HUMAIN: Human-Aware Implicit Social Robot Navigation

Daeun Song[1], Nhat Le[2], Jeffrey Chen[3], Mohammad Nazeri[2], Amirreza Payandeh[2], Rohan Chandra[3], Reuth Mirsky[4], Ross Mead[5], Ling Xiao[6], Xuesu Xiao[2]

***Abstract*— Effective social robot navigation requires sensitivity to human behavior, often revealed through subtle skeletal cues like gait and orientation. We present Human-Aware Implicit Social Robot Navigation (HUMAIN), a novel framework that fuses implicit social cues directly into the planning loop via knowledge distillation. We first employ a transformer-based teacher model that fuses rich multi-modal inputs, including historic images, skeletal keypoints, robot state, and a robot's target goal, to learn robust, human-aware representations for the robot's future trajectory planning. To enable real-time deployment, we then distill this knowledge into a lightweight student model. By optimizing for both trajectory reconstruction and latent feature alignment with the teacher, the student learns to infer complex social dynamics from minimal inputs. Bridging the prediction–planning gap with an efficient distilled architecture, our method enables robots to reason about human behavior in a manner that is adaptive, robust, and socially compliant. We validate HUMAIN through extensive experiments, where it improves trajectory prediction metrics by an average of 29.8% across all metrics compared to state-of-the-art baselines. These results highlight the benefit of using implicit, whole-body cues to achieve human-like navigation awareness on resource-constrained platforms.**

I. INTRODUCTION

Social robot navigation is a critical area of robotics, driven by the growing demand for robots to operate autonomously in human-populated environments, from hospitals [1] to public spaces [2]. Unlike navigation in static or industrial domains, social navigation requires sensitivity to human behaviors and comfort [3]. To navigate fluently, a robot must go beyond collision avoidance and reason about others' intended paths, such as differentiating a pedestrian continuing straight from one preparing to turn [4]. Human motor control studies suggest that pedestrians implicitly reveal navigational goals through body pose, head orientation, and gait [5]. For example, body pose can often indicate whether a person is walking or standing, and their intended walking direction.

However, despite the richness of these signals, many motion planning approaches oversimplify humans as low-dimensional 2D points or moving circles [6]. They typically react only to position and velocity using distance keeping [7] or proxemic rules [8], [9]. While efficient, such representations discard body language, making it difficult to perceive preparatory cues. As a result, pedestrian turns can appear more difficult to anticipate.

[1]Ewha Womans University, Korea. songd@ewha.ac.kr; [2]George Mason University, USA. {nle47, mnazerir, apayande, xiao}@gmu.ac.kr; [3]University of Virginia, USA. {fyy2wsm, rohanchandra} @virginia.edu; [4]Tufts University, USA. reuth.mirsky@tufts.edu; [5]Semio, USA. ross@semio.ai; [6]Hokkaido University, Japan. ling@ist.hokudai.ac.jp

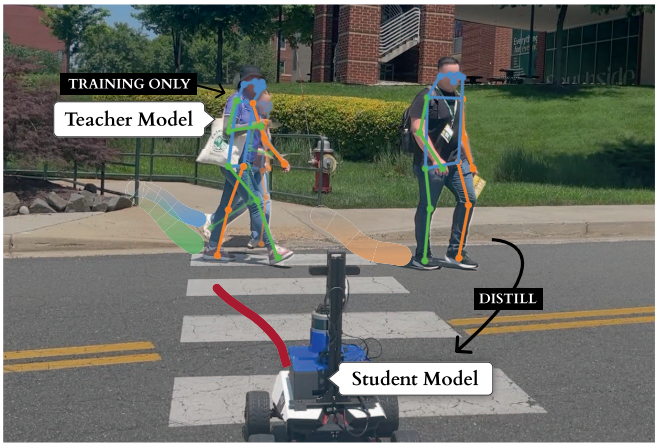


Fig. 1: HUMAIN distills a privileged knowledge from a teacher model into a student model. While the teacher uses these keypoints to plan robot trajectories, the student learns to infer social cues directly from RGB input to plan safe, socially compliant trajectories end-to-end.

While human trajectory prediction methods have leveraged pose to improve forecasting [10], [11], integrating these predictions into robot trajectory planning remains challenging. Many systems operate in a decoupled manner by forecasting a human path and then planning a robot trajectory to avoid it [12]. This modular approach treats forecasts as fixed constraints and often yields overly conservative maneuvers or inefficient detours.

In contrast, humans implicitly use social cues such as body pose, hip orientation, gaze, and gait to infer others' short-term motion and adjust their own trajectories accordingly [13]. Yet robots often lack the sensors or compute budget to track such cues reliably at deployment time, which creates a gap between the rich information needed for social reasoning and the limited sensorimotor inputs available at deployment. A natural solution is to learn representations with a computationally powerful teacher that uses privileged information during training as scaffolding, then distill this knowledge into an efficient onboard student that reasons from raw observations.

Inspired by this intuition, we propose a Teacher–Student framework that incorporates implicit social cues into robot planning (Fig. 1). We train an Oracle Teacher with privileged multi-modal data, including skeletal keypoints, to learn a rich latent representation of human behavior. We then distill this representation into a lightweight sensorimotor student that operates only on raw images and a goal, enabling socially compliant behavior without explicit pose tracking.

Main Results: We present **HUMAIN**, **Hum**an-**A**ware **I**mplicit social robot **N**avigation, a social navigation approach that incorporates whole-body pose cues via response-based knowledge distillation [14]. We employ a two-stage training strategy. A Teacher learns multi-modal representations from images and skeletal keypoints. A Student is trained to match the Teacher latent representation and imitate its behavior using only raw images. This process enables leveraging skeletal supervision during training while remaining efficient and sensor-agnostic at deployment.

The main contributions of this work are as follows:

- HUMAIN, an end-to-end framework for social navigation that leverages whole-body pose via skeletal keypoints as an implicit cue for robot trajectory planning.
- A distillation strategy that aligns a Student with a privileged Teacher, enabling inference of complex social cues from raw visual inputs.
- Extensive quantitative and qualitative evaluations and ablations showing the impact of pose cues on trajectory planning and overall navigation performance. HUMAIN improves social navigation performance by an average of 29.8% over state-of-the-art baselines and all metrics.

II. Related Work

We review work on social robot navigation, human trajectory prediction, and representation learning for navigation.

A. Social Robot Navigation

Classical social navigation methods based on proxemics [8], [9], social force models [7], cost maps [8], and handcrafted rules [4] offer safety and explainability [15], but often require substantial tuning and heuristics to handle complex scenarios [16], [17]. Learning-based approaches reduce manual design via imitation learning [18], [19] and reinforcement learning [20]–[22], but can be less interpretable and brittle in edge cases [15]. Prior work incorporated orientation or gait cues with handcrafted perception [23], [24]. More recently, VLM-based methods [25]–[27] provide semantic grounding, but they are computationally heavy and sensitive to prompts and domain shifts [28], and they often require costly language supervision. These limitations motivate our approach, which learns implicit human cues via representation learning and distillation without handcrafted rules or expensive labels.

B. Human Trajectory Prediction

Human trajectory prediction is closely related to social navigation, since anticipating motion supports safe and natural robot behaviors. Early generative approaches modeled interactions with recurrent networks or GANs, such as Social LSTM [10] and Social GAN [29], which capture multi-agent dependencies. More recently, Transformer-based and graph-based methods have advanced the field by explicitly reasoning over spatial–temporal relations. For instance, Human Scene Transformer [11] models multi-human interactions with attention, SocialPose [30] leverages pose features for forecasting in crowds, Social-Transmotion [31] exploits human poses and bounding boxes for trajectory forecasting, and UPTor [32] jointly predicts 3D body pose dynamics and trajectories for human–robot interaction scenarios. Some works integrate predictors into MPC for robot deployment [12], yet strong prediction accuracy does not necessarily yield socially compliant navigation behaviors.

C. Representation Learning for Navigation

Representation learning [33], often achieved through self-supervised learning (SSL), enables robots to acquire robust features without costly labels. In navigation, representation learning has been applied in diverse ways. Prior work includes vision–action pretraining [34], language-weak supervision [35], and trajectory-centric SSL [36]. Large-scale supervised policies such as ViNT [37] and NoMaD [38] further improve transfer, but most still focus on low-level control, and do not explicitly capture social cues.

Overall, recent work has shown that incorporating human body pose improves trajectory prediction, demonstrating the value of pose and gait cues for modeling human motion [30], [32]. However, these advances have not yet translated into social robot navigation, where many methods focus on geometric safety, semantic grounding, or general representations from images. Our approach extends the use of skeletal keypoints from trajectory prediction to navigation, and employs SSL and distillation to embed implicit human cues directly into robot planning.

III. Approach

In this section, we present details of our proposed HUMAIN, a hierarchical learning framework for socially compliant navigation. We first define the problem as a trajectory planning problem, given the observation and goal. We then provide an overview of our architecture, followed by detailed descriptions of the multi-modal representation learning in the Teacher model and the knowledge distillation strategy used for the student model.

A. Problem Definition

We formally define social robot navigation as a goal-conditioned trajectory planning problem. The robot receives an observation history $O \in \mathcal{O}^{T_{\text{obs}}}$ and a local navigation goal $g \in \mathcal{G}$, where $\mathcal{O}$ is the observation space, T_{obs} the observation horizon, and $\mathcal{G}$ the goal space. The objective is to generate a future trajectory $\hat{Y} \in \mathcal{Y}^{T_{\text{pred}}}$, where $\mathcal{Y}$ represents the robot's trajectory space and T_{pred} is the prediction horizon. The trajectory must progress toward g while adhering to social norms. To address the challenge of learning social compliance without manual cost engineering, we formulate this process under two distinct information regimes:

Privileged Representation Learning (Teacher): The primary objective of the first stage is to learn a robust latent representation that encapsulates complex social dynamics from privileged observation data. During training, the system has access to an oracle observation set $O_t^{\text{teacher}} = (I_t, R_t, H_t)$, comprising egocentric RGB images $I_t \in \mathcal{I}$, the robot's state

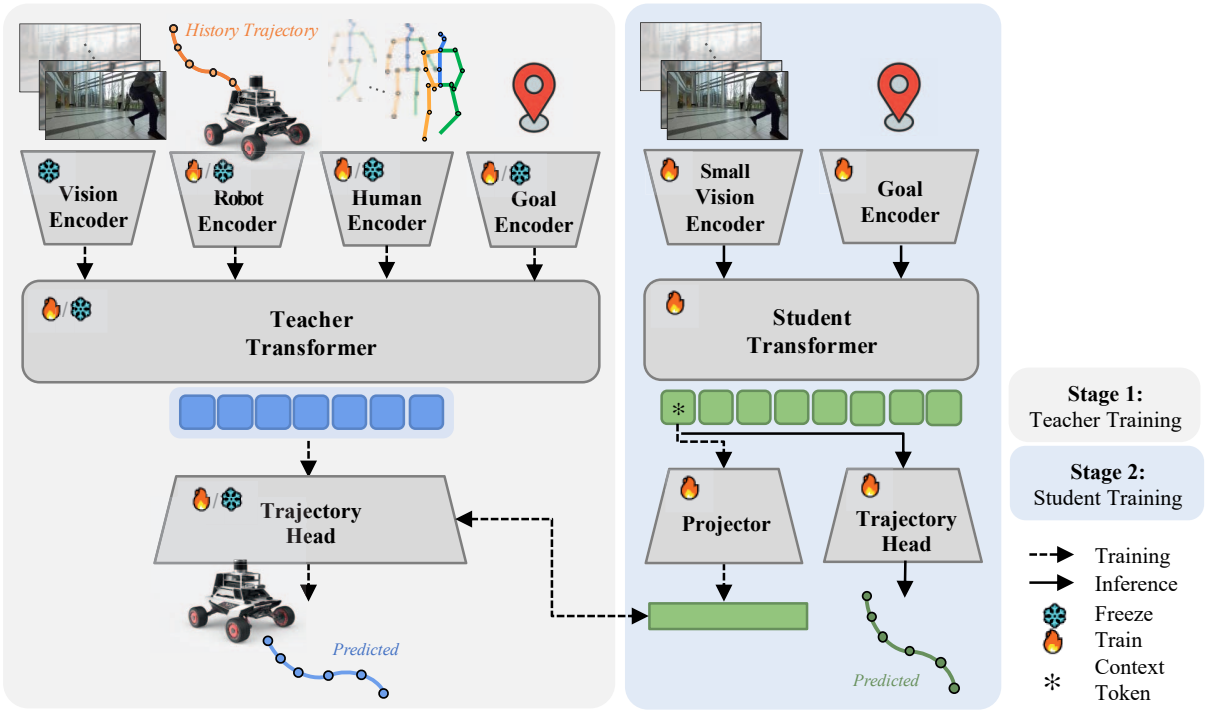


Fig. 2: **HUMAIN Architecture.** Two-stage framework. The Oracle Teacher uses privileged multi-modal inputs including dense keypoints and trajectory history to learn social representations and predict trajectories. The Student is distilled to navigate from raw images and goal by matching the frozen Teacher latent features. Solid lines denote inference. Dotted lines denote training-only distillation.

history $R_t \in \mathcal{R}$, and privileged human states $H_{t,m} \in \mathcal{H}$ for up to M humans. Specifically, we define the state of each human m at time t as a tuple $H_{t,m} = (\mathbf{k}_{t,m}, \mathbf{p}_{t,m}, \mathbf{d}_{t,m})$, comprising the 3D skeletal keypoints, the 2D position relative to the robot, and the 2D facing direction relative to the robot. We seek to learn a mapping function $F_T : (\mathcal{I} \times \mathcal{R} \times \mathcal{H})^{T_{\text{obs}}} \times \mathcal{G} \to \mathcal{Y}^{T_{\text{pred}}}$ parameterized by θ_{teacher}. The Teacher leverages this full spectrum of spatio-temporal data as scaffolding to generate a future trajectory $\hat{Y}_{\text{teacher}}$ that minimizes the prediction error against the ground truth Y_{GT}. Crucially, while the explicit output is the predicted trajectory, the learned latent representation $\mathcal{Z}_{\text{teacher}}$ captures social cues encoded within the skeletal data in $\mathcal{H}$ along with other observations. This latent representation is later used as the distillation target for the Student.

Sensorimotor Policy Distillation (Student): The Student operates under realistic deployment constraints where human tracking data is unavailable or computationally prohibitive. The observation is restricted to the sensorimotor subset $O_t^{\text{student}} = I_t$, containing only the visual history. We seek a deployment policy $F_S : \mathcal{I}^{T_{\text{obs}}} \times \mathcal{G} \to \mathcal{Y}^{T_{\text{pred}}}$, parameterized by θ_{student}. Since the sensorimotor input space is a subset of the privileged input space, minimizing trajectory error alone is often insufficient to recover complex social behaviors. Therefore, the Student's objective is two-fold: (1) to generate a socially-compliant trajectory $\hat{Y}_{\text{student}}$ that matches the ground truth Y_{GT}, and (2) to align its internal belief state with the Teacher's privileged representation $\mathcal{Z}_{\text{teacher}}$. This dual objective forces the Student to implicitly reason about the missing social cues, such as human whole-body pose, solely from the available visual inputs.

B. Model Architecture

In Fig. 2, we illustrate the overall architecture of HUMAIN. Our framework consists of two Transformer-based networks: the Oracle Teacher and the Sensorimotor Student.

The Oracle Teacher is a high-capacity model designed to process multi-modal inputs, including dense human keypoints and robot state history, to generate socially aware trajectories. Each modality is processed by a dedicated encoder to project heterogeneous inputs into a shared embedding space. These embeddings are then fused through a hierarchical, robot-centric attention pipeline. The resulting tokens are passed to the Trajectory Head, implemented as a Multilayer Perceptron (MLP), for trajectory prediction.

The Sensorimotor Student is a lightweight model for real-time inference on mobile robots. It takes only a sequence of raw images and the navigation goal as input. These inputs are encoded by a lightweight vision encoder and a goal encoder, respectively, before being fused by the Student Transformer. A key architectural distinction lies in the token processing strategy. Unlike the Teacher, which utilizes the full sequence of Transformer output tokens, the Student extracts only a single global context token. This token is passed through a projector and a lightweight MLP head to regress to the final trajectory. This design forces the Student to compress complex social reasoning into a compact latent representation, which is explicitly aligned with the Teacher's privileged representation via knowledge distillation.

C. Multi-Modal Representation Learning

The primary goal of the first training stage is to learn the Teacher mapping function F_T and, more importantly, to construct a regularized latent manifold $\mathcal{Z}_{\text{teacher}}$ that encapsulates high-fidelity social semantics. The Oracle Teacher processes the privileged observation sequence $O^{\text{teacher}} = \{(I_t, R_t, H_t)\}_{t=1}^{T_{\text{obs}}}$ through modality-specific encoders, followed by a hierarchical fusion pipeline composed of two Transformer-based modules that operate in sequence.

Modality-Specific Encoding: Each modality is projected into a shared embedding dimension D to form tokens.

- *Visual Tokens:* A frozen pretrained backbone (e.g., DINOv3 [39]) encodes egocentric RGB frames. We use all tokens, not only the global token, to preserve spatial context. The resulting visual tokens are defined as $\mathbf{Z}^I \in \mathbb{R}^{T\times N\times D}$, where N includes patch, class, and register tokens per frame.
- *Robot Trajectory Tokens:* The robot kinematic history (e.g., waypoints) is encoded by an MLP into $\mathbf{Z}^R \in \mathbb{R}^{T\times 1\times D}$.
- *Goal Token:* The goal is projected into a single token $\mathbf{z}^g \in \mathbb{R}^D$, providing the directional context.

Body-Part-Aware Human Encoder: To extract expressive social cues, the human encoder partitions the J skeletal keypoints of M humans into a set of semantic body parts, $b \in \mathcal{B} = \{\text{Head, Torso, Arms, Legs}\}$. For each human m at time t, the skeletal keypoints $\mathbf{k}_{t,m}$ of each group are processed by a part-specific MLP to extract part-level features. We add a learnable part-type embedding to each of the four outputs, explicitly tagging each vector as belonging to a specific body region. These tagged tokens are then concatenated with the human's relative position $\mathbf{p}_{t,m}$ and facing direction $\mathbf{d}_{t,m}$, grounding the postural representation in the robot's egocentric frame. Finally, a fusion MLP projects this combined vector into a single D-dimensional human token $\mathbf{z}^H_{t,m}$, which encapsulates the human's posture and intent while preserving a fixed structural hierarchy. The resulting output is a sequence $\mathbf{Z}^H \in \mathbb{R}^{T\times M\times D}$.

Hierarchical Robot-Centric Fusion: We employ a hierarchical fusion strategy that mirrors the cognitive stages of navigation: social awareness, scene understanding, and goal-directed reasoning. The process is structured into three distinct stages:

- *Social-Robot Interaction:* Inspired by the Human Scene Transformer (HST) [11], we model the joint dynamics of the robot and humans. While we adopt the dual-stage temporal and social attention paradigm from HST, we extend this by enabling bidirectional robot-human attention. This ensures the robot's latent state is contextualized by human intent while explicitly modeling the robot's own influence on surrounding agents.
- *Vision-Robot Grounding:* The social-aware robot tokens from the preceding interaction layer act as queries in a cross-attention mechanism over the visual token sequence $\mathbf{Z}^I$. This attention is strictly time-aligned: the robot state at time t queries only the visual features from the corresponding frame. This ensures temporal consistency, allowing the model to ground its social reasoning within the physical constraints of the environment, such as obstacles or navigable pathways.
- *Goal-Robot Reasoning:* In the final stage, the enriched robot tokens are concatenated with the time-invariant goal token $\mathbf{z}^g$. A self-attention block facilitates a global reasoning step where temporal social-visual features are synthesized with the destination target to produce the final decision-making tokens.

By decoupling these interactions into a hierarchy, the model avoids the dilution of signal common in massive flat sequences and ensures that the robot remains the central agent throughout the reasoning process.

Trajectory Planning and Objective Function: The Transformer Encoder outputs a sequence of contextualized tokens. After hierarchical fusion, we obtain a sequence of $T+1$ tokens, comprising T temporally-contextualized robot tokens and one goal token. We flatten this sequence into a single vector of dimension $(T_{obs}+1)\times D$, which serves as the input to the Trajectory Head. We pass this vector to a Trajectory Head, composed of a 4-layer MLP. This head maps high-dimensional scene features to the robot's future path. The first two layers perform initial feature expansion and processing, which are then followed by a critical third layer, which later serves as the distillation target. This layer produces the latent representation $\mathcal{Z}_{\text{teacher}} \in \mathbb{R}^D$. This vector serves as a condensed semantic embedding, which encapsulates the high-fidelity social semantics extracted from skeletal and visual data. Finally, the fourth layer functions as the output head, generating the coordinate ($\hat{y} \in \mathbb{R}^2$) sequence for the predicted future trajectory $\hat{Y} \in \mathbb{R}^{T_{\text{pred}}\times 2}$. The training objective is to minimize the supervised trajectory prediction error. We employ a Mean Squared Error (MSE) loss between the predicted trajectory $\hat{Y}$ and the ground truth future trajectory Y_{GT}. By minimizing this loss, the Teacher model learns to optimize the latent manifold $\mathcal{Z}_{\text{teacher}}$ to be maximally predictive of socially compliant robot motion.

D. Knowledge Distillation

In this stage, the Student model operates under deployment constraints where human skeletal tracking is unavailable. The Student receives only the visual history $\mathbf{I}$ and the navigational goal g to plan the future trajectory $\hat{Y}_{\text{student}}$. To ensure real-time performance onboard mobile robots, we replace the Teacher's heavy backbone with a lightweight, trainable convolutional neural network (e.g., ResNet [40]). The Student transformer model maintains a similar structural flow, mapping visual tokens and the goal to a fused context. However, the internal Transformer and MLP layers are scaled down to reduce computational overhead for onboard real-time inference. We construct the input sequence by processing the available modalities into a streamlined token set. The visual history is encoded into T tokens, which are concatenated with a single token representing the goal embedding.

Feature Distillation Strategy: After the Student Transformer fuses the visual and goal embeddings, we extract *only* the output of a learnable CLS token, which is prepended to the input sequence and aggregates information from all other tokens through self-attention. This forces the model to compress the entire scene's social affordances into a compact belief state z_{student}. While inherently lossy, this compression is essential for low-latency inference. The burden of ensuring this single token captures complex social dynamics is handled by the distillation process. The distillation process leverages the frozen Teacher model to guide this compression. We extract the Teacher's latent representation $\mathcal{Z}_{\text{teacher}}$ from the third layer of its trajectory MLP, which encapsulates high-level reasoning about social interactions. Since the Student's compact belief state z_{student} may differ in dimensionality from the Teacher's latent space, we introduce a learnable projector ϕ. This projector maps the Student's token into the Teacher's feature space, allowing for direct alignment.

Objective Function: The Student is trained using a dual-objective loss function to implicitly consider the missing social cues, solely from visual inputs. The first component is the Feature Alignment Loss, $\mathcal{L}_{\text{feat}}$, which minimizes the distance between the Student's embedding $\mathcal{Z}_{\text{student}}$ and the Teacher's latent representation $\mathcal{Z}_{\text{teacher}}$. By targeting the manifold extracted from the third layer of the Teacher's trajectory head using an MSE criterion, the Student learns to approximate the Teacher's privileged social reasoning. Simultaneously, the Student must maintain accuracy in the primary navigation task through a Trajectory Imitation Loss, $\mathcal{L}_{\text{traj}}$. This component applies an MSE loss between the Student's predicted path $\hat{Y}_{\text{student}}$ and the ground truth trajectory Y_{GT}. The final Student policy is optimized by minimizing a weighted combination of these losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{traj}} + \lambda \mathcal{L}_{\text{feat}} \tag{1}$$

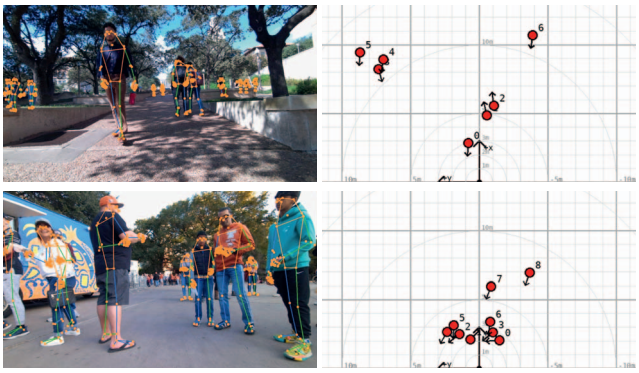
This training regime ensures that even without access to skeletal data at test time, the Student's navigation remains socially compliant by mimicking the Teacher's privileged understanding of the scene.

IV. Experimental Results

In this section, we outline the implementation details and datasets, followed by quantitative and qualitative results.

A. Experimental Setup

We implement HUMAIN with PyTorch and train our model on a single NVIDIA GeForce RTX 4070 Ti SUPER GPU with 16GB memory. Training is performed with the AdamW optimizer. We process the data at a frequency of 4Hz. The model utilizes an observation window of $T_{\text{obs}} = 6$ frames (approximately 1.5s) to capture immediate historical context and forecasts a future trajectory over a prediction horizon of $T_{\text{pred}} = 12$ frames (approximately 3.0s). For goal conditioning, the ground truth position at 15m ahead is defined as the target goal.



(a) Skeletal Keypoints (b) Position and Direction

Fig. 3: **Visualization of the Human State Representation.** We employ SAM3DBody [41] to extract social cues. (a) 3D skeletal keypoints overlaid on the robot's egocentric view. (b) A corresponding top-down perspective, where numbered nodes represent the relative 2D position, and arrows indicate the facing direction of each tracked human.

Dataset: We train our model on SCAND [42], a social robot navigation dataset, where robots were teleoperated to demonstrate socially compliant behaviors in human-populated environments. To reduce ambiguity and noise, we filter out scenes without humans within 15m or with minimal robot orientation change. The resulting dataset comprises about 12,000 samples for training, with an additional 2,000 samples reserved for testing. For fair comparison, we also evaluate on an independent Out-of-Distribution (OOD) dataset of about 2,000 samples collected by teleoperating an AgileX Scout Mini robot with similar socially compliant behaviors.

To obtain the privileged human-state inputs for the Teacher model, we apply SAM3DBody [41] together with ByteTrack [43] offline on RGB images to extract dense 3D human keypoints, as well as per-person 2D positions and facing directions used as training social cues (Fig. 3). We discard detailed hand and foot joints and retain 30 core keypoints corresponding to the head, torso, arms, and legs, supporting up to $M = 10$ tracked humans per scene.

Baselines: To validate HUMAIN, we compare against four state-of-the-art methods. We include ViNT [37], NoMaD [38], and GNM [44] as representative vision-only navigation baselines that share the same input modality as our Student. We also include HST [11] as a modular baseline that first predicts human trajectories and then plans robot motion, enabling a direct comparison between prediction-then-planning pipelines and our end-to-end distillation approach. Unlike modular systems that require explicit human tracking at inference time, HUMAIN runs from RGB images only.

To leverage HST [11]'s probabilistic forecasts, we construct time-varying keep-out zones by inflating each predicted mean of human positions into an obstacle whose radius is proportional to the forecast variance. A Model Predictive Controller (MPC) [45] planner then optimizes a robot trajectory that balances goal progress, smoothness, and control effort while avoiding these dynamic safety zones.

TABLE I: **Quantitative Results**: We compare HUMAIN with state-of-the-art baselines. **Bold** indicates best.

Method	ADE (m) ↓	FDE (m) ↓	AOE (rad) ↓
ViNT	0.756 ± 0.396	1.304 ± 0.788	0.203 ± 0.199
NoMaD	0.853 ± 0.486	1.561 ± 0.979	0.251 ± 0.265
GNM	0.753 ± 0.397	1.254 ± 0.778	0.219 ± 0.194
HST+MPC	0.891 ± 0.825	1.886 ± 1.009	0.313 ± 0.210
HUMAIN	**0.578 ± 0.374**	**1.064 ± 0.703**	**0.159 ± 0.153**

Metrics: Evaluating social navigation is challenging because factors like human comfort and perceived compliance are subjective [46]. Since our data consist of expert teleoperation demonstrations, we evaluate performance by measuring divergence from the ground-truth trajectories, treating them as a proxy for socially compliant behavior. We report three standard metrics:

- **Average Displacement Error (ADE)**: mean position error over all time steps.
- **Final Displacement Error (FDE)**: position error at the final prediction horizon.
- **Average Orientation Error (AOE)**: mean heading error relative to ground-truth orientation [47].

B. Quantitative Results

Table I summarizes the comparative performance of each method. HUMAIN consistently outperforms all baselines across the three metrics. Compared to the strongest baseline on each metric, HUMAIN reduces ADE by 23.2%, FDE by 14.2%, and AOE by 21.7%. The gains in ADE and AOE suggest that our model better captures the subtle, socially aware maneuvers demonstrated by human teleoperators. The improved FDE is largely due to training with explicit goal positions. Vision-only baselines such as ViNT, NoMaD, and GNM condition on goal images, which can be visually ambiguous and lack geometric precision for long-horizon navigation, whereas HUMAIN uses spatial coordinates that provide a more precise geometric objective.

Finally, while HST with MPC leverages high-fidelity human keypoints, its lower accuracy underscores a critical challenge. Integrating these forecasts into a decoupled MPC planner is inherently difficult, as the optimizer often struggles to balance goal-seeking with the rapidly shifting constraints of predicted keep-out zones. By learning social affordances in an end-to-end manner, HUMAIN bypasses the sensitivity of these decoupled pipelines, resulting in more robust and socially compliant performance.

C. Qualitative Results

Fig. 4 illustrates qualitative results by projecting predicted trajectories onto the robot's egocentric view. While quantitative analysis is conducted on out-of-distribution samples to ensure fairness, we also include results from SCAND test samples to demonstrate the results across diverse environmental contexts. We highlight several challenging social interaction scenarios to evaluate the model's behavior.

TABLE II: **Ablation Study Results**: We compare the effect of individual components of HUMAIN. **Bold** indicates best.

Method	ADE (m) ↓	FDE (m) ↓	AOE (rad) ↓
HUMAIN-*vision*	0.585 ± 0.367	1.078 ± 0.692	**0.157 ± 0.141**
HUMAIN-*hum*	0.594 ± 0.377	1.095 ± 0.712	0.157 ± 0.143
HUMAIN	**0.578 ± 0.374**	**1.064 ± 0.703**	0.159 ± 0.153

We observe that purely vision-based baselines, such as ViNT, NoMaD, and GNM, suffer from significant performance degradation in scenarios with high scene variance, like involving turns. These baselines tend to simply maintain a straight trajectory regardless of the context, often failing to account for human presence. In contrast, our method, along with HST+MPC, remains strictly goal-oriented. HST+MPC often produces non-smooth trajectories because keep-out zones severely restrict feasible paths, causing abrupt steering and oscillations. This suggests that our approach is better suited for robot navigation, where providing a positional goal is more effective for consistent movement. Furthermore, our method better captures social compliance, proactively adjusting its path and timing to minimize interference with pedestrians while maintaining steady goal progress.

D. Ablation Study Results

To evaluate the contribution of different components in HUMAIN, we perform two sets of ablation studies. First, we analyze the role of human whole-body pose cues to answer the question: *"Is Body Language Actually Useful?"* Second, we analyze the effect of auxiliary human trajectory prediction to answer the question: *"Does Explicitly Predicting Human Trajectories Help?"*

Ablations on Keypoints: We compare students distilled from our full Teacher model against those distilled from HUMAIN-*vision*, a teacher trained without skeletal input. As shown in Table II, the model distilled from the skeleton-aware teacher achieves lower ADE and FDE while HUMAIN-*vision* yields a marginally better AOE. The overall differences are modest, which is expected given that the student operates purely from visual input at test time and has no direct access to skeletal cues. Nevertheless, the improvement in displacement metrics suggests that skeleton-aware teacher representations encode richer spatial social semantics that are partially transferable through distillation. The slightly higher AOE of HUMAIN reflects more active heading adjustments when navigating around humans. HUMAIN adjusts its heading more proactively, resulting in more human-aware but slightly less direct trajectories.

Ablations on Human Trajectory Prediction: We investigate whether supervising the Teacher model with an additional auxiliary human trajectory prediction objective improves navigation performance. In this configuration, a lightweight MLP head is added to predict future positions of each tracked human from the contextualized human tokens produced by the social-robot interaction stage. We name this variant HUMAIN-*hum*. As shown in Table II, HUMAIN-

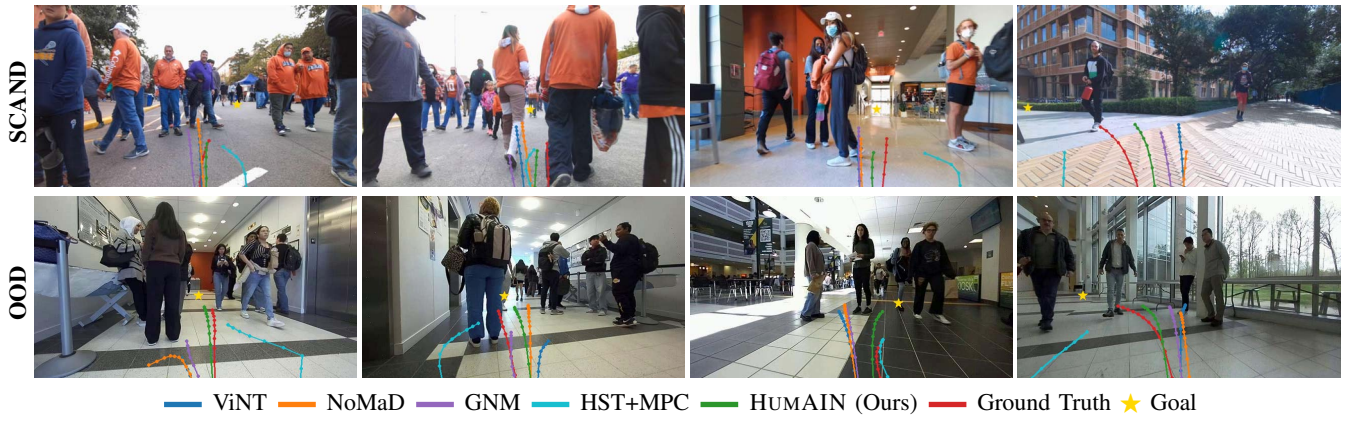


Fig. 4: **Qualitative Results:** Generated trajectories are projected onto the robot ego-centric view images using all the methods, ViNT [37] in blue, NoMaD [38] in orange, GNM [44] in purple, HST [11] with MPC [45] in cyan, HUMAIN in green, and the ground truth trajectory in red. The target goal is marked with a yellow star. The top row shows SCAND test dataset results, and the bottom row shows the results on the out-of-distribution dataset that we collected.

hum causes performance to degrade by 2.8% on average compared to our full model. We attribute this to the difficulty of explicitly predicting human trajectories, which is a distinct and non-trivial problem from robot trajectory planning. Jointly optimizing both objectives introduces conflicting gradient signals, ultimately hurting navigation performance. We therefore exclude it from our final model.

E. Robot Deployment

To validate the practicality of our approach, we deployed our model on a Clearpath Jackal equipped with a ZED2 RGB camera and NVIDIA GTX 1050 GPU. We compared our method against baseline methods in real-world scenarios. The navigation goal is fixed at a position 10m ahead of the robot. We evaluate on two representative social navigation scenarios: frontal human approach and intersection crossing. In both scenarios, baseline methods tend to prioritize following a straight path toward the goal, often failing to yield to nearby humans in a timely manner. In contrast, HUMAIN exhibits more conservative and human-aware behavior, proactively adjusting its path to avoid humans while maintaining progress toward the goal.

To further assess whether our model captures human motion intent, we test a scenario where a human approaches the robot diagonally, once from the right and once from the left. We observe that the robot consistently yields in the direction that minimizes interference with the approaching human: steering left when the human approaches from the right, and steering right when the human approaches from the left. This directionally-aware avoidance behavior suggests that HUMAIN appears to infer human motion intent from visual cues alone at deployment time, without any access to explicit skeletal or tracking information. Despite these promising behaviors, our current policy is not yet robust enough for general-purpose open-world deployment.

V. CONCLUSION

In this paper, we present HUMAIN, an end-to-end social robot navigation approach that leverages whole-body pose as an implicit cue for trajectory planning. We use a Teacher–Student strategy to distill privileged pose knowledge into a lightweight vision-only Student suitable for real-robot deployment, enabling socially compliant navigation without expensive skeletal tracking at runtime. Extensive comparisons against state-of-the-art baselines show that HUMAIN achieves better social compliance, with an average gain of 29.8%, demonstrating that implicit social awareness can be deployed on resource-constrained mobile platforms.

Despite these encouraging results, reliable real-world social navigation remains an open challenge. Although our policy exhibits promising behaviors in representative deployment scenarios, it is not yet robust enough for reliable deployment in unconstrained real-world environments. In particular, crowded scenes, occlusions, multiple interacting pedestrians, and unexpected human behaviors remain challenging. Furthermore, our approach is limited by the quality and diversity of available datasets. Since SCAND is collected through teleoperation, demonstrated trajectories may reflect operator preferences and interaction biases, making it difficult to learn broadly generalizable social behaviors. Future work will focus on improving robustness through larger-scale, more diverse datasets with crowded, multi-person interactions under natural human behavior. We also plan to explore more effective representation learning and distillation strategies that capture complex social dynamics while remaining suitable for real-time deployment.

VI. ACKNOWLEDGEMENT

George Mason University is supported by the National Science Foundation (NSF 2350352), the Army Research Office (W911NF2320004, W911NF2520011), Google DeepMind, Clearpath Robotics, FrodoBots Lab, Raytheon Technologies, Tangenta, Mason Innovation Exchange, and Walmart.

Ewha Womans University is supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. IITP-2026-RS-2020-II201460, Information Technology Research Center (ITRC)).

REFERENCES

[1] E. Ackerman, "How diligent's robots are making a difference in texas hospitals," *IEEE Spectrum*, March 2020. [Online]. Available: https://spectrum.ieee.org/how-diligents-robots-are-making-a-difference-in-texas-hospitals

[2] Starship Technologies, "Starship delivery robots: Revolutionizing last mile logistics," *Company White Paper*, 2021. [Online]. Available: https://www.starship.xyz/

[3] R. Mirsky, *et al.*, "Conflict avoidance in social navigation—a survey," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–36, 2024.

[4] R. Mirsky and E. Shpiro, "Recognition and identification of intentional blocking in social navigation," in *International Symposium on Technological Advances in Human-Robot Interaction*, 2024, pp. 101–110.

[5] T. Imai, *et al.*, "Interaction of the body, head, and eyes during walking and turning," *Experimental brain research*, vol. 136, no. 1, pp. 1–18, 2001.

[6] G. Ferrer, *et al.*, "Robot companion: A social-force based approach with human awareness-navigation in crowded environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 1688–1694.

[7] ——, "Social-aware robot navigation in urban environments," in *2013 European Conference on Mobile Robots*. IEEE, 2013, pp. 331–336.

[8] R. Mead and M. J. Matarić, "Autonomous human-robot proxemics: Socially aware navigation based on interaction potential," *Autonomous Robots*, vol. 41, pp. 1189–1201, 2017.

[9] K. Charalampous, *et al.*, "Recent trends in social aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 93, pp. 85–104, 2017.

[10] A. Alahi, *et al.*, "Social lstm: Human trajectory prediction in crowded spaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971.

[11] T. Salzmann, *et al.*, "Robots that can see: Leveraging human pose for trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7090–7097, 2023.

[12] S. Poddar, *et al.*, "From crowd motion prediction to robot navigation in crowds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 6765–6772.

[13] R. Kitagawa, *et al.*, "Human-inspired motion planning for omnidirectional social robots," in *ACM/IEEE international conference on human-robot interaction*, 2021, pp. 34–42.

[14] J. Gou, *et al.*, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[15] A. H. Raj, *et al.*, "Rethinking social robot navigation: Leveraging the best of two worlds," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 330–16 337.

[16] X. Xiao, *et al.*, "Appl: Adaptive planner parameter learning," *Robotics and Autonomous Systems*, vol. 154, p. 104132, 2022.

[17] Z. Wang, *et al.*, "Appli: Adaptive planner parameter learning from interventions," in *2021 IEEE international conference on robotics and automation (ICRA)*, 2021, pp. 6079–6085.

[18] D. M. Nguyen, *et al.*, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 7442–7447.

[19] H. Karnan, *et al.*, "Voila: Visual-observation-only imitation learning for autonomous navigation," in *International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2497–2503.

[20] W. Wang, *et al.*, "Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 11 348–11 355.

[21] A. Faust, *et al.*, "Prm-rl: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5113–5120.

[22] R. Chandra, *et al.*, "Multi-agent inverse reinforcement learning in real world unstructured pedestrian crowds," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 18 668–18 675.

[23] P. Ratsamee, *et al.*, "Social navigation model based on human intention analysis using face orientation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1682–1687.

[24] V. Narayanan, *et al.*, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8200–8207.

[25] D. Song, *et al.*, "Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, 2025.

[26] S. Narasimhan, *et al.*, "Olivia-nav: An online lifelong vision language approach for mobile robot social navigation," in *IEEE international conference on robotics and automation (ICRA)*, 2025, pp. 9130–9137.

[27] A. Payandeh, *et al.*, "Social-llava: Enhancing social robot navigation through human-language reasoning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 17 192–17 198.

[28] Z. Wang, *et al.*, "Maction-socialnav: Multi-action socially compliant navigation via reasoning-enhanced prompt tuning," *IEEE Robotics and Automation Letters*, 2026.

[29] A. Gupta, *et al.*, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2255–2264.

[30] Y. Gao, *et al.*, "Social-pose: Enhancing trajectory prediction with human body pose," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[31] S. Saadatnejad, *et al.*, "Social-transmotion: Promptable human trajectory prediction," in *International conference on learning representations*, vol. 2024, 2024, pp. 38 299–38 316.

[32] N. Nilavadi, *et al.*, "Uptor: Unified 3d human pose dynamics and trajectory prediction for human-robot interaction," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 13 927–13 933.

[33] Y. Bengio, *et al.*, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[34] M. Nazeri, *et al.*, "Vanp: Learning where to see for navigation with self-supervised vision-action pre-training," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 2741–2746.

[35] A. Payandeh, *et al.*, "Narrate2nav: Real-time visual navigation with implicit language reasoning in human-centric environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[36] M. Nazeri, *et al.*, "Verticoder: Self-supervised kinodynamic representation learning on vertically challenging terrain," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 6536–6543.

[37] D. Shah, *et al.*, "Vint: A foundation model for visual navigation," in *7th Annual Conference on Robot Learning*, 2023.

[38] A. Sridhar, *et al.*, "Nomad: Goal masked diffusion policies for navigation and exploration," in *2024 IEEE International Conference on Robotics and Automation*, 2024, pp. 63–70.

[39] O. Siméoni, *et al.*, "DINOv3," 2025. [Online]. Available: https://arxiv.org/abs/2508.10104

[40] K. He, *et al.*, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] X. Yang, *et al.*, "Sam 3d body: Robust full-body human mesh recovery," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 7209–7219.

[42] H. Karnan, *et al.*, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 807–11 814, 2022.

[43] Y. Zhang, *et al.*, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[44] D. Shah, *et al.*, "Gnm: A general navigation model to drive any robot," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7226–7233.

[45] C. E. Garcia, *et al.*, "Model predictive control: Theory and practice—a survey," *Automatica*, vol. 25, no. 3, pp. 335–348, 1989.

[46] A. Francis, *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–65, 2025.

[47] X. Liu, *et al.*, "Citywalker: Learning embodied urban navigation from web-scale videos," in *Computer Vision and Pattern Recognition Conference*, 2025, pp. 6875–6885.