GRACE: Group-Aware Socially Compliant Visual Navigation in Human Environments

Lulu Huang[1] **Wenbo Wang**[1] **Guo Cheng**[1] **Hao Cheng**[1]
Jian-Jian Jiang[1] **Wei-Shi Zheng**[1] **Jingke Meng**[1,2*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Key Laboratory of Tibetan Information Processing, Ministry of Education, Xining

{huangllu7,wangwb37,chengg33,chengh226,jiangjj35}@mail2.sysu.edu.cn
wszheng@ieee.org mengjk3@mail.sysu.edu.cn

Abstract: Visual navigation has made substantial progress in learning goal-reaching policies from egocentric observations. Most existing approaches primarily focus on geometric traversability, however, they do not explicitly account for the social constraints induced by human-populated environments where collision-free space can still be socially occupied, such as the space between walking partners or within conversational formations. To address this, we present GRACE, a group-aware framework for socially compliant visual navigation. The key insight of GRACE is to explicitly model social relations and group interaction spaces, allowing the robot to avoid paths that are physically feasible but socially undesirable. Specifically, our social perception module builds a visual pedestrian–robot graph from RGB observations and uses soft F-formation targets to supervise group-aware attention. This social representation guides a visual trajectory generator, which we further finetune using group-relative reward-weighted flow matching to preserve personal and group spaces. Real-world experiments on a Unitree Go2 demonstrate that GRACE significantly reduces group-space intrusions and collision events compared with representative baselines, while maintaining competitive navigation efficiency.

Keywords: Social Navigation, Visual Navigation

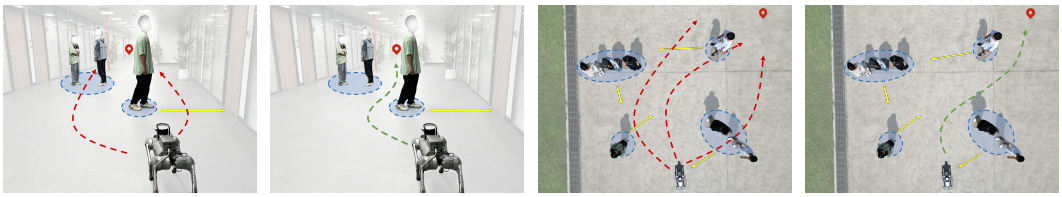
1 Introduction

Visual navigation [1, 2, 3, 4] aims to guide mobile robots toward specified destinations in real-world environments. Given RGB observations, a mobile agent perceives the surrounding scene and generates actions to reach the goal. Previous visual navigation works often use generative paradigms to improve navigation accuracy. Although they effectively model trajectory multi-modality and successfully reach destinations in static scenes, their performance is limited in human-populated environments where navigation requires understanding not only free space but also human motion and social interaction.

Navigation in human-populated scenes requires more than simply identifying collision-free spaces [5, 6, 7]. Using only RGB cameras, a robot must estimate not only pedestrian locations but also social cues, such as relative motions, group affiliations, and shared interaction spaces. However, raw RGB images do not directly provide these group-level structures. Furthermore, group boundaries are ambiguous, human motion is noisy, and interaction spaces change dynamically as people move. As illustrated in Fig. 1, goal-reaching visual navigation policies may still produce socially intrusive trajectories when pedestrian motion and group-level interaction spaces are not explicitly modeled.

*Corresponding author.

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.



(a) FlowNav (baseline) **(b)** GRACE (ours) **(c)** FlowNav (baseline) **(d)** GRACE (ours)

Figure 1: Illustration of human-shared navigation scenarios. FlowNav may reach the goal with socially intrusive paths, while GRACE uses pedestrian dynamics and group-level context to avoid moving pedestrians and preserve social spaces.

To address these challenges, we propose a group-aware framework for socially compliant visual navigation in human-populated environments, termed GRACE. The key insight is that we explicitly model the social relations and group-level interaction spaces of pedestrians, allowing the trajectory generation to avoid paths that are physically feasible but socially undesirable, such as cutting through a conversational group or intruding into a shared space.

Specifically, GRACE augments a goal-conditioned visual navigation backbone with a group-aware social perception module and a reward-weighted trajectory fine-tuning stage. Given monocular RGB observations, GRACE constructs a visual pedestrian–robot graph and uses soft F-formation targets to supervise group-aware graph attention. The resulting social representation conditions the visual trajectory generator alongside visual and goal tokens, allowing group structures and predicted human motions to shape the goal-directed trajectory generation. Furthermore, we fine-tune GRACE using group-relative reward-weighted flow matching. This favors trajectories that preserve personal and group spaces while reducing predicted robot-pedestrian conflicts. At deployment, GRACE uses only forward-facing RGB observations, without requiring external tracking or ground-truth group annotations.

Our contributions are summarized as follows:

- We introduce an RGB-based group-aware social perception module that learns group-aware graph attention from soft F-formation supervision, predicts short-horizon pedestrian motion, and summarizes group relations into a social token for trajectory generation.
- We propose a group-relative reward-weighted flow-matching finetuning procedure that ranks sampled trajectories according to goal progress, personal- and group-space preservation, and predicted robot-pedestrian conflicts.
- We evaluate GRACE on real-world Unitree Go2 deployments in four types of human-shared environments. The results demonstrate that our method reduces group-space intrusions and collision events compared with representative visual navigation baselines, while maintaining competitive navigation efficiency.

2 Related Work

Visual Navigation. Goal-conditioned visual navigation learns robot actions or waypoints from image observations and goals, reducing reliance on hand-designed maps or privileged state inputs. Recent large-scale visual navigation models, such as GNM and ViNT [1, 2], learn generalizable representations for goal-conditioned navigation across different robot platforms. Generative navigation policies further formulate action generation as trajectory or waypoint sampling. NoMaD [3] uses diffusion-based waypoint generation, while recent flow- and diffusion-based policies improve the efficiency, flexibility, and scalability of visual trajectory generation [4, 8, 9]. Beyond imitation learning, recent studies have explored reinforcement learning or reward-based fine-tuning for generative robot policies, including diffusion policy optimization, reinforcement fine-tuning of flow-matching policies, and reward-weighted flow-matching objectives [10, 11, 12]. These works suggest

a promising direction for aligning generative policies with task-level preferences. However, existing visual navigation objectives mainly emphasize goal reaching, visual feasibility, or geometric traversability, rather than social admissibility in human-shared environments. GRACE addresses this gap by conditioning generative waypoint prediction on group-aware social cues inferred from onboard RGB observations.

Social Navigation. Social robot navigation has been studied through both analytical and learning-based approaches. Analytical methods include reciprocal collision avoidance for velocity-space collision avoidance [13] and social-force models for hand-designed pedestrian interaction modeling [14]. Learning-based methods further improve adaptability around pedestrians through reinforcement learning [15, 16] and attention-based crowd interaction modeling [17, 18]. Such methods encode interactions from structured geometric and kinematic cues, including positions, velocities, and headings, while learning relations implicitly through task-oriented objectives.

Visual social-navigation methods also learn directly from human–robot interactions. SACSoN [19] minimizes counterfactual perturbations to pedestrian motion, and SELFI [20] combines offline model-based learning with online reinforcement learning to improve a visual policy around people. Group-aware approaches explicitly account for social structure: Katyal et al. [21] learn a reinforcement-learning policy that respects dynamic pedestrian groups, while *Following Social Groups* [22] tracks groups and joins crowd flows to support navigation in dense crowds. These works motivate preserving both individual motion and group structure.

Recent VLM-based methods further explore social navigation by leveraging pretrained vision-language models for semantic scene understanding, social rule grounding, or high-level navigation decision making [23, 24, 25, 26, 27]. While these methods benefit from the semantic understanding ability of VLMs, their large model scale and inference latency may hinder real-time deployment on resource-constrained robotic platforms.

Prior social navigation methods commonly model human interactions either implicitly from geometric states or through large vision-language models, but efficient relation modeling with explicit social-structure supervision remains less explored. GRACE addresses this gap by combining lightweight appearance-level visual features with geometric cues for human-relation modeling, and by introducing F-formation [28] theory as an auxiliary supervision signal that injects structured social-interaction priors.

3 Method

3.1 Problem Formulation and Overview

Problem Formulation. We study goal-conditioned visual navigation in human environments. At time t, the robot receives a history of forward-facing RGB observations $O_t = \{o_{t-T_{\text{obs}}+1}, \ldots, o_t\}$ and a goal image g. The robot also maintains the corresponding robot-state history $S_t^r = \{s_{t-T_{\text{obs}}+1}^r, \ldots, s_t^r\}$, where $s_t^r = (p_t^r, v_t^r, \psi_t^r)$, the superscript r denotes the robot, and p, v, and ψ denote position, velocity, and heading, respectively. The history supports ego-motion compensation; s_t^r defines the current planning frame. The policy outputs an H-step robot-centric waypoint trajectory $\hat{\tau}_t^r = (\hat{p}_{t+1}^r, \ldots, \hat{p}_{t+H}^r)$, where each waypoint $\hat{p}_{t+h}^r \in \mathbb{R}^2$ lies on the robot's local ground plane. Our objective is to learn a goal-conditioned policy $\pi_\theta(\hat{\tau}_t^r \mid O_t, g, S_t^r)$ that reaches the visual goal while generating socially compliant trajectories that avoid predicted robot–pedestrian conflicts and preserve both personal and group interaction spaces.

Overview. We propose **GRACE** (Group-Aware Socially Compliant Visual Navigation in Human Environments), an RGB-based goal-conditioned visual navigation framework for robots in human-shared spaces. As shown in Fig. 2, GRACE augments a visual trajectory-generation backbone with group-aware social perception. From monocular RGB observations, it estimates robot-centric pedestrian states, constructs a visual pedestrian–robot graph with group-aware attention supervised by soft F-formation targets, and uses the resulting social representation to condition waypoint generation. The model is trained by imitation learning and further aligned with social behavior through reward-

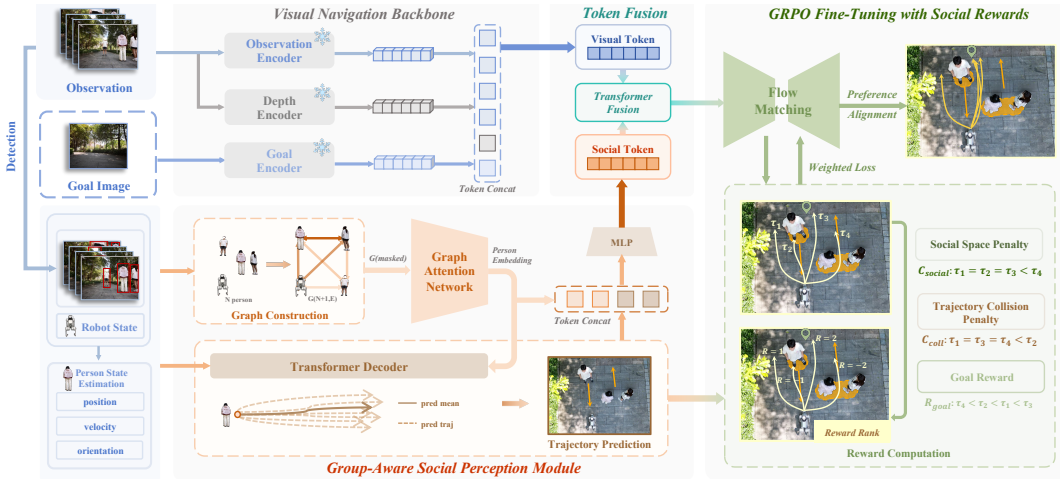


Figure 2: Overview of GRACE. The framework combines a goal-conditioned visual navigation backbone with group-aware social perception, which estimates pedestrian relations and future trajectories for visual-social action generation. GRPO fine-tuning further aligns the policy with goal progress, social-space preservation, and collision avoidance.

weighted flow matching. At deployment, GRACE uses only forward-facing RGB observations and the goal image, without LiDAR, depth sensors, external tracking, or ground-truth group annotations.

3.2 Group-Aware Social Perception Module

Socially compliant navigation requires more than avoiding pedestrians as independent moving obstacles. In human-shared environments, pedestrians may form socially coupled groups, and the space between interacting members may be geometrically free but socially occupied. GRACE therefore builds a robot-centric pedestrian–robot graph whose nodes represent pedestrians and the robot, and whose edges encode relative geometric, kinematic, and group-relation cues. The graph is encoded by an edge-aware Graph Attention Network (GAT) [29]. The resulting embeddings are used to infer soft pedestrian relations, forecast short-horizon human motion, and construct a social token for visual-social trajectory generation.

F-Formation-guided Relation Learning. F-formation theory describes how people engaged in a shared activity, such as conversation or walking together, arrange their positions and body orientations around a common interaction region, known as the *o-space* [28, 30]. We use this principle as an auxiliary social supervision signal for learning pedestrian relations. Instead of treating F-formation scores as hard group labels or directly imposing them as planning costs, we use them to guide a visual graph encoder to learn soft social affinities between pedestrians.

Formally, we construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N pedestrian nodes and one robot node. Each pedestrian node is initialized with a visual feature $\mathbf{g}_i$ extracted from its detected bounding box [31], while the robot node is initialized with a global ego-centric feature $\mathbf{g}_r$ extracted from the full current RGB observation. Pairwise edges are described by geometric and kinematic cues, $\mathbf{e}_{ij} = [\, d_{ij}, \Delta\psi_{ij}, \Delta v_{ij}, \Delta\theta_{ij}]$, where d_{ij} is the pairwise distance, $\Delta\psi_{ij}$ is the wrapped heading difference, Δv_{ij} is the relative speed, and $\Delta\theta_{ij}$ denotes the relative bearing. These edge features provide explicit spatial and motion context, while the node features preserve visual information about pedestrians and the surrounding scene.

We encode the visual-social graph with an edge-aware GAT. At layer ℓ, node features and edge attributes are projected into a shared attention space as $\tilde{h}_i^{(\ell)}$ and $\tilde{e}_{ij}^{(\ell)}$, with $h_i^{(0)} = \mathbf{g}_i$. The attention weight from node j to node i is computed by

$$a_{ij}^{(\ell)} = \mathrm{softmax}_{j\in\mathcal{N}(i)}\, \phi(\tilde{h}_i^{(\ell)}, \tilde{h}_j^{(\ell)}, \tilde{e}_{ij}^{(\ell)}), \qquad h_i^{(\ell+1)} = \sum_{j\in\mathcal{N}(i)} a_{ij}^{(\ell)} \tilde{h}_j^{(\ell)}, \tag{1}$$

where $\phi(\cdot)$ is a learnable attention function and $\mathcal{N}(i)$ contains valid neighboring nodes. After several GAT layers, the encoder outputs pedestrian embeddings $\{h_i\}_{i=1}^{N}$ and a robot embedding h_r.

To provide social supervision, we compute a soft F-formation target $\tilde{r}_{ij} = \exp\left(-\frac{d_{ij}}{\sigma_F}\right)\gamma_{ij}$, for each pedestrian pair according to their spatial distance and heading compatibility, where σ_F controls the spatial decay and γ_{ij} measures whether two pedestrians are facing each other or moving with compatible headings. This target is computed from geometric and motion cues and is used only as an auxiliary supervision signal. The full definition of γ_{ij} is given in Appendix B. We use this soft target to guide the final-layer graph attention. Let $\bar{a}_{ij}^{(L)}$ denote the attention weight from pedestrian j to pedestrian i in the last GAT layer, averaged over attention heads. We supervise the learned attention weights with a masked binary cross-entropy loss $\mathcal{L}_{\text{rel}} = -\sum_{i,j} f_{ij}\left[\tilde{r}_{ij}\log\bar{a}_{ij}^{(L)} + (1-\tilde{r}_{ij})\log(1-\bar{a}_{ij}^{(L)})\right]$, where f_{ij} filters invalid pedestrian (zero-padded placeholders used to align variable-size pedestrian sets in a batch) pairs. The relation loss is optimized jointly with the imitation learning objective, encouraging the visual graph encoder to assign stronger attention to socially related pedestrians without requiring manually annotated group labels.

Group-Conditioned Pedestrian Forecasting. For each pedestrian i, we predict future trajectories by conditioning its motion history on the group-aware GAT embedding h_i. Given the observed positions $p_{i,t-T_{\text{obs}}+1:t}$, we encode the history as $m_i = \text{Enc}_{\text{hist}}(p_{i,t-T_{\text{obs}}+1:t})$ and fuse it with the social context h_i. The decoder produces M future hypotheses using trainable queries Q:

$$\{\hat{\tau}_i^m\}_{m=1}^{M} = \text{Dec}_\omega\left(Q, m_i + W_f[\text{pool}(m_i), h_i]\right), \qquad Q = [q_1, \ldots, q_M] \in \mathbb{R}^{M\times d}. \tag{2}$$

Here, $\hat{\tau}_i^m = (\hat{p}_{i,t+1}^m, \ldots, \hat{p}_{i,t+H}^m)$ denotes the future trajectory predicted from the m-th query q_m. The learnable queries Q enable the shared decoder to generate M plausible future trajectories for pedestrian i. The decoder is trained with standard best-of-M ADE/FDE supervision [32, 33] against the ground-truth future trajectory, with invalid pedestrians masked.

We further convert the predicted pedestrian futures into a compact social token. For each pedestrian, we summarize its M hypotheses by their cross-mode mean and standard deviation, fuse this motion summary with h_i, pool over valid pedestrians, and concatenate the pooled human representation with the robot embedding h_r:

$$z_t^{\text{soc}} = \text{MLP}_{\text{soc}}\left(\left[\text{Pool}_i\left(\{[h_i,\ \text{Mean}_m(\hat{\tau}_i^m),\ \text{Std}_m(\hat{\tau}_i^m)]\}_{i=1}^{N}\right),\ h_r\right]\right). \tag{3}$$

The resulting token z_t^{soc} captures group-conditioned pedestrian motion and is used to condition the downstream visual-social trajectory generator. At each future step, the mean and sample standard deviation are computed across the M pedestrian hypotheses separately for each x/y coordinate (standard-deviation denominator $M-1$). Disagreement between modes therefore appears as uncertainty in the social token. This aggregation summarizes pedestrian forecasts; the policy generates the robot trajectory conditioned on the token.

3.3 GRPO Fine-Tuning with Social Rewards

Starting from a pretrained FlowNav policy [4] as the visual navigation backbone, we initialize GRACE by attaching the pretrained group-aware social perception module and conditioning the trajectory generator on the resulting social token. We then fine-tune this policy with reward-weighted flow matching [34, 35]. For each observation, GRACE samples a set of candidate robot trajectories from the current flow-based generator policy, evaluates them with a social reward, and normalizes the rewards within the same sampled set. The resulting group-relative weights encourage the trajectory generator to increase the likelihood of candidates that are both goal-directed and socially compliant, while retaining the goal-reaching prior inherited from the base FlowNav policy.

Social-Space-Inspired Reward. At each planning step t, the flow generator samples robot trajectory candidates $\hat{\tau}_t^{r,k} = (\hat{p}_{t+1}^{r,k}, \ldots, \hat{p}_{t+H}^{r,k})$. Our reward primarily encourages candidates to respect human social spaces and avoid time-aligned collisions with predicted pedestrian motion. We construct social spaces from the current pedestrian states using an F-formation detector consistent with

the soft formation score in Sec. 3.2. Each pedestrian induces a personal-space circle, while each detected group induces bridge-shaped capsule regions between socially related members. Let Ω_t denote all social-space elements, where each $\omega \in \Omega_t$ has radius or effective thickness ρ_ω.

The social-space intrusion and dynamic-collision costs are defined as

$$C_{\text{social}}(\hat{\tau}_t^{r,k}) = \sum_{h=1}^{H} \sum_{\omega \in \Omega_t} \left[1 - \frac{d(\hat{p}_{t+h}^{r,k}, \omega)}{\rho_\omega} \right]_+, \qquad C_{\text{coll}}(\hat{\tau}_t^{r,k}) = \sum_{h=1}^{H} \sum_{i=1}^{N} \left[1 - \frac{\|\hat{p}_{t+h}^{r,k} - \hat{p}_{i,t+h}\|_2}{r_c} \right]_+, \tag{4}$$

Here, $d(\hat{p}_{t+h}^{r,k}, \omega)$ measures the distance from the robot waypoint to a social-space element, $\hat{p}_{i,t+h}$ is the predicted pedestrian position at the same future step, r_c is the collision radius, and $[a]_+ = \max(a, 0)$.

The final reward penalizes these social violations while retaining a goal-progress term:

$$R(\hat{\tau}_t^{r,k}) = -\lambda_s C_{\text{social}}(\hat{\tau}_t^{r,k}) - \lambda_c C_{\text{coll}}(\hat{\tau}_t^{r,k}) + \lambda_g R_{\text{goal}}(\hat{\tau}_t^{r,k}), \tag{5}$$

where $\lambda_s, \lambda_c, \lambda_g > 0$, and $R_{\text{goal}}(\hat{\tau}_t^{r,k}) = \|p_t^r - p_g\|_2 - \|\hat{p}_{t+H}^{r,k} - p_g\|_2$ simply measures whether the candidate ends closer to the goal, p_g denotes the goal position .

Group-Relative Flow Matching Fine-Tuning. For the same condition (O_t, g, S_t^r), we sample K candidate trajectories $\mathcal{T}_t = \{\hat{\tau}_t^{r,k}\}_{k=1}^K$ from the current conditional flow generator and score each of them using Eq. (5). Rewards are normalized within this candidate set and converted into positive weights:

$$w_{t,k} = \min\left(\exp\left(\eta \, \text{Norm}_{\mathcal{T}_t}\left(R(\hat{\tau}_t^{r,k}) \right) \right), w_{\max} \right), \tag{6}$$

This relative normalization compares candidates under the same visual-social condition, making the update less sensitive to reward-scale differences across scenes.

The flow model is conditioned on the fused visual-social token $z_t = \text{Fuse}_\theta([z_t^{\text{vis}}, z_t^{\text{soc}}])$, where z_t^{vis} is the goal-conditioned visual token produced by the visual navigation backbone from the RGB observation history and the goal image. For each candidate, conditional flow matching (CFM) samples an interpolation time $\xi \sim \mathcal{U}(0, 1)$ and constructs an intermediate trajectory $x_\xi^{t,k}$ between Gaussian noise and the demonstration trajectory, together with its target velocity $u_\xi^{t,k}$. We fine-tune the conditional flow model by weighting this velocity-matching loss with $w_{t,k}$:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{t,k,\xi}\left[w_{t,k} \left\| v_\theta(x_\xi^{t,k}, \xi, z_t) - u_\xi^{t,k} \right\|_2^2 \right]. \tag{7}$$

This objective increases the likelihood of trajectories with higher relative social reward while retaining the standard flow-matching training structure. To preserve the imitation prior from demonstrations, we jointly optimize the reward-weighted objective and the original CFM loss:

$$\mathcal{L} = \lambda_{\text{grpo}} \mathcal{L}_{\text{GRPO}} + \lambda_{\text{cfm}} \mathcal{L}_{\text{CFM}}. \tag{8}$$

Here, $\mathcal{L}_{\text{CFM}}$ is the standard unweighted flow-matching loss [34], which stabilizes fine-tuning and prevents the policy from drifting too far from the demonstrated navigation behavior.

4 Experiments

We evaluate GRACE through real-world deployment on a Unitree Go2 quadruped robot in human-shared indoor and outdoor environments. Our experiments are designed to answer three questions: (i) whether GRACE reduces unsafe interactions and motion interference with moving pedestrians, (ii) whether it preserves group interaction spaces in dense human-shared environments, and (iii) whether it maintains goal-reaching performance while improving social compliance.

4.1 Experimental Setup

Robot platform and deployment. All experiments use a Unitree Go2 quadruped, RGB observations, and a goal image. Methods share the same controller, velocity limits, and waypoint-to-velocity

(a) Static FlowNav (b) Static GRACE (c) Dynamic FlowNav (d) Dynamic GRACE

Figure 3: Representative real-world rollouts of FlowNav and GRACE in static and dynamic multi-pedestrian scenarios. Compared with FlowNav, GRACE better preserves group interaction spaces in static scenes and anticipates pedestrian motion in dynamic scenes, leading to more socially compliant navigation.

conversion. GRACE runs on an external RTX 4060 workstation with approximately 51 ms end-to-end latency and 15 Hz closed-loop control.

Training data. We train on public real-world datasets and evaluate in unseen environments and pedestrian configurations. Data and training details are provided in the appendix.

Evaluation scenarios. The primary evaluation crosses pedestrian motion (static/dynamic) with crowd density (sparse/dense) (Table 1). Sparse settings contain three pedestrians; dense settings contain five to seven. Static scenes include isolated pedestrians and conversational groups. Dynamic scenes include crossings, pedestrians walking in the same or opposite direction, and moving groups. Methods share start–goal pairs and scripted pedestrian behaviors within each setting. Start and goal are approximately 10 m apart in static scenes and 15 m in dynamic scenes, with a 90 s time limit. Each method is evaluated over eight trials per static setting and six per dynamic setting. Additional social baselines and fine-grained ablations follow the same protocol in *both Static-Dense and Dynamic-Dense*, combined in Table 2.

Baselines. The primary comparison uses the official pretrained checkpoints of NoMaD [3] and FlowNav [4]. The additional comparison includes GSON [27] and SACSoN [19], adapted to our monocular RGB-only setting. Asterisks in Table 2 identify these adaptations.

Metrics and statistics. We report navigation time (NT), success rate (SR), personal-space violation duration (PSV), pedestrian collision rate (CR), group-space intrusion rate (GSIR), and pedestrian direction interference rate (PDIR). SR measures collision-free goal reaching within the time limit. PSV accumulates time within 0.25 m of a pedestrian [20, 36]; GSIR measures intrusions into group interaction spaces. PDIR applies only to dynamic scenes and counts encounters within 1 m that cause a pedestrian to stop or change direction. CR and PDIR use encountered pedestrians as denominators, whereas GSIR uses encountered groups; each pedestrian or group is counted once per trial upon entering the robot's field of view. Table 1 averages NT over all trials, assigning 90 s to failures. Table 2 instead reports $\mathrm{NT}_{\mathrm{succ}}$ over successful trials only; it must be read alongside SR and cannot be directly compared with NT. $\mathrm{NT}_{\mathrm{succ}}$/PSV are mean$\pm$SD(SD is omitted for a single successful trial), and CR/GSIR/PDIR are pooled encounter rates$\pm$trial-bootstrap SE; daggers mark $p < 0.05$ versus GRACE under exact paired tests.

4.2 Real-world Navigation Results

We conduct real-world trials across static and dynamic pedestrian scenarios, as summarized in Table 1, where GRACE achieves the lowest social-violation metrics among the compared methods. We further compare GRACE with two social navigation baselines, GSON and SACSoN, and conduct fine-grained ablations in both Static-Dense and Dynamic-Dense settings, as summarized in Table 2.

Static scenes. GRACE substantially reduces personal- and group-space violations compared with NoMaD and FlowNav. In the static-sparse setting, it achieves *zero* collisions and *zero* group-space intrusions. The improvement is more evident in dense layouts, where stationary pedestrians create constrained interaction spaces. As shown in Fig. 3(a,b), FlowNav may pass through the interaction space between seated pedestrians, while GRACE detours around the group and preserves progress

Table 1: Real-world navigation performance across controlled social navigation scenarios. The table compares GRACE with visual navigation baselines and ablated variants.

(a) Static pedestrian scenarios

	Static-Sparse					Static-Dense				
Method	NT↓	SR(%)↑	PSV↓	CR(%)↓	GSIR(%)↓	NT↓	SR(%)↑	PSV↓	CR(%)↓	GSIR(%)↓
Baselines										
NoMaD [3]	81.56	12.50	8.63	62.50	87.50	84.38	12.50	10.60	71.43	88.89
FlowNav [4]	65.53	37.50	7.04	56.25	75.00	74.50	25.00	7.97	58.82	81.81
Ablation Study										
GRACE w/o GRPO	49.06	62.50	3.63	33.33	25.00	56.42	62.50	3.88	35.00	28.57
GRACE w/o SP	36.47	75.00	1.97	18.75	12.50	51.50	62.50	3.38	25.00	16.67
GRACE	**27.13**	**87.50**	**0.56**	**0.00**	**0.00**	**42.63**	**75.00**	**2.81**	**11.76**	**7.69**

(b) Dynamic pedestrian scenarios

	Dynamic-Sparse						Dynamic-Dense					
Method	NT↓	SR(%)↑	PSV↓	CR(%)↓	GSIR(%)↓	PDIR(%)↓	NT↓	SR(%)↑	PSV↓	CR(%)↓	GSIR(%)↓	PDIR(%)↓
Baselines												
NoMaD [3]	79.05	16.67	6.67	68.75	83.33	77.78	81.17	16.67	11.50	43.33	75.00	46.67
FlowNav [4]	79.50	16.67	5.33	60.00	66.67	61.11	90.00	0.00	7.66	53.33	58.33	63.33
Ablation Study												
GRACE w/o GRPO	58.50	50.00	3.83	28.57	33.33	11.11	64.60	50.00	3.17	13.33	33.33	20.00
GRACE w/o SP	59.57	50.00	3.67	38.46	16.67	38.89	73.27	33.33	6.83	23.33	25.00	36.67
GRACE	**35.83**	**83.33**	**1.67**	**15.38**	**0.00**	**5.56**	**47.33**	**66.67**	**1.83**	**10.00**	**8.33**	**6.67**

Table 2: Dense-scene social baselines and fine-grained ablations: eight Static-Dense and six Dynamic-Dense trials per method. $*$: adapted baseline.

	Static-Dense					Dynamic-Dense					
Method	NT_{succ} ↓	SR(%)↑	PSV↓	CR(%)↓	GSIR(%)↓	NT_{succ} ↓	SR(%)↑	PSV↓	CR(%)↓	GSIR(%)↓	PDIR(%)↓
GSON*	$64.02_{\pm 3.88}$	**75.0**	**2.73**$_{\pm 3.03}$	$14.3_{\pm 8.4}$	$8.3_{\pm 7.4}$	$78.47_{\pm 1.82}$	50.0	$7.35_{\pm 0.54}^{\dagger}$	$23.3_{\pm 3.0}$	$36.4_{\pm 12.1}$	$40.0_{\pm 4.7}^{\dagger}$
SACSoN*	$31.72_{\pm 3.77}$	62.5	$5.88_{\pm 4.36}$	$42.9_{\pm 8.5}$	$63.6_{\pm 15.1}$	38.60	16.7	$6.97_{\pm 1.79}^{\dagger}$	$30.0_{\pm 4.1}$	$55.6_{\pm 10.4}$	$43.3_{\pm 5.6}$
Hard F	$30.42_{\pm 1.84}$	62.5	$2.93_{\pm 1.36}$	$18.8_{\pm 7.9}$	$10.0_{\pm 10.1}$	$27.57_{\pm 1.07}$	50.0	$2.54_{\pm 0.45}$	$16.7_{\pm 5.6}$	$18.2_{\pm 10.1}$	$16.7_{\pm 5.6}$
w/o Pred.	$30.42_{\pm 1.84}$	62.5	$3.18_{\pm 1.61}$	$26.3_{\pm 11.4}$	$15.4_{\pm 8.8}$	$41.95_{\pm 0.35}$	33.3	$6.58_{\pm 2.94}^{\dagger}$	$33.3_{\pm 7.7}$	$20.0_{\pm 10.6}$	$33.3_{\pm 3.8}^{\dagger}$
w/o GAT	$30.68_{\pm 1.88}$	50.0	$3.81_{\pm 0.79}$	$31.6_{\pm 12.6}$	$23.1_{\pm 9.3}$	$36.87_{\pm 0.70}$	50.0	$7.19_{\pm 1.19}^{\dagger}$	$20.0_{\pm 4.7}$	$22.2_{\pm 13.6}$	$26.7_{\pm 7.7}$
GRACE	**26.83**$_{\pm 1.17}$	**75.0**	$2.81_{\pm 2.00}$	**11.8**$_{\pm 7.4}$	**7.7**$_{\pm 6.9}$	**26.00**$_{\pm 1.41}$	**66.7**	**1.83**$_{\pm 2.23}$	**10.0**$_{\pm 6.3}$	**8.3**$_{\pm 7.6}$	**6.7**$_{\pm 3.8}$

toward the goal.The comparison with social navigation baselines in Table 2 provides a more nuanced picture. In Static-Dense, GSON matches GRACE's success rate and achieves slightly lower PSV, indicating comparable performance in respecting stationary pedestrians and groups. GRACE nevertheless completes successful trials considerably faster, with slightly lower CR and GSIR. SACSoN improves on NoMaD and FlowNav, but still exhibits lower success and more group-space intrusions than GRACE. Thus, GRACE combines competitive static social compliance with efficient goal reaching.

Dynamic scenes. GRACE shows its clearest advantage in dynamic scenes, achieving the lowest PDIR in both sparse and dense settings. This highlights the importance of anticipating pedestrian motion rather than relying solely on reactive avoidance. FlowNav performs well in static scenes, where reactive maneuvers can bypass stationary pedestrians, but becomes less reliable with moving pedestrians. As shown in Fig. 3(c,d), its immediate avoidance turn may steer the robot into the pedestrian's future path, whereas GRACE anticipates motion and adjusts earlier to avoid such conflicts. This advantage does not compromise group-space compliance: GRACE achieves *zero* group-space intrusions in Dynamic-Sparse and the lowest GSIR in Dynamic-Dense. Although GSON is competitive in Static-Dense, it incurs more group-space intrusions than GRACE in Dynamic-Dense. Overall, GRACE achieves lower PDIR and higher success rates than both social baselines, preserving group spaces while reducing interference with pedestrian motion.

4.3 Ablation Studies

To disentangle the contributions of social perception and reward-guided fine-tuning, we evaluate two ablated variants of GRACE. **w/o GRPO** retains the social perception stream but removes GRPO-based fine-tuning. **w/o SP** removes the social perception stream and applies GRPO fine-tuning without prediction-based conflict rewards. As shown in Table 1, both variants exhibit degraded navigation performance, with distinct effects on group-space preservation and interference with pedestrian motion.

Effect of social perception. Removing social perception mainly degrades performance in dynamic scenes. GRACE w/o SP exhibits higher PDIR and lower success rates than full GRACE, indicating that rewards based on the current social layout alone are insufficient when pedestrians are moving. Its higher GSIR across all four settings further suggests that explicit relation modeling helps the policy preserve group interaction spaces.

The fine-grained ablations in Table 2 clarify the contributions of the two components within this stream. Removing trajectory prediction (**w/o Pred.**) increases Dynamic-Dense CR and PDIR, highlighting the importance of anticipating pedestrian motion beyond recognizing the current social layout. Removing the graph attention network (**w/o GAT**) increases GSIR and PDIR, suggesting that pedestrian relations support both group-space preservation and motion-aware navigation. Both variants also have lower success rates and higher CR and GSIR in Static-Dense, consistent with the whole-stream ablation.

The **Hard F** variant replaces soft F-formation targets with binary labels. It yields similar PSV and GSIR in Static-Dense, but higher GSIR and PDIR in Dynamic-Dense. Despite similar group-level F1, soft supervision reduces forecasting errors (Appendix E), suggesting that graded relations benefit prediction and navigation beyond group detection.

Effect of GRPO fine-tuning. GRACE w/o GRPO retains social perception and achieves lower dynamic PDIR than GRACE w/o SP, showing that pedestrian and group-level cues remain useful without reward-guided fine-tuning. However, it incurs more collisions and group-space intrusions, and achieves lower success rates than full GRACE. Thus, providing social information alone is insufficient to obtain the full benefit of the framework. GRPO explicitly optimizes the policy to respect personal and group spaces and avoid predicted motion conflicts while maintaining progress toward the goal.

Overall, the ablations show that social perception and GRPO make complementary contributions: the former models group relations and anticipates pedestrian motion, while the latter encourages the policy to translate this information into socially compliant navigation.

5 Conclusion

We presented GRACE, a group-aware visual navigation framework that enables RGB-based robots to navigate more socially in human environments. GRACE incorporates group-aware social perception into a goal-conditioned visual navigation policy and uses reward-weighted flow matching to align trajectory generation with pedestrian motion, personal spaces, and group interaction spaces. We demonstrate its effectiveness through real-world experiments on a Unitree Go2 robot, where GRACE achieves more socially compliant navigation than representative visual navigation baselines while maintaining competitive goal-reaching performance.

6 Limitation

GRACE focuses on local, group-aware social navigation around the robot. While it captures short-horizon pedestrian interactions and group structures within the observed scene, it does not explicitly maintain long-term social context across long-horizon navigation. Future work could further improve visual social understanding by modeling finer-grained cues such as gaze, gestures, pose dynamics, and interaction context, while extending GRACE with a social memory for complex open-world environments.

Acknowledgments

This work was supported partially by the National Key Research and Development Program of China (No. 2023YFA1008502), Guangdong NSF Project (No. 2024A1515010101, No. 2023B1515040025), Guangdong Key Research and Development Program (No. 2024B0101040004, No. 2025B0909020002), the Open Project Program (Grant No. QHSF-CS-2601) of Key Laboratory of Tibetan Information Processing, Ministry of Education, and the Research and Application Project on Key Technologies for Embodied Intelligent Inspection of Power Grids in Mountainous Areas of Guizhou Province.

References

[1] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. GNM: A general navigation model to drive any robot. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2023.

[2] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. ViNT: A foundation model for visual navigation. In *Proceedings of the Conference on Robot Learning*, 2023.

[3] A. Sridhar, D. Shah, C. Glossop, and S. Levine. NoMaD: Goal masked diffusion policies for navigation and exploration. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2024.

[4] S. Gode, A. Nayak, D. N. Oliveira, M. Krawez, C. Schmid, and W. Burgard. FlowNav: Combining flow matching and depth priors for efficient navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025.

[5] A. Francis, C. Pérez-D'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, H.-T. L. Chiang, M. Everett, S. Ha, J. Hart, J. P. How, H. Karnan, T.-W. E. Lee, L. J. Manso, R. Mirsky, S. Pirk, P. T. Singamaneni, P. Stone, A. V. Taylor, P. Trautman, N. Tsoi, M. Vázquez, X. Xiao, P. Xu, N. Yokoyama, A. Toshev, and R. Martín-Martín. Principles and guidelines for evaluating social robot navigation algorithms. *ACM Transactions on Human-Robot Interaction*, 2025.

[6] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 2023.

[7] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami. A survey on socially aware robot navigation: Taxonomy and future challenges. *The International Journal of Robotics Research*, 2024.

[8] W. Cai, J. Peng, Y. Yang, Y. Zhang, M. Wei, H. Wang, Y. Chen, T. Wang, and J. Pang. NavDP: Learning sim-to-real navigation diffusion policy with privileged information guidance. *arXiv preprint arXiv:2505.08712*, 2025.

[9] D. Gao, B. Zhao, A. Lee, I. Chuang, H. Zhou, H. Wang, Z. Zhao, J. Zhang, and I. Soltani. VITA: Vision-to-action flow matching policy. *arXiv preprint arXiv:2507.13231*, 2025.

[10] A. Z. Ren, J. Lidard, L. Ankile, A. Simeonov, P. Agrawal, A. Majumdar, B. Burchfiel, H. Dai, and M. Simchowitz. Diffusion policy policy optimization. In *Proceedings of the International Conference on Learning Representations*, 2025.

[11] T. Zhang, C. Yu, S. Su, and Y. Wang. ReinFlow: Fine-tuning flow matching policy with online reinforcement learning. *arXiv preprint arXiv:2505.22094*, 2025.

[12] S. Pfrommer, Y. Huang, and S. Sojoudi. Reinforcement learning for flow-matching policies. *arXiv preprint arXiv:2507.15073*, 2025.

[13] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha. Reciprocal n-body collision avoidance. In *Robotics Research: The 14th International Symposium*, 2011.

[14] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 1995.

[15] Y. F. Chen, M. Everett, M. Liu, and J. P. How. Socially aware motion planning with deep reinforcement learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[16] M. Everett, Y. F. Chen, and J. P. How. Collision avoidance in pedestrian-rich environments with deep reinforcement learning. *IEEE Access*, 2021.

[17] C. Chen, Y. Liu, S. Kreiss, and A. Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2019.

[18] S. Liu, P. Chang, Z. Huang, N. Chakraborty, K. Hong, W. Liang, D. L. McPherson, J. Geng, and K. Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2023.

[19] N. Hirose, D. Shah, A. Sridhar, and S. Levine. SACSoN: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 2023.

[20] N. Hirose, D. Shah, K. Stachowicz, A. Sridhar, and S. Levine. SELFI: Autonomous self-improvement with RL for vision-based navigation around people. In *Proceedings of the Conference on Robot Learning*, 2025.

[21] K. Katyal, Y. Gao, J. Markowitz, S. Pohland, C. Rivera, I.-J. Wang, and C.-M. Huang. Learning a group-aware policy for robot navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022.

[22] X. Yao, J. Zhang, and J. Oh. Following social groups: Socially compliant autonomous navigation in dense crowds. *arXiv preprint arXiv:1911.12063*, 2019.

[23] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha. VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 2025.

[24] K. Weerakoon, M. Elnoor, G. Seneviratne, V. Rajagopal, S. H. Arul, J. Liang, M. K. M. Jaffar, and D. Manocha. BehAV: Behavioral rule guided autonomy using VLMs for robot navigation in outdoor scenes. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2025.

[25] Z. Fang, A. Xiao, D. Hsu, and G. H. Lee. From obstacles to etiquette: Robot social navigation with VLM-informed path selection. *IEEE Robotics and Automation Letters*, 2026.

[26] A. J. Sathyamoorthy, K. Weerakoon, M. Elnoor, A. Zore, B. Ichter, F. Xia, J. Tan, W. Yu, and D. Manocha. CoNVOI: Context-aware navigation using vision language models in outdoor and indoor environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.

[27] S. Luo, P. Sun, J. Zhu, Y. Deng, C. Yu, A. Xiao, and X. Wang. Gson: a group-based social navigation framework with large multimodal model. *IEEE Robotics and Automation Letters*, 2025.

[28] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*. CUP Archive, 1990.

[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

[30] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PLOS ONE*, 2015.

[31] MMPose Contributors. MMPose: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[32] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[33] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[34] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations*, 2023.

[35] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[36] S. Narasimhan, A. H. Tan, D. Choi, and G. Nejat. OLiVia-Nav: An online lifelong vision language approach for mobile robot social navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2025.

Appendix

Training data overview. All training data come from public real-world datasets, while real-world evaluation is conducted in unseen environments and pedestrian configurations. The FlowNav backbone is trained on Go Stanford [1], RECON [2], Tartan Drive [3], SACSoN [4], SCAND [5], and MuSoHu [6]. For social perception, we use 61,628 relation samples from SCAND and MuSoHu to train the relation encoder, and 12,712 trajectories from SiT [7] to train the pedestrian trajectory predictor. Relation targets are constructed from RGB-estimated pedestrian geometry and headings, whereas forecasting ground truth is obtained from SiT 3D annotations. GRPO fine-tuning uses 74,017 samples from SCAND and MuSoHu.

A RGB-based Pedestrian Localization

The social perception module uses the same RGB-only pedestria1n-localization pipeline during training and deployment. We use this pipeline on two robot navigation datasets, SCAND [5] and MuSoHu [6], which have different camera configurations. Pedestrian positions are estimated using the dataset-specific camera calibration parameters.

For each RGB observation, we apply MMPose [8] in a top-down pose-estimation pipeline. A pedestrian detector first produces a bounding box $b_i = (x_1, y_1, x_2, y_2)$ for pedestrian i, and the pose estimator then predicts body keypoints within the box. We determine the image-space foot point from reliable ankle keypoints. When both ankles are reliable, we use their midpoint; when only one ankle is reliable, we use the corresponding ankle location. If neither ankle is reliable, we use the bottom centre of the bounding box, $u_i = (x_1 + x_2)/2$ and $v_i = y_2$. This fallback provides a stable approximation of the pedestrian–ground contact point under partial occlusion or low-confidence pose estimates.

Figure 4: Examples of RGB-based pedestrian localization on MuSoHu (top) and SCAND (bottom). The RGB context frames show the detected pedestrians, their image-space foot points, and the corresponding metric positions. The rightmost panel shows the estimated pedestrian trajectories in the reference robot frame used by the social perception module.

We project each image-space foot point onto the ground plane to estimate its metric position. For a context frame $t' \in [t - T_{\text{obs}} + 1, t]$, let $\tilde{u}_{i,t'} = [u_{i,t'}, v_{i,t'}, 1]^\top$ denote the homogeneous image coordinate of the foot point. Let K_{cam} be the camera intrinsic matrix, R_{rc} the camera-to-robot rotation matrix, and $o^r = (0, 0, h)^\top$ the camera origin in the robot frame. The ground-plane position

is computed as

$$\bar{p}_{i,t'} = \left(o^r - \frac{h}{d_{i,t',z}} d_{i,t'} \right)_{xy}, \qquad d_{i,t'} = R_{rc} K_{\text{cam}}^{-1} \tilde{u}_{i,t'}. \tag{9}$$

Here, $\bar{p}_{i,t'}$ is the pedestrian position in the robot frame at context frame t'. The vector $d_{i,t'}$ is the corresponding viewing-ray direction in this frame, and $d_{i,t',z}$ is its vertical component. The scale factor $-h/d_{i,t',z}$ gives the intersection of the viewing ray with the ground plane. We discard invalid projections and physically implausible ground-plane intersections.

To construct a temporally consistent pedestrian history, we associate detections across context frames using Hungarian matching. The matching cost combines image-space bounding-box overlap with metric distance. Because the robot moves during the observation window, we first transform the associated detections from their frame-specific robot coordinates into a common world frame. We then express the resulting positions in the current robot frame:

$$p^w_{i,t'} = R(\psi^r_{t'}) \bar{p}_{i,t'} + p^r_{t'}, \qquad p_{i,t'} = R(\psi^r_t)^\top \left(p^w_{i,t'} - p^r_t \right). \tag{10}$$

Here, $p^r_{t'}$ and $\psi^r_{t'}$ denote the robot position and heading at context frame t', respectively. The current robot frame defined by s^r_t in Sec. 3 serves as the reference frame for downstream social perception and planning. The resulting sequence $p_{i,t-T_{\text{obs}}+1:t}$ forms the pedestrian history used by the social perception module in Sec. 3.2.

B Soft F-formation and GAT Relation Learning

We pretrain the GAT relation encoder using soft supervision targets derived from F-formation theory [9]. F-formations describe groups of people who maintain a shared interaction space by standing near one another and orienting their bodies toward a common conversational or walking configuration. Following this principle, we model pairwise social relatedness as a continuous affinity: spatially close and orientation-compatible pedestrians receive larger values, while weakly related pairs receive smaller values instead of being assigned binary group labels.

For each valid pedestrian pair (i, j) in the last observed frame, we construct a soft relation score that gives larger values to pairs that are more likely to belong to the same social interaction. The first component measures spatial proximity, since F-formation members are expected to remain close to one another. Let $d_{ij} = \|p_i - p_j\|_2$ be their distance in the current robot frame and let $\phi_{ij} = \text{atan2}(p_{i,y} - p_{j,y}, p_{i,x} - p_{j,x})$ denote the bearing from pedestrian j to pedestrian i. The proximity term follows the same exponential decay as in Sec. 3.2, with $\sigma_F = 0.45/\log 2$ making the affinity decay to one half at the same 0.45 m distance used by the social-reward group detector. Since F-formations are characterized not only by spatial proximity but also by body orientation patterns, we further compute an orientation compatibility score:

$$c_{ij} = \max \left[\frac{\cos(\psi_i - (\phi_{ij} + \pi)) + \cos(\psi_j - \phi_{ij})}{2}, \cos(\psi_i - \psi_j) \right], \tag{11}$$

where ψ_i and ψ_j are pedestrian headings estimated from the observation history. The score c_{ij} measures heading compatibility but is not directly suitable as a multiplicative factor because it can be negative. We therefore map it to the bounded non-negative orientation term used in Sec. 3.2, $\gamma_{ij} = \text{clip}(0.2 + 0.8(c_{ij} + 1)/2, 0.2, 1.0)$, which keeps orientation as a continuous modulation of the distance-based affinity while avoiding hard binary decisions from noisy visual heading estimates. The final soft F-formation target is then

$$\tilde{r}_{ij} = \exp\left(-\frac{d_{ij}}{\sigma_F} \right) \gamma_{ij}, \qquad \tilde{r}_{ii} = 0, \tag{12}$$

with invalid or padded pedestrian pairs filtered by the mask f_{ij} in the relation loss below. This target assigns the strongest supervision to nearby pedestrians whose distance and heading cues suggest a shared social interaction, and weak or zero supervision to distant or invalid pairs.

The relation encoder is a three-layer edge-aware GAT with four attention heads per layer, hidden dimension 128, and four-dimensional edge attributes containing relative distance, heading difference, velocity difference, and relative bearing. The GAT produces attention-derived relation scores from the final-layer attention weights. Following Sec. 3.2, let $\bar{a}_{ij}^{(L)}$ denote the attention weight from pedestrian j to pedestrian i in the last GAT layer, averaged over attention heads. We supervise these scores with the same masked soft-label BCE relation loss:

$$\mathcal{L}_{\text{rel}} = -\sum_{i,j} f_{ij} \left[\tilde{r}_{ij} \log \bar{a}_{ij}^{(L)} + (1 - \tilde{r}_{ij}) \log(1 - \bar{a}_{ij}^{(L)}) \right], \tag{13}$$

where f_{ij} filters invalid pedestrian pairs. This objective treats $\tilde{r}_{ij}$ as a soft target rather than a hard group label, encouraging socially related pairs to receive larger attention-derived scores and weakly related pairs to receive smaller scores. We also add a small auxiliary margin ranking term that preserves the relative ordering of raw soft F-formation strengths:

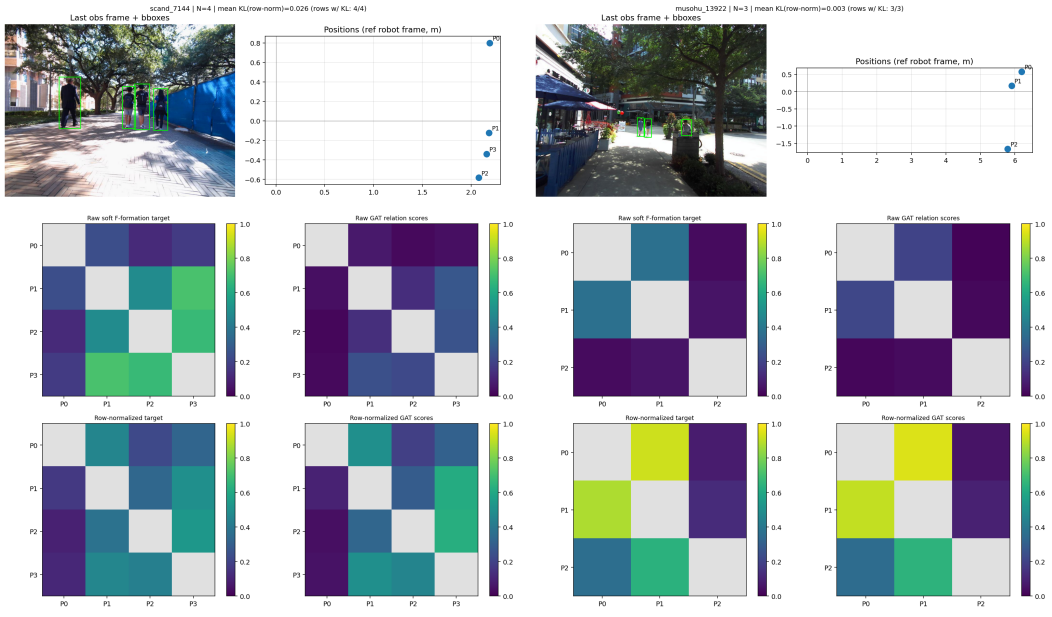
$$\mathcal{L}_{\text{margin}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \left[\bar{a}_i^{\text{far}} - \bar{a}_i^{\text{near}} + 0.3 \right]_+ . \tag{14}$$

The GAT relation pretraining objective is

$$\mathcal{L}_{\text{GAT}} = \lambda_{\text{rel}} \mathcal{L}_{\text{rel}} + \lambda_{\text{m}} \mathcal{L}_{\text{margin}}, \tag{15}$$

with $\lambda_{\text{rel}} = \lambda_{\text{m}} = 1$ in our default configuration. Training uses cached SCAND and MuSoHu perception samples, including robot-frame pedestrian histories, RGB context images, and detected bounding boxes. The learned attention-derived relation scores are later used by the trajectory predictor and by the social-token aggregation module.

Figure 5 shows qualitative examples of GAT relation pretraining on SCAND and MuSoHu. In each heatmap, row i denotes the anchor pedestrian and column j denotes the neighbor being attended to; brighter colors indicate stronger relations. The raw matrices show the pairwise soft targets and attention-derived scores used by the BCE supervision. The row-normalized matrices are included only for visualization, making it easier to compare each pedestrian's relative neighbor-attention pattern.



(a) SCAND (b) MuSoHu

Figure 5: Visualization of GAT relation pretraining on SCAND and MuSoHu examples. Each panel shows the last RGB observation with detected pedestrian boxes, the estimated pedestrian positions in the reference robot frame, the raw soft F-formation target $\tilde{r}_{ij}$, the raw final-layer attention scores $\bar{a}_{ij}^{(L)}$, and their row-normalized versions for visual comparison.

C Trajectory Predictor Pretraining

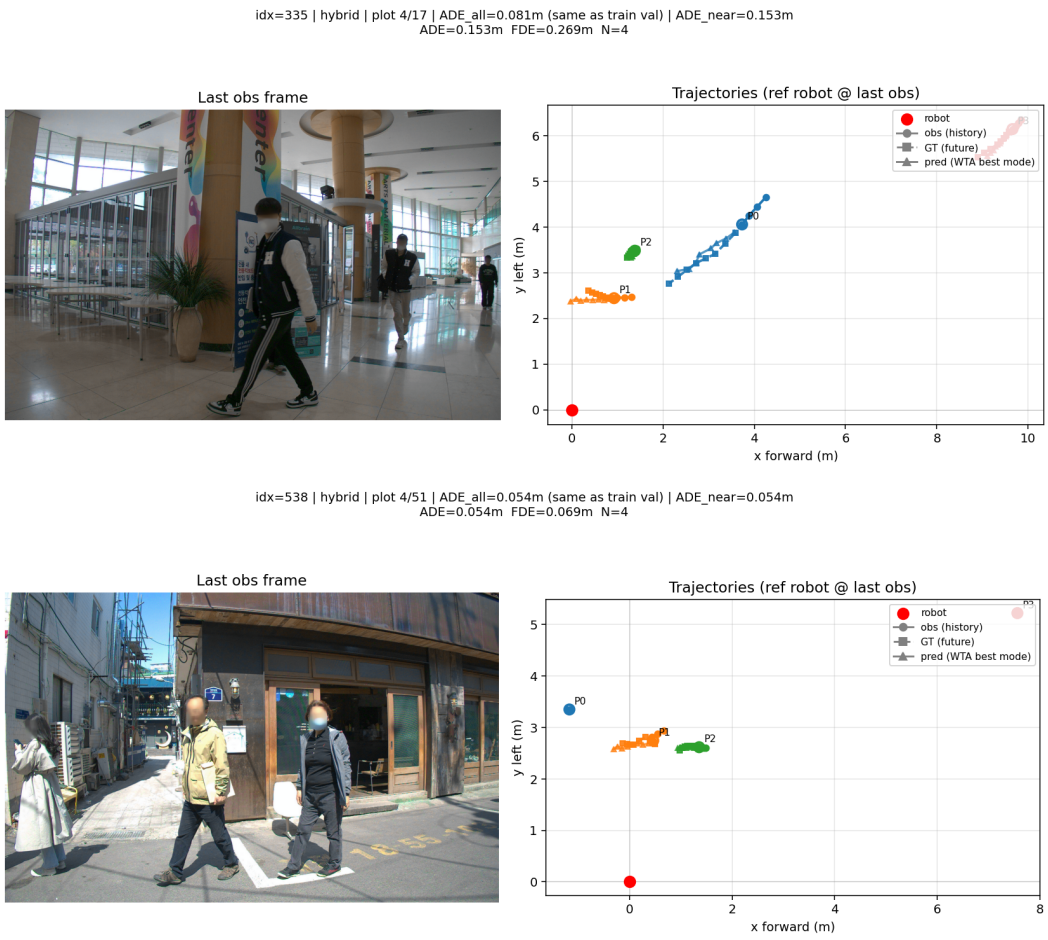


Figure 6: Qualitative pedestrian-trajectory predictions on the SiT validation split. Each example shows the final observed RGB frame and the corresponding trajectories in the reference robot frame. Circles indicate observed pedestrian positions, squares indicate ground-truth future positions, and triangles indicate the winner-takes-all selected prediction.

After pretraining the GAT relation encoder, we pretrain the pedestrian trajectory predictor on the SiT dataset [7]. This stage uses 12,712 trajectories. We use the hybrid SiT format, in which pedestrian trajectories are obtained from the 3D annotations. The pedestrian and ego trajectories are transformed into the vehicle frame of the final observed context frame. This coordinate convention matches the reference robot frame used by the RGB perception pipeline in Appendix A.

Each SiT sample also provides RGB context images and 2D pedestrian bounding boxes from the selected camera. These inputs are processed by the image node encoder when visual node features are enabled. Each training window retains pedestrians that remain visible throughout the complete observation and prediction horizons. The resulting sample contains the observed pedestrian histories $\{p_{i,t-T_{\text{obs}}+1:t}\}_{i=1}^{N}$, the future trajectories $\{\tau_i\}_{i=1}^{N}$, and a valid-pedestrian mask, where $\tau_i = (p_{i,t+1}, \ldots, p_{i,t+H})$ denotes the ground-truth future trajectory of pedestrian i. The observed and future trajectories form tensors in $\mathbb{R}^{N\times T_{\text{obs}}\times 2}$ and $\mathbb{R}^{N\times H\times 2}$, respectively.

We observe $T_{\text{obs}} = 4$ frames and use a prediction horizon of $H = 8$ frames. When visual node features are enabled, pedestrian crops are resized to 96×96 pixels. The GAT encoder is initialized

from its relation-pretraining checkpoint and frozen by default. The image node encoder is also frozen, and only the trajectory predictor is optimized.

Figure 6 shows qualitative predictions on the SiT validation split. The selected modes closely follow the corresponding ground-truth trajectories.

For each pedestrian i, the observed positions $p_{i,t-T_{\text{obs}}+1:t}$ are encoded into a sequence of history tokens from frame-to-frame displacements and speeds, while the pretrained GAT encoder provides the corresponding group-aware pedestrian embedding h_i and attention-derived relation scores. The trajectory predictor uses a three-layer, eight-head Transformer decoder with 128-dimensional hidden states. A relation-weighted aggregation of GAT embeddings is fused with the history summary and injected into the history-token memory. The decoder uses $M = 6$ learnable mode queries to generate future trajectory hypotheses $\{\hat{\tau}_i^m\}_{m=1}^M$, where $\hat{\tau}_i^m = (\hat{p}_{i,t+1}^m, \ldots, \hat{p}_{i,t+H}^m)$. Given the ground-truth future trajectory τ_i, we select the mode with the lowest average displacement error (ADE):

$$m_i^\star = \arg\min_m \text{ADE}(\hat{\tau}_i^m, \tau_i)\,. \tag{16}$$

We then optimize a winner-takes-all objective that combines the ADE and final displacement error (FDE) of the selected mode:

$$\mathcal{L}_{\text{traj}} = \text{ADE}\Big(\hat{\tau}_i^{m_i^\star}, \tau_i\Big) + \text{FDE}\Big(\hat{\tau}_i^{m_i^\star}, \tau_i\Big)\,. \tag{17}$$

The loss is averaged over valid pedestrians only. We select the best pretraining checkpoint according to ADE and FDE on the SiT validation split.

We train the trajectory predictor for 80 epochs using AdamW with a learning rate of 3×10^{-4}, a weight decay of 10^{-4}, and a gradient-norm clipping threshold of 1.0. This pretraining stage adapts the trajectory predictor to multimodal pedestrian motion represented in the same robot-centric coordinate frame used by the downstream navigation policy.

D Main Navigation Training

After loading the pretrained social perception modules described above, we train the complete navigation policy on 74,017 samples from the public SCAND and MuSoHu datasets. Each training sample contains RGB context frames, a goal image, robot trajectory and pose information, and social observations extracted by the social perception pipeline.

The social perception module operates online during navigation training. The RGB context observations are first converted into robot-centric pedestrian histories using the pipeline in Appendix A. These histories are encoded by the soft F-formation-guided GAT and passed to the pretrained trajectory predictor. The predicted pedestrian futures and graph representations are then aggregated into the social token that conditions the navigation model.

The main training objective follows Eq. (8) and combines the standard conditional flow matching (CFM) loss with the reward-weighted GRPO loss:

$$\mathcal{L} = \lambda_{\text{grpo}}\mathcal{L}_{\text{GRPO}} + \lambda_{\text{cfm}}\mathcal{L}_{\text{CFM}}. \tag{18}$$

We set $\lambda_{\text{cfm}} = 0.375$ and $\lambda_{\text{grpo}} = 0.625$. For each training input, we sample $K = 10$ candidate robot trajectories for the GRPO objective. The GAT encoder and trajectory predictor are initialized from their best pretraining checkpoints. During main training, the pretrained social modules are optionally fine-tuned with a small learning rate.

The GRPO reward is computed from the social layout inferred from the current observation. To obtain group-level social spaces, we first cluster pedestrians using the same $0.45\,\text{m}$ distance scale as the soft F-formation supervision in Appendix B. The detector keeps groups with 2–8 members, splits spatially elongated components, and uses body orientations to separate inconsistent groups. It also merges nearby groups when their members face one another.

The detected groups are then converted into social-space constraints. Each pedestrian defines a circular personal space with radius $0.5\,\text{m}$. For each group, we construct group interaction spaces as

bridge-shaped capsules between selected member pairs. When a group has more than two members, the selected pairs are given by a minimum spanning tree, with edges longer than $2.5\,\mathrm{m}$ discarded.

For each sampled robot trajectory, we compute the reward using the terms defined in Eq. (5):

$$R(\hat{\tau}_t^{r,k}) = -\lambda_s C_{\mathrm{social}}(\hat{\tau}_t^{r,k}) - \lambda_c C_{\mathrm{coll}}(\hat{\tau}_t^{r,k}) + \lambda_g R_{\mathrm{goal}}(\hat{\tau}_t^{r,k}), \qquad \lambda_s = 10.0, \quad \lambda_c = 5.0, \quad \lambda_g = 1.0. \tag{19}$$

The social-space cost C_{social} sums normalized intrusions of the robot waypoints into both personal-space circles and group-interaction capsules. We record the personal-space and group-space components separately for analysis, but they share the same weight in the training objective. The goal-progress reward R_{goal} is the reduction in distance to the goal from the first to the final trajectory waypoint. When predicted pedestrian futures are available, the collision cost C_{coll} sums time-aligned robot–pedestrian intrusions within radius $r_c = 0.8\,\mathrm{m}$.

During training, we independently mask the goal image with probability 0.5, following the configuration of the visual navigation backbone.

E Perception Evaluation

We evaluate group perception on the SCAND and MuSoHu test sets and pedestrian forecasting on the SiT test set. The pairwise group-relation precision/recall/F1 are 88.6/39.5/54.6%; group-level precision/recall/F1 are 75.4/68.4/71.7%, using an IoU threshold of at least 0.5 for group matching. The lower pairwise recall indicates that some social links remain undetected, despite the stronger group-level scores.

Table 3: Perception results with soft and hard F-formation supervision. Group F1 is evaluated on SCAND/MuSoHu; forecasting errors on SiT are best-of-six and measured in meters.

Supervision	Group F1(%)↑	minADE↓	minFDE↓
Hard F	71.3	0.1908	0.2833
Soft F (GRACE)	**71.7**	**0.0596**	**0.1040**

The reported minADE/minFDE are best-of-six average/final displacement errors, evaluated on valid pedestrian forecasts with the observation and prediction horizons in Appendix C. These best-of-six metrics measure coverage by the predicted hypotheses, rather than the error of their average or of the robot's executed trajectory. Soft supervision gives similar group F1 but lower forecast errors than hard supervision (Table 3), consistent with the Dynamic-Dense ablations in Table 2.

F Participants, Safety, and Non-Scripted Demonstrations

Pedestrians in the controlled navigation experiments were recruited volunteers who followed pre-scripted behaviors to make comparisons across methods consistent. Static behaviors included conversations and standing or seated pedestrians; dynamic behaviors included individual and group crossings and approaches. All participants provided informed consent. A safety operator and an emergency-stop procedure were used throughout the experiments.

We additionally demonstrate GRACE in non-scripted human environments with naturally occurring pedestrian behaviors (Fig. 7). These rollouts provide qualitative context beyond the controlled evaluation; they are not additional trials in Tables 1 and 2.

Figure 7: GRACE rollouts in non-scripted human environments with naturally occurring pedestrian behavior.

G Perceived Social Compliance: Study Details

We collect ratings from two cohorts: 10 naive raters without knowledge of the study aims or design, and 10 navigation-experiment volunteers blinded to method identity. The naive raters assess anonymized navigation videos presented in randomized order; the volunteers provide a complementary perspective from direct interaction with the robot. Both cohorts rate Safety, Comfort, Naturalness, Social Appropriateness, and Responsiveness on five-point scales. Figure 8 compares NoMaD, FlowNav, SACSoN, GSON, and GRACE. GRACE has the highest rating profile across all five dimensions in both cohorts. Volunteers show a consistent preference, with lower absolute Safety and Comfort ratings than the naive cohort. The small cohorts and descriptive summaries limit conclusions about statistical significance and broader population preferences.

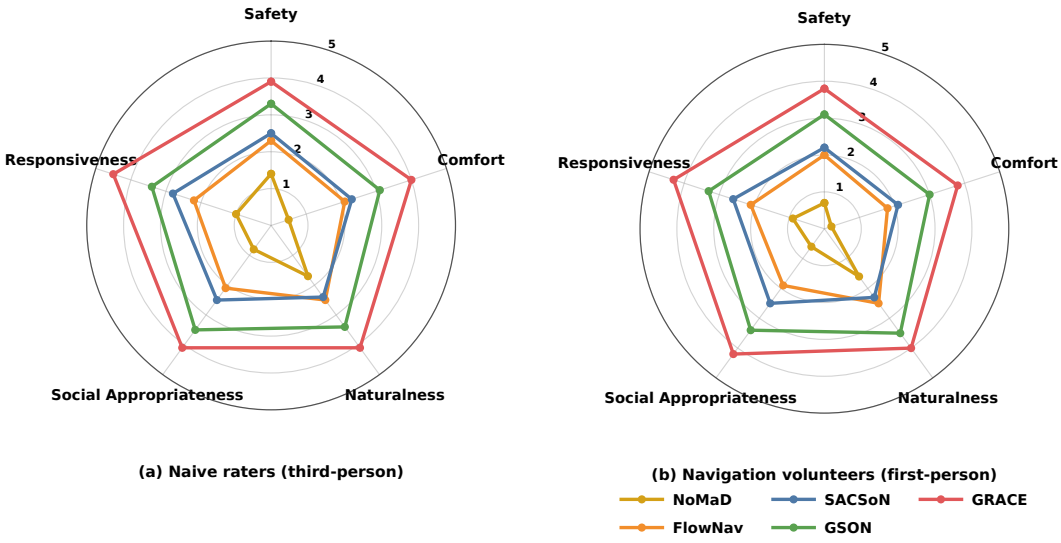


Figure 8: Perceived social compliance on five-point scales: (a) 10 naive video raters and (b) 10 navigation-experiment volunteers blinded to method identity. Larger values indicate more favorable ratings. The figure reproduces the supplied study summary.

Appendix References

[1] N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese. GONet: A semi-supervised deep learning approach for traversability estimation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018.

[2] D. Shah, B. Eysenbach, N. Rhinehart, and S. Levine. Rapid exploration for open-world navigation with latent goal models. In *Proceedings of the Conference on Robot Learning*, 2021.

[3] S. Triest, M. Sivaprakasam, S. J. Wang, W. Wang, A. M. Johnson, and S. Scherer. TartanDrive: A large-scale dataset for learning off-road dynamics models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2022.

[4] N. Hirose, D. Shah, A. Sridhar, and S. Levine. SACSoN: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 2023.

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

[6] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023.

[7] J. Bae, J. Kim, J. Yun, C. Kang, J. Choi, C. Kim, J. Lee, J. Choi, and J.-W. Choi. Sit dataset: Socially interactive pedestrian trajectory dataset for social navigation robots. In *Advances in Neural Information Processing Systems*, 2023.

[8] MMPose Contributors. MMPose: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[9] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*. CUP Archive, 1990.