Related Work

Human trajectory prediction. The prospect of predicting pedestrian trajectories from third-person or BEV has been extensively studied. Prior methods include deterministic predictors that estimate one most likely path using scene information and social cues [2, 28, 53], and stochastic models that generate multimodal futures or full distributions from probabilistic frameworks such as diffusion models [4, 19, 39] and normalizing flows [38, 56]. However, third-person views often neglect first-person natural perceptual and cognitive signals, which can limit the development of high-fidelity modeling of human motion in cluttered environments [42]. Despite substantial progress in egocentric human motion estimation [1,30,60,68], trajectory and motion forecasting that incorporate human cognitive states remain underexplored. Wang and colleagues [64] use a diffusion framework to generate multiple plausible future trajectories from RGB-D video with a learned visual-memory of the surroundings. Pan and colleagues [46] predict 6-DoF head poses to learn collision-aware, information-seeking behaviors by projecting 3D-feature volumes into BEV. However, these approaches largely assume homogeneous behaviors and overlook individual differences and internal cognitive states that are salient in real-world navigation. Integrating experiential factors can reveal why and when people hesitate or backtrack to improve behavioral fidelity and enable assistance and design interventions that anticipate difficulties rather than merely reacting to observed motion.

Perceived uncertainty in human wayfinding models. Perceived path uncertainty is a key cognitive variable closely linked to wayfinding performance which emerges from limited knowledge about forthcoming events [11, 49]. The Entropy Model of Uncertainty (EMU) [22] posits that uncertainty is affected by the range of perceived choices and peaks when perceived probability for each choice is equal. Prior strategies to integrate human cognition typically include translating empirical observations like "go-to" affordances into rule-based deterministic models [51], or defining utility parameters such as path choice and individual preference for agents to optimize during navigation [23, 65, 72]. For instance, Huang and colleagues [23] present a two-layer floor-field cellular automaton with three intertwined sub-modules (exit choice, locomotion, and exit-choice switching) to model risk- and uncertainty-aware decisions to capture pedestrian dynamics. Yang and colleagues [67] introduce a probabilistic data-driven wayfinding agent that integrates isovist-derived spatial metrics, signage

visibility/recency, route-choice counts, and individual factors to predict continuous perceived path uncertainty for simulating human trajectories in indoor settings. In contrast to prior BEV- or rule-based formulations, we employ a data-driven and learning-based method that predicts perceived path uncertainty from multimodal first-person cues.

Memory-augmented motion prediction. Recent works have explored memory mechanisms to extend the effective context of motion prediction models. Xu and colleagues [66] introduce an instance memory bank (MemoNet) that stores and retrieves past trajectory patterns to improve pedestrian forecasting. Shi and colleagues [57] introduces MTR++ that utilizes learnable intention queries that capture recurring motion patterns across agents that provides a form of parametric memory that the decoder attends to during prediction. Similarly, Yu and colleagues [69] use learnable trajectory anchors as reference patterns that encode frequently observed movement modes, enabling the decoder to attend to relevant anchors and produce diverse yet plausible future trajectories. In the generative domain, Zhang and colleagues [70] augment diffusion-based motion generation (ReMoDiffuse) with a retrieval mechanism that queries a database of motion clips. While these approaches demonstrate the value of memory for motion forecasting, few focus on conditioning memory retrieval or processing on predicted human cognition state.

Egocentric dataset for human navigation. Many egocentric datasets [9, 18, 31, 35] provide real-world monocular RGB videos of diverse daily activities such as household, workplace, and outdoor scenes. To enrich human sensory inputs, subsequent efforts were made to incorporate wearable and instrumented platforms like motion-capture suits and Project Aria glasses [15] to deliver calibrated multimodal signals, including head poses [46, 63], gaze [32, 35], hand and object tracking [5, 59], full-body motion [36], and detailed scene and action annotations [33]. However, these datasets are not explicitly curated for human navigation. Few works utilized synthetic pipelines to simulate virtual human navigation and collisions in 3D environments from multi-view body-mounted cameras, but resulting motions are often simplified or unnatural due to constraints in existing motion generation methods. Closer to our setting, Wang and colleagues [64] and Pan and colleagues [46] develop egocentric human-navigation datasets, yet the coverage is comparatively limited and the focus is primarily trajectory or head-pose forecasting without leveraging varied sensory inputs or human cognitive factors. Moreover, neither dataset is publicly released to date. Therefore, we introduce a new dataset that unifies sensory signals, cognitive indicators, and accurate localization across diverse indoor and outdoor scenes.

3 Method

3.1 Problem Formulation

As illustrated in Fig. 1, given a past egocentric video $\mathbf{X}_{1:T_1} \in \mathbb{R}^{T_1 \times H \times W \times 3}$, past body-frame motion $\mathbf{S}_{1:T_1} = [\Delta x, \Delta y, \Delta \psi] \in \mathbb{R}^{T_1 \times 3}$, 6D continuous head

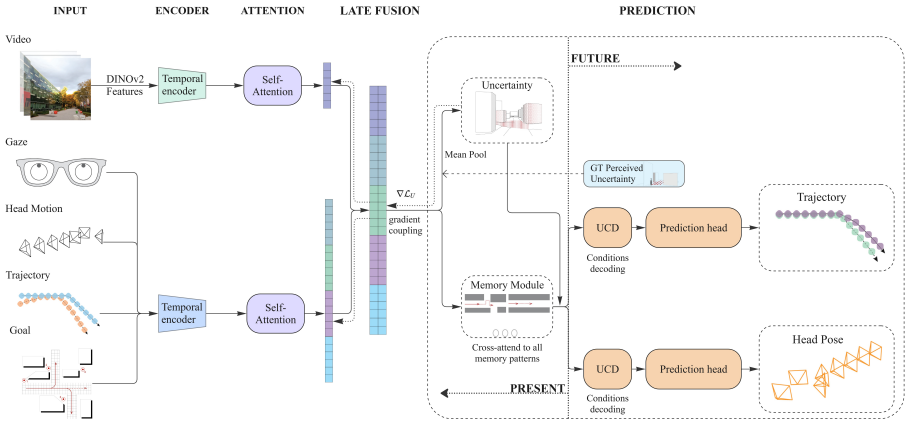

Fig. 2: EgoCogNav architecture. Given past egocentric video and sensor inputs, the model encodes perception (video features from a pre-trained vision transformer) and action (past motion, gaze, and goal) into two self-attention streams and fuses them by late concatenation. A cognition module predicts perceived uncertainty and retrieves situation-relevant context from learnable memory patterns, and then conditions decoding via adaptive layer normalization to forecast future trajectory and head motion.

rotations [71] $\mathbf{H}_{1:T_1} \in \mathbb{R}^{T_1 \times 6}$, normalized gaze points $\mathbf{G}_{1:T_1} \in \mathbb{R}^{T_1 \times 2}$, and body-frame navigation goal $\mathbf{q} = [d, \sin\beta, \cos\beta] \in \mathbb{R}^3$ where d is distance and β the goal bearing, we jointly predict the future trajectory $\widehat{\mathbf{S}}_{T_1+1:T_1+T_2} \in \mathbb{R}^{T_2 \times 3}$, head-pose sequence $\widehat{\mathbf{H}}_{T_1+1:T_1+T_2} \in \mathbb{R}^{T_2 \times 6}$, and the current perceived uncertainty $\hat{U}_{T_1} \in [0, 1]$. At 10 Hz, we use a past window $T_1 = 30$ steps (3 s) and future horizon $T_2 = 10$ steps (1 s) which matches the timescale of short-term navigation behaviors such as hesitation and rapid head reorientation [7, 27].

3.2 EgoCogNav Architecture

Our framework is organized into three modules (Fig. 2): (1) a perception module that extracts spatio-temporal features from recent video using a pre-trained vision transformer, (2) an action module that encodes past body-frame motion, head rotation, and gaze together with goal conditioning, and (3) a cognition module that predicts the current perceived uncertainty $\hat{U}_t$ and uses it to condition decoding via adaptive layer normalization and to augment features with learnable memory modules. The perception and action streams are encoded independently with self-attention and then fused through late concatenation into a shared representation from which we forecast body-frame trajectory and head motion, together with a cognition head that estimates perceived uncertainty.

Perception module. The perception module processes the past RGB frames. Each frame is resized to 224×224 and passed through a frozen pre-trained DINOv2 [8, 45] vision transformer. We use the per-frame `CLS` token as a global visual descriptor to produce a temporal stack of video features $\mathbf{F}^{\text{vid}} \in \mathbb{R}^{T_1 \times 384}$.

These features are linearly projected to a shared model dimension d and fed into the subsequent fusion module.

Action module. We encode three synchronized cues over the past T_1 steps: body-frame trajectory deltas $\mathbf{S}_{1:T_1} = [\Delta x, \Delta y, \Delta\psi]$, head rotations $\mathbf{H}_{1:T_1}$, and gaze points $\mathbf{G}_{1:T_1} = (u, v)$ in normalized image coordinates, together with the navigation goal $\mathbf{q} = [d, \sin\beta, \cos\beta]$ expressed in the current body frame. All streams are aligned with sinusoidal positional encoding and projected to the same width d before fusion.

Multi-modal fusion. We process the action and perception streams independently before fusion. The action stream concatenates body-frame motion, head rotation, gaze, and goal features into a single temporal sequence projecting to width d, and encodes it with a 4-layer transformer encoder using sinusoidal positional encoding. The perception stream similarly projects and encodes the DINOv2 feature sequence with a 2-layer transformer encoder. Both streams are temporally mean-pooled and concatenated to produce a fused representation $\mathbf{f} \in \mathbb{R}^{2d}$, which is then projected to $\mathbf{h} \in \mathbb{R}^d$ via a linear layer with layer normalization. This late fusion strategy lets each stream learn modality-specific temporal patterns through self-attention before combining them into a shared representation.

Cognition module. The cognition module is the core module of our architecture. It captures the internal cognitive state of the navigator and uses it to guide both how and what information the prediction heads use. It consists of three sub-components: (1) gradient-coupled uncertainty prediction, (2) memory-augmented prediction, and (3) uncertainty-conditioned decoding (UCD) as described below.

Gradient-coupled uncertainty estimation. Conceptually aligned with the EMU theory of competing scene interpretations and action choices [22], we model the navigator's current internal state as a single-step prediction of perceived uncertainty $\hat{U}_t \in [0, 1]$, a state signal that summarizes moment-to-moment choice difficulty preceding navigation behaviors. The cognition head operates on the projected fused representation $\mathbf{h} \in \mathbb{R}^d$ with a two-layer MLP and sigmoid activation:

$$\hat{U}_t = \sigma\big(\mathrm{MLP}(\mathbf{h})\big) \in [0, 1]. \tag{1}$$

Since $\hat{U}_t$ is predicted from shared encoder features, the task objectives jointly shape the encoder representation to encourage features that support both motion prediction and uncertainty estimation.

Memory-augmented prediction. While predicting $\hat{U}_t$ provides useful gradient signal to the encoder, it does not extend what information is available to the forecasting heads. Navigation under perceived uncertainty often requires context beyond the immediate T_1-step input window with patterns from similar past situations. Inspired by learnable intention queries in motion prediction [57], we augment the model with $N_m = 16$ learnable navigation pattern vectors $\mathbf{M} \in \mathbb{R}^{N_m \times d_m}$ that capture recurring navigation situations from the training data. The current navigation state $\mathbf{h}$ queries these patterns via cross-attention to retrieve

situation-relevant context:

$$\mathbf{c} = \mathrm{CrossAttn}\big(W_q \cdot \mathbf{h},\ \mathbf{M},\ \mathbf{M}\big) \tag{2}$$

$$\mathbf{h}^{\mathrm{mem}} = \mathbf{h} + W_{\mathrm{out}} \cdot \mathbf{c} \tag{3}$$

where W_q projects to the pattern space and W_{out} projects back, with W_{out} zero-initialized so the module starts as identity and gradually learns to contribute.

Uncertainty-conditioned decoding. While memory extends what information is available, it does not inform the prediction heads about the current level of perceived uncertainty of the navigator. To this end, we introduce UCD that modulates the shared latent representation based on the predicted cognitive cost. We utilize the adaptive layer normalization from [48] to condition on the predicted $\hat{U}_t$. Since $\hat{U}_t$ is predicted from the shared encoder rather than provided as an external input, UCD treats uncertainty as a learned internal state and lets the model learn how to map it into modulation parameters. Given the memory-augmented features $\mathbf{h}^{\mathrm{mem}} \in \mathbb{R}^d$, UCD produces uncertainty-aware features via:

$$\tilde{\mathbf{h}} = (1+\gamma) \odot \mathrm{LN}(\mathbf{h}^{\mathrm{mem}}) + \beta, \quad (\gamma, \beta) = \mathrm{MLP}(\hat{U}_t) \tag{4}$$

where LN is layer normalization, $\odot$ is element-wise multiplication, and the MLP maps $\hat{U}_t \in \mathbb{R}$ to $\mathbb{R}^{2d}$ via a two-layer network with SiLU activation [14]. Following [48], the MLP is zero-initialized so that $\gamma = \mathbf{0}$ and $\beta = \mathbf{0}$ at the start of training to ensure identity behavior until the model learns meaningful modulation.

Memory and UCD address complementary limitations. Memory extends the information available with learned navigation patterns, and UCD adjusts how the prediction heads internally process this information based on perceived uncertainty of the current scene.

Trajectory and head motion prediction. On top of the uncertainty-aware features $\tilde{\mathbf{h}}$, two prediction heads produce task-specific forecasts: a 3-DOF body-frame trajectory $\widehat{\mathbf{S}}_{T_1+1:T_1+T_2}$ and a 6D head-motion sequence $\widehat{\mathbf{H}}_{T_1+1:T_1+T_2}$.

3.3 Training Objectives

Task losses. For trajectory, we prioritize near-future steps [3] with discounted ℓ_1 loss and a variance regularization term:

$$\mathcal{L}_{\mathrm{traj}} = \textstyle\sum_{i=1}^{T_2} \gamma^i \left\| \widehat{\mathbf{S}}_i - \mathbf{S}_i^{\mathrm{gt}} \right\|_1 + \lambda_{\mathrm{var}} \left\| \mathrm{std}_i(\widehat{\mathbf{S}}) - \mathrm{std}_i(\mathbf{S}^{\mathrm{gt}}) \right\|_2^2 \tag{5}$$

where $\gamma = 0.98$ and $\lambda_{\mathrm{var}} = 0.3$, and $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^3$ is the per-coordinate standard deviation over the T_2 future steps. Following [46], head rotations use the rotation matrix ℓ_1 distance:

$$\mathcal{L}_{\mathrm{head}} = \frac{1}{T_2} \sum_{i=1}^{T_2} \left\| \widehat{\mathbf{R}}_{t+i}\, \mathbf{R}_{t+i}^{\top} - \mathbf{I} \right\|_1 \tag{6}$$

where $\widehat{\mathbf{R}}_{t+i}$ and $\mathbf{R}_{t+i}$ are rotation matrices recovered from the predicted and ground-truth 6D representations. We regress the human self-reported perceived uncertainty with mean squared error:

$$\mathcal{L}_U = \left\| \hat{U}_t - U_t^{\text{human}} \right\|_2^2. \tag{7}$$

Multi-task objective. The total training loss combines all three task losses with equal weighting:

$$\mathcal{L} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{head}} + \mathcal{L}_U \tag{8}$$

3.4 Implementation Details

We train on a single NVIDIA RTX 4090 with AdamW [34], using cosine annealing over 300 epochs with a maximum learning rate of 1×10^{-4}, weight decay 8×10^{-5}, and batch size 64. DINOv2 features are precomputed and cached and the vision backbone remains frozen throughout training. The model dimension is $d=512$. The action encoder uses 4 transformer layers and the perception encoder uses 2 layers. The memory module contains $N_m = 16$ slots with internal dimension $d_m = 256$.

4 Cognition-aware Egocentric Navigation (CEN) Dataset

As discussed in Section 2, there is currently no publicly available dataset to support research into egocentric human navigation with cognitive factors. We therefore collect a multimodal egocentric navigation dataset combining rich sensory inputs with moment-to-moment annotations of perceived path uncertainty. We detail the data-collection pipeline and dataset statistics below.

4.1 Data Collection

Hardware. To accommodate the distinct characteristics of indoor and outdoor environments, we employ two complementary setups. For outdoors, we use Tobii Pro Glasses capturing 20fps RGB with high-quality binocular gaze and 6-axis IMU commonly used across studies [44], and a reliable Garmin handheld GPS providing precise global positions. For indoor scenarios, we use Project Aria glasses [15] that capture 20fps RGB video with accurate SLAM via two monochrome cameras, eye-tracking, and an IMU sensor. In both settings, participants hold an Xbox controller [3] to continuously self-report perceived path uncertainty on a normalized [0,1] scale.

Recording procedure. The procedure was approved by the Institutional Review Board (IRB). Before each session, participants were briefed on the study goals and received a tutorial on how to continuously report perceived uncertainty with the joystick for full range. Participants were explicitly instructed to

[3] https://xboxdesignlab.xbox.com/en-us/controllers/xbox-wireless-controller

perform route-finding behaviors when uncertain such as scanning the surroundings for cues, confirming signage or landmarks, and were reminded to check for passing vehicles before crossing streets. During recording, participants navigated from a fixed starting point through an identical sequence of waypoints per scenario toward predefined goals for each scenario. A researcher trailed each participant to present the next waypoint image upon arrival at previous one to ensure proper task progression.

Data processing. Outdoor recordings from Tobii Pro Glasses are processed in Tobii Pro Lab[4] and exported as `.tsv` files with multimodal streams including 2D gaze coordinates, 6-axis IMU signals, and egocentric videos. Trajectories are obtained from GPS logs in `.gpx` files and are smoothed with a Savitzky–Golay filter [55]. Joystick signals are saved as `.csv` with magnitude of each press. Indoor sessions are stored in `.vrs` files and processed with Aria Machine Perception Services[5] to extract RGB video, 6D head-pose trajectories, scene point clouds, and eye-gaze estimates. All recordings are synchronized to a 10 Hz timeline via down-sampling with nearest neighbor or interpolation. We additionally annotate videos with behavior types and environment categories to provide labels for supervised learning and validation.

Privacy. All recordings were de-identified by blurring faces and removing audio to protect privacy of participants.

4.2 Dataset Statistics.

The dataset comprises approximately 6 hours from 17 participants with total 226k RGB frames across 42 distinct sites. Diverse sites were selected across indoor and outdoor settings, including university campuses, healthcare facilities, urban commercial streets, and natural routes. The sites are recorded at varied times of day with varying lighting, crowding, and traffic. Wayfinding routes are designed to integrate four uncertainty-inducing spatial types. We further annotate three label groups for stratified analysis and auxiliary supervision. Labels were selected from pilot observations and theory that cause choice difficulty and subsequent behavioral responses. Environment types include multiple-route junctions (`JCT`), occluded/poor-signage segments (`OCC`), dynamic/crowded areas (`CROWD`), and spatial transitions/sudden changes (`ST`); trajectory behaviors include hesitation/pausing (`HES`), wrong turn (`WRONG`), and backtrack (`BACK`); head-movement behaviors include information gathering (`SCAN`), information confirmation (`CONFIRM`), and look-back (`LB`).

The four environment types account for 42.9% of the recorded traversal, including junctions (`JCT`, 24.4%), occluded or poor-signage areas (`OCC`, 10.6%), spatial transitions or sudden scene changes (`ST`, 3.4%), and crowded or dynamic areas (`CROWD`, 4.5%). The remaining 57.1% is treated as neutral. Mean self-reported uncertainty is substantially higher in navigation-challenging regions,

[4] https://www.tobii.com/products/software/behavior-research-software/tobii-pro-lab
[5] https://huggingface.co/projectaria

reaching 0.47 at junctions and 0.44 in occluded or poorly signed areas, compared with 0.19 in neutral segments. This distribution shows that the joystick signal is not uniform across the route, but systematically increases in spatial conditions that require route choice, visual search, or additional confirmation. We further validate that the uncertainty signal reflects perceived navigation difficulty by comparing it against independently video-annotated behaviors labeled by two independent reviewers. Reported uncertainty is consistently elevated during decision-difficulty behaviors, with medium-to-large effect sizes for wrong turns (Cohen's $d = 0.80$), look-backs ($d = 0.66$), and backtracking ($d = 0.55$).

5 Experiments

We evaluate EgoCogNav with baselines and ablate key design choices in Section 5.1. We also provide qualitative analyses that visualize trajectory and head-pose forecasts and uncertainty estimation in Section 5.2. Finally, we discuss limitations and future directions in Section 5.3. All results are reported on a held-out test split. Since participants follow the same waypoint sequences within each scenario, this split evaluates generalization to held-out traversal instances, head and gaze movements, and local egocentric observations.

5.1 Quantitative Evaluation

Metrics. For trajectory, we report Average Displacement Error (ADE), which captures the mean distance between predicted and actual positions at each timestep; and Final Displacement Error (FDE), which measures the distance at the trajectory endpoint. We evaluate the L1 loss [46] used during training for head rotations. As for uncertainty, we report mean absolute error (MAE), Spearman rank correlation coefficient [20] over all scenarios and on top-20% high-uncertainty subset. ΔU measures the mean elevation of predicted uncertainty during annotated navigation behaviors relative to neutral segments to capture the model's sensitivity to behavioral difficulty.

Baselines. Since this is a novel task, there are limited baselines to the best of our knowledge. The closest related works [46, 50, 64] are pre-printed or have no publicly released code. As a result, we compare against the following baselines: (1) Constant Velocity (`Const_Vel`) [41], which extrapolates future body-frame translation from the last linear velocity and future head rotation from the last angular velocity, (2) Linear Extrapolation (`Lin_Ext`) [46], which fits a per-axis linear model to the past translation and rotation sequences and projects them into the future, (3) a standard Multimodal Transformer (`M_Transformer`) baseline that uses the same inputs and training protocol as our model but performs early fusion by concatenating embeddings with a single temporal transformer decoder, followed by linear heads for the three comprehensive tasks, and (4) EgoCast [16] where we adapt its forecasting module which was originally designed for full-body pose to our perceived uncertainty, trajectory, and head-motion prediction settings with additional layers, using the same frozen DINOv2 features as EgoCogNav to isolate the forecasting module from the visual backbone.

Method	All Scenarios			High-U Scenarios			Uncertainty	
	ADE ↓	FDE ↓	Head ↓	ADE ↓	FDE ↓	Head ↓	MAE ↓	ρ ↑
Const_Vel	.1892	.4257	.0875	.2170	.4778	.0877	–	–
Lin_Ext	–	–	.1224	–	–	.1235	–	–
M_Transformer	.1536	.3213	.0776	.1684	.3469	.0778	.1247	.683
EgoCast* [16]	.1092	.2184	.0712	.1198	.2369	.0716	.1029	.752
EgoCogNav	**.1051**	**.2074**	**.0698**	**.1155**	**.2256**	**.0704**	**.0986**	**.788**

Table 1: Baseline comparison. Trajectory (ADE/FDE), head rotation (L1), and uncertainty value error (MAE) on the full test set and on a high-uncertainty subset.

Method	All Scenarios		High-U Scenarios		
	MAE ↓	ρ ↑	MAE ↓	ρ ↑	ΔU ↑
EMU [22]	.1887	.081	.1842	.100	.0122
PATH_U [67]	.1857	.195	.1814	.210	.0212
EgoCogNav	**.0986**	**.788**	**.1017**	**.636**	**.0829**

Table 2: Uncertainty prediction. MAE and Spearman ρ are reported for all samples and for the high-uncertainty subset. ΔU measures mean predicted uncertainty elevation during annotated navigation behaviors relative to neutral segments.

As for perceived uncertainty prediction, we compare with two baselines: (1) an EMU-entropy (`EMU`) theory proxy, which computes two signals of perceptual ambiguity from visual scenes and behavioral variability from short-horizon motion, and learns a linear combination to the human uncertainty label; and (2) a PATH-U-adapted (`PATH_U`) [67] model, as it was originally designed for indoor wayfinding using signage, we adapt it to the egocentric setting by composing a 5-dimensional feature vectors capturing decision complexity and behavioral variability (e.g., junction count, occlusion/poor-signage count, goal distance), and fitting a linear regressor to predict perceived uncertainty.

Results. The comparison results are presented in Table 1. Our model achieves the best trajectory and head-motion performance on both the full test set and the high-uncertainty subset, reducing ADE/FDE by 3.8%/5.0% relative to the EgoCast-adapted baseline [16]. The M_Transformer baseline uses early fusion without cognition-aware modules and shows degraded performance, which suggests that late fusion with modality-specific temporal encoding better preserves the sensory patterns needed for accurate forecasting. Table 2 directly compares our learned uncertainty against dedicated baselines to evaluate whether the model captures meaningful cognitive states. The EMU entropy proxy [22] and PATH_U heuristic [67] both achieve near-chance rank correlation, indicating that hand-crafted rules based on scene ambiguity or decision complexity cannot capture the subjective, person-specific nature of perceived uncertainty. In contrast, our model achieves $\rho = 0.788$ by learning to map multimodal sensory-

				All Scenarios			High-U Scenarios		
Method	U	UCD	Mem	ADE ↓	FDE ↓	Head ↓	ADE ↓	FDE ↓	Head ↓
Base module	–	–	–	.1168	.2443	.0785	.1286	.2630	.0790
+ U prediction	✓	–	–	.1121	.2217	.0721	.1212	.2401	.0722
+ UCD	✓	✓	–	.1096	.2188	.0712	.1202	.2381	.0716
+ Memory	✓	–	✓	.1114	.2193	.0707	.1228	.2384	.0710
EgoCogNav	✓	✓	✓	**.1051**	**.2074**	**.0698**	**.1155**	**.2256**	**.0704**

Table 3: Ablation study. Each design choice contributes and UCD and memory show complementary gains when combined.

motor patterns to individual cognitive states through joint training. The behavioral sensitivity gap is equally strong as EgoCogNav produces higher elevation of perceived uncertainty during annotated navigation behaviors compared to the baselines.

Ablation study. Table 3 presents the ablation results. The full model achieves the best trajectory performance across all scenarios and high-uncertainty moments. The largest improvement comes from adding uncertainty prediction alone which reduces FDE by 9.2% and head error by 8.2%. This suggests that uncertainty supervision provides useful gradient signal to the shared encoder to encourage representations that are sensitive to decision difficulty and behavior transitions (e.g., scanning, hesitation, and backtracking). Since moments of high perceived uncertainty systematically co-occur with hesitation, direction changes, and head confirming, the encoder learns features that support both cognitive-state estimation and short-horizon forecasting. Neither UCD nor memory alone substantially improves performance, but their combination produces the largest trajectory gains. This complementarity arises because the two modules address different aspects of the forecasting problem. The memory module augments the representation with learned navigation patterns, while UCD modulates decoding according to the predicted uncertainty state. Their combination allows the model to use uncertainty not only as an auxiliary supervised signal, but also as an internal conditioning variable for trajectory and head-motion prediction.

Table 4 reports the contribution of each input modality. Motion and goal alone provide limited gains, while adding video substantially improves trajectory and uncertainty prediction. Head history further improves head-motion forecasting, and the full multimodal model achieves the best overall performance. Table 5 further breaks results down by annotated behaviors. EgoCogNav achieves the best trajectory performance on `HES`, `WRONG`, and `BACK`, indicating that the method most benefits decision-intensive moments. For head behaviors, improvements are distributed across components where memory alone performs best on `CONF` and `LB`, and the full model provides the strongest gain on `SCAN`.

Input	ADE↓	FDE↓	Head L1↓	ρ_U ↑	MAE_U ↓
Motion Only	0.156	0.314	0.0794	0.313	0.178
Motion+Goal	0.155	0.313	0.0791	0.314	0.181
+Video	0.112	0.213	0.0713	0.778	0.0968
+Head	0.115	0.217	0.0708	0.780	0.0971
Full (+Gaze)	**0.105**	**0.207**	**0.0698**	**0.788**	**0.0986**

Table 4: Modality ablation. Contribution of each input modality to trajectory, head-motion, and uncertainty prediction.

	Trajectory Behaviors						Head Behaviors		
	HES		WRONG		BACK		SCAN	CONF	LB
Method	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓	Head↓	Head↓	Head↓
Base module	.1192	.2442	.0929	.2013	.1178	.2471	.0991	.0820	.1278
+ U prediction	.1136	.2214	.0952	.1938	.1082	.2205	.0877	.0768	**.1244**
+ UCD	.1160	.2232	.0945	.1893	.1092	.2218	.0872	.0789	.1248
+ Memory	.1134	.2188	.0946	.1909	.1064	.2169	.0869	**.0749**	**.1244**
EgoCogNav	**.1112**	**.2115**	**.0905**	**.1825**	**.1040**	**.2100**	**.0864**	.0751	.1277

Table 5: Ablation for navigation behaviors. Performance during annotated navigation behaviors for trajectories and head movements.

5.2 Qualitative Evaluation

We visualize predictions alongside ground truth for trajectories and head motion, and overlay uncertainty as a color-coded intensity along the predicted path shown in Fig. 3. Overall, our model closely tracks ground-truth trajectories and head orientations across conditions while producing uncertainty values that align with decision difficulty environments. In multi-junction scenes, we observe elevated uncertainty before hesitation and scanning. And in occlusion-heavy cases, uncertainty peaks before backtracking. Conversely, in well-specified corridors with clear sight-lines, uncertainty remains low and motion is smooth. These patterns qualitatively support our analysis and indicate that the model captures both route-finding behaviors, and their coupling with perceived uncertainty. From an environment-centric perspective, we observe that scenes rated by participants as more confusing or cognitively demanding can systematically induce higher predicted uncertainty, while visually simple regions correspond to low predicted uncertainty. This alignment between environmental structures and subjective feelings from model outputs suggests that EgoCogNav captures not only route-finding behaviors but also how different environments are experienced.

Failure cases. We also identify failure cases as illustrated in Fig. 4. In the first scenario, the participant suspected a wrong turn due to heavy occlusion

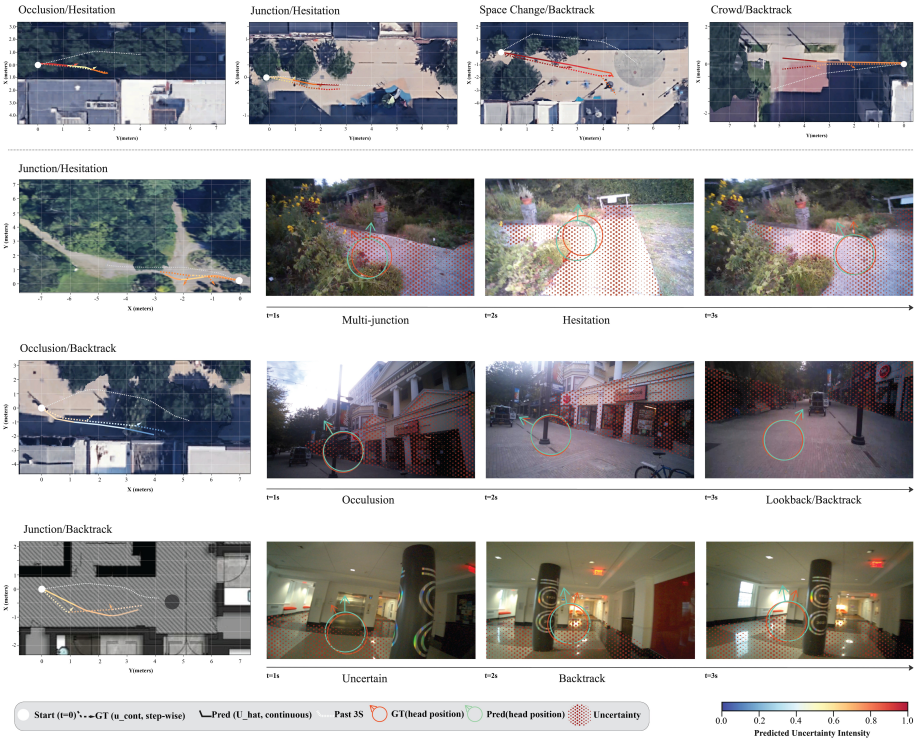

Fig. 3: Qualitative visualizations. The top row presents BEV examples under high-uncertainty and behavior-eliciting scenarios. For each scenario on the bottom row, the left panel shows a BEV overlay with past trajectory (gray), ground-truth future, predicted future, and overlaid path uncertainty. The right panels show time-aligned egocentric frames (t+1 to t+3s) with ground-truth (red) and predicted (green) head positions. Environments were also highlighted with red dots for those that trigger uncertain behaviors.

and backtracked to seek an alternative route. Our model instead predicted a return along the same path segment rather than backtracking to other decision points. This indicates insufficient use of long-horizon visual context and episodic scene memory. In the second example, although the predicted path heads in the correct direction, it fails to capture the brief hesitation and look-back behaviors triggered by the changing scene context. These failure cases emphasize the need for stronger global context beyond the immediate scene representations and for explicit multi-hypothesis futures when comparable uncertainty covers multiple plausible routes.

5.3 Limitations and Future Work

There are several limitations noted. First, when salient cues are outside the camera frustum or are heavily occluded, the model operates with partial scene knowl-

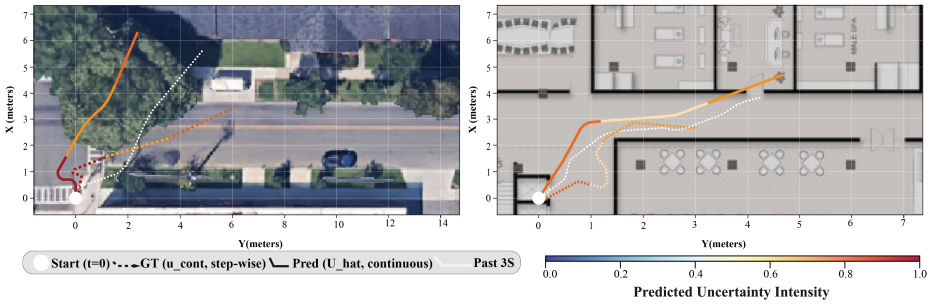

Fig. 4: Failure cases. Two failure cases highlight limits in long-horizon scene memory.

edge and degrades performance. Future work can incorporate richer 3D/semantic context to improve disambiguation at complex environments. Second, we predict a single best trajectory and head sequence, and can utilize generative models to sample a distribution of futures can better capture diverse navigation styles under similar cognitive states. Third, this work features uncertainty as main cognitive factor and future work will incorporate additional cognitive/experiential signals such as affect and spatial memory and extend from short-horizon forecasting to hierarchical, longer-horizon planning. Last, CEN is intended as a first-step dataset for cognition-aware egocentric navigation. Future extensions will expand the participant pool, scene diversity, and route types, and incorporate richer annotations such as signage visibility, semantic landmarks, and social dynamics. These additions would support stronger analysis of personalization, environment-level difficulty, and assistive navigation policies.

6 Conclusion

In this paper, we present several key contributions to the goal of cognition-aware egocentric navigation. First, we introduce the challenging task of jointly forecasting body-centered trajectory, head motion, and perceived path uncertainty from multimodal inputs. This formulation enables the model to reason not only about future motion but also about underlying cognitive states to explain onset behaviors and environmental evaluation. Second, we propose EgoCogNav, an architecture that effectively integrates scene features and sensory cues to accurately forecast human motion. To support this task, we collect the CEN dataset, a 6-hour collection of real-world recordings across 42 diverse environments with synchronized video, gaze, and self-reported uncertainty. Our results indicate that EgoCogNav learns the coupling between perception, perceived uncertainty, and human-like navigation behaviors and improves forecasting performance on held-out navigation recordings. We also discuss the limitations of the project and important future work that is needed toward real-world deployment in social and assistive navigation technologies.

References

1. Akada, H., Wang, J., Golyanik, V., Theobalt, C.: 3d human pose perception from egocentric stereo videos. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 767–776 (2024)
2. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961–971 (2016)
3. Amit, R., Meir, R., Ciosek, K.: Discount factor as a regularizer in reinforcement learning. In: International conference on machine learning. pp. 269–278. PMLR (2020)
4. Bae, I., Park, Y.J., Jeon, H.G.: Singulartrajectory: Universal trajectory predictor using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17890–17901 (2024)
5. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7061–7071 (2025)
6. Bosse, T., Hoogendoorn, M., Klein, M.C., Treur, J., Van Der Wal, C.N., Van Wissen, A.: Modelling collective decision making in groups and crowds: Integrating social contagion and interacting emotions, beliefs and intentions. Autonomous Agents and Multi-Agent Systems **27**(1), 52–84 (2013)
7. Brunyé, T.T., Gardony, A.L., Holmes, A., Taylor, H.A.: Spatial decision dynamics during wayfinding: Intersections prompt the decision-making process. Cognitive Research: Principles and Implications **3**(1), 13 (2018)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018)
10. De Iuliis, M., Battegazzorre, E., Domaneschi, M., Cimellaro, G.P., Bottino, A.G.: Large scale simulation of pedestrian seismic evacuation including panic behavior. Sustainable Cities and Society **94**, 104527 (2023)
11. Devlin, A.S.: Wayfinding in healthcare facilities: Contributions from environmental psychology. Behavioral sciences **4**(4), 423–436 (2014)
12. Dubey, R.K., Sohn, S.S., Hoelscher, C., Kapadia, M.: Fusion-based wayfinding prediction model for multiple information sources. In: 2019 22th international conference on information fusion (FUSION). pp. 1–8. IEEE (2019)
13. Edward, L., Lourdeaux, D., Barthes, J.P.: Cognitive modeling of virtual autonomous intelligent agents integrating human factors. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. vol. 3, pp. 353–356. IEEE (2009)
14. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural networks **107**, 3–11 (2018)
15. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023)

16. Escobar, M., Puentes, J., Forigua, C., Pont-Tuset, J., Maninis, K.K., Arbelaez, P.: Egocast: Forecasting egocentric human pose in the wild. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5831–5841. IEEE (2025)
17. Farr, A.C., Kleinschmidt, T., Yarlagadda, P., Mengersen, K.: Wayfinding: A simple concept, a complex process. Transport Reviews **32**(6), 715–743 (2012)
18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
19. Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17113–17122 (2022)
20. Hauke, J., Kossowski, T.: Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones geographicae **30**(2), 87–93 (2011)
21. Hirose, N., Shah, D., Sridhar, A., Levine, S.: Sacson: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters **9**(1), 49–56 (2023)
22. Hirsh, J.B., Mar, R.A., Peterson, J.B.: Psychological entropy: a framework for understanding uncertainty-related anxiety. Psychological review **119**(2), 304 (2012)
23. Huang, R., Zhao, X., Yuan, Y., Yu, Q., Liu, C., Daamen, W.: Modeling pedestrian tactical and operational decisions under risk and uncertainty: A two-layer model framework. IEEE Transactions on Intelligent Transportation Systems **24**(5), 5259–5281 (2023)
24. Jain, A., Casas, S., Liao, R., Xiong, Y., Feng, S., Segal, S., Urtasun, R.: Discrete residual flow for probabilistic pedestrian behavior prediction. In: Conference on Robot Learning. pp. 407–419. PMLR (2020)
25. Jeong, J., Lee, S., Park, D., Lee, G., Yoon, K.J.: Multi-modal knowledge distillation-based human trajectory forecasting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24222–24233 (2025)
26. Kaya, S.D., Ileri, Y.Y., Yuceler, A.: Importance of hospital way-finding system on patient satisfaction. In: Business Challenges in the Changing Economic Landscape-Vol. 2: Proceedings of the 14th Eurasia Business and Economics Society Conference. pp. 33–40. Springer (2016)
27. Keller, A.M., Taylor, H.A., Brunyé, T.T.: Uncertainty promotes information-seeking actions, but what information? Cognitive Research: Principles and Implications **5**(1), 42 (2020)
28. Kothari, P., Kreiss, S., Alahi, A.: Human trajectory forecasting in crowds: A deep learning perspective. IEEE Transactions on Intelligent Transportation Systems **23**(7), 7386–7400 (2021)
29. Kretzschmar, H., Spies, M., Sprunk, C., Burgard, W.: Socially compliant mobile robot navigation via inverse reinforcement learning. The International Journal of Robotics Research **35**(11), 1289–1307 (2016)
30. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)
31. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6943–6953 (2021)

32. Li, Y., Liu, M., Rehg, J.M.: In the eye of the beholder: Gaze and actions in first person video. IEEE transactions on pattern analysis and machine intelligence **45**(6), 6731–6747 (2021)
33. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: Egoexofitness: Towards egocentric and exocentric full-body action understanding. In: European Conference on Computer Vision. pp. 363–382. Springer (2024)
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
35. Lv, Z., Charron, N., Moulon, P., Gamino, A., Peng, C., Sweeney, C., Miller, E., Tang, H., Meissner, J., Dong, J., et al.: Aria everyday activities dataset. arXiv preprint arXiv:2402.13349 (2024)
36. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024)
37. Mackett, R.L.: Mental health and wayfinding. Transportation research part F: traffic psychology and behaviour **81**, 342–354 (2021)
38. Maeda, T., Ukita, N.: Fast inference and update of probabilistic density estimation on trajectory prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9795–9805 (2023)
39. Mao, W., Xu, C., Zhu, Q., Chen, S., Wang, Y.: Leapfrog diffusion model for stochastic trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5517–5526 (2023)
40. Maruyama, T., Kanai, S., Date, H., Tada, M.: Simulation-based evaluation of ease of wayfinding using digital human and as-is environment models. ISPRS International Journal of Geo-Information **6**(9), 267 (2017)
41. Mavrogiannis, C., Alves-Oliveira, P., Thomason, W., Knepper, R.A.: Social momentum: Design and evaluation of a framework for socially competent robot navigation. ACM Transactions on Human-Robot Interaction (THRI) **11**(2), 1–37 (2022)
42. Mavrogiannis, C., Baldini, F., Wang, A., Zhao, D., Trautman, P., Steinfeld, A., Oh, J.: Core challenges of social robot navigation: A survey. ACM Transactions on Human-Robot Interaction **12**(3), 1–39 (2023)
43. Nguyen, D.M., Nazeri, M., Payandeh, A., Datar, A., Xiao, X.: 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). pp. 7442–7447. IEEE (2023)
44. Onkhar, V., Dodou, D., De Winter, J.: Evaluating the tobii pro glasses 2 and 3 in static and dynamic conditions. Behavior Research Methods **56**(5), 4221–4238 (2024)
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
46. Pan, B., Harley, A.W., Engelmann, F., Liu, C.K., Guibas, L.J.: Lookout: Real-world humanoid egocentric navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24977–24988 (2025)
47. Pan, X.: Computational modeling of human and social behaviors for emergency egress analysis. Stanford University (2006)
48. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)

49. Pouyan, A.E., Ghanbaran, A.H., Hosseinzadeh, A., Shakibamanesh, A.: The elderly wayfinding performance in an informative healthcare design indoors. Journal of Building Engineering **87**, 108843 (2024)
50. Qiu, J., Chen, L., Gu, X., Lo, F.P.W., Tsai, Y.Y., Sun, J., Liu, J., Lo, B.: Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion. IEEE Robotics and Automation Letters **7**(4), 8799–8806 (2022)
51. Raubal, M.: Human wayfinding in unfamiliar buildings: a simulation with a cognizing agent. Cognitive Processing **2**(3), 363–388 (2001)
52. Raubal, M., Worboys, M.: A formal model of the process of wayfinding in built environments. In: International conference on spatial information theory. pp. 381–399. Springer (1999)
53. Saadatnejad, S., Gao, Y., Messaoud, K., Alahi, A.: Social-transmotion: Promptable human trajectory prediction. arXiv preprint arXiv:2312.16168 (2023)
54. Salzmann, T., Chiang, H.T.L., Ryll, M., Sadigh, D., Parada, C., Bewley, A.: Robots that can see: Leveraging human pose for trajectory prediction. IEEE Robotics and Automation Letters **8**(11), 7090–7097 (2023)
55. Schafer, R.W.: What is a savitzky-golay filter?[lecture notes]. IEEE Signal processing magazine **28**(4), 111–117 (2011)
56. Schöller, C., Knoll, A.: Flomo: Tractable motion prediction with normalizing flows. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7977–7984. IEEE (2021)
57. Shi, S., Jiang, L., Dai, D., Schiele, B.: MTR++: Multi-agent motion prediction with symmetric scene modeling and pair-wise interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9787–9803 (2024)
58. Song, D., Liang, J., Payandeh, A., Raj, A.H., Xiao, X., Manocha, D.: Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. IEEE Robotics and Automation Letters (2024)
59. Tang, H., Liang, K.J., Grauman, K., Feiszli, M., Wang, W.: Egotracks: A long-term egocentric visual object tracking dataset. Advances in Neural Information Processing Systems **36**, 75716–75739 (2023)
60. Tome, D., Peluse, P., Agapito, L., Badino, H.: xr-egopose: Egocentric 3d human pose from an hmd camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7728–7738 (2019)
61. Vellenga, K., Steinhauer, H.J., Falkman, G., Björklund, T.: Evaluation of video masked autoencoders' performance and uncertainty estimations for driver action and intention recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7429–7437 (2024)
62. Walker, N., Mavrogiannis, C., Srinivasa, S., Cakmak, M.: Influencing behavioral attributions to robot motion during task execution. In: Conference on Robot Learning. pp. 169–179. PMLR (2022)
63. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023)
64. Wang, W., Liu, C.K., Kennedy III, M.: Egonav: Egocentric scene-aware human trajectory prediction. arXiv preprint arXiv:2403.19026 (2024)
65. Xie, W., Lee, E.W.M., Lee, Y.Y.: Simulation of spontaneous leader–follower behaviour in crowd evacuation. Automation in Construction **134**, 104100 (2022)
66. Xu, C., Mao, W., Zhang, W., Chen, S.: Remember intentions: Retrospective-memory-based trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2022)

67. Yang, Q., Dubey, R.K., Kalantari, S.: Path-u: A data-driven agent-based wayfinding model incorporating perceived path uncertainty and cognitive strategies in unfamiliar indoor environments. In: Building Simulation. vol. 18, pp. 449–471. Springer (2025)
68. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7072–7084 (2025)
69. Yu, J., Zhu, J.: FMTP: Fine-grained multi-modal trajectory prediction. arXiv preprint arXiv:2407.02558 (2024)
70. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: ReMoDiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
71. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019)
72. Zhu, R., Becerik-Gerber, B., Lin, J., Li, N.: Behavioral, data-driven, agent-based evacuation simulation for building safety design using machine learning and discrete choice models. Advanced Engineering Informatics **55**, 101827 (2023)
73. Zhu, Y., Chen, T., Ding, N., Chraibi, M., Fan, W.C.: Follow people or signs? a novel way-finding method based on experiments and simulation. Physica A: Statistical Mechanics and its Applications **573**, 125926 (2021)