SOCIOWM: SOCIAL-CONDITIONED WORLD MODEL FOR SOCIAL CONTEXT VISION-AND-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional vision-and-language navigation (VLN) enables embodied agents to navigate static scenes by following natural-language instructions. Existing Social-VLN works further introduce moving pedestrians into navigation scenarios, but primarily model them as dynamic obstacles or generic visual entities, overlooking the social context that influences pedestrian behavior, including personal attributes and human–robot interactions. Consequently, existing methods may fail to account for attribute-dependent differences among pedestrians and their interaction responses to robot actions, leading to navigation behavior that prioritizes goal progress while neglecting social norms. To address this limitation, we formulate Social Context Vision-and-Language Navigation (Social Context-VLN), a new VLN setting that incorporates pedestrian motion and social context into navigation decision-making. To support this setting, we construct SocialBench-VLN, the first benchmark for Social Context-VLN. It explicitly annotates Pedestrian Attribute, including age and gender, and Human–Robot Interaction as social context, and contains more than 35K language-guided navigation episodes involving 1.4K pedestrians. To validate the effectiveness of these structured annotations, we further introduce SocioWM, a socially conditioned world model that predicts the social consequences of candidate routes and selectively modulates candidate world states for action selection. Experiments on SocialBench-VLN show that SocioWM improves the social success rate (SR) from 18.27% to 27.07%, while reducing collisions and improving personal-space compliance. Evaluation of existing VLN methods further shows that strong conventional navigation performance does not necessarily imply socially compliant behavior. Real-world experiments verify our method transfers to unseen indoor environments and real-world pedestrian interactions.Ours project page:https://anonymous.4open.science/r/iclr2027anony

1 INTRODUCTION

Conventional vision-and-language navigation (VLN) aims to enable embodied agents to navigate static scenes by following natural-language instructions. Recent advances in vision-language models and generative action planners have substantially improved instruction understanding, long-horizon reasoning, and generalization to unseen scenes. Social-VLN extends VLN by introducing moving pedestrians, but primarily represents them through observable motion or geometric states without explicitly modeling social context. As summarized in Figure 1, this setting provides an important step toward navigation in human-populated environments, yet motion alone does not fully explain how pedestrians behave during robot navigation.

In real-world social navigation, effective decision-making also depends on social context, including pedestrian attributes and Human–Robot Interactions. Pedestrians with different attributes may exhibit different behavioral characteristics, while their behavior may also change in response to robot actions. Without such context, an agent may prioritize progress toward the instructed goal and select a seemingly efficient route, but fail to account for pedestrian-specific behavioral differences and responses to the robot. This can result in human–robot collisions and ultimately prevent the agent from completing the navigation task.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

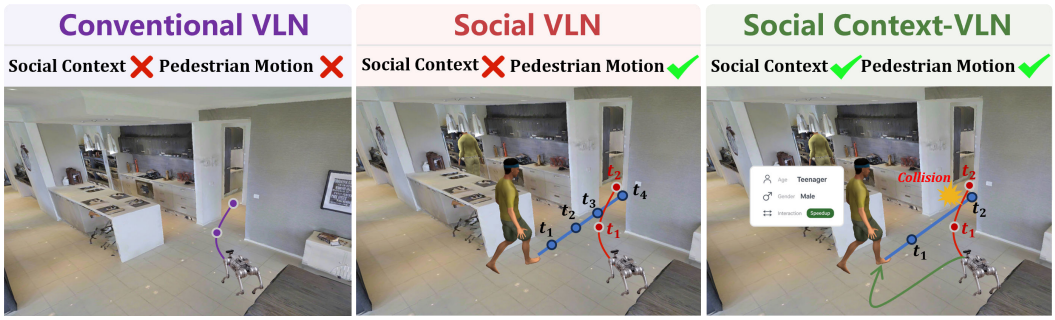

Figure 1: **Motivation of Social Context-VLN.** Conventional VLN does not model pedestrian motion or social context. Social-VLN incorporates pedestrian motion but lacks explicit social context. Social Context-VLN jointly considers pedestrian motion and social context, including age, gender, and human–robot interaction, enabling the agent to anticipate collision risks and select a socially compliant detour.

These limitations motivate navigation models that go beyond motion-based pedestrian modeling and explicitly account for the contextual factors affecting pedestrian behavior and robot–pedestrian interaction. This introduces two key challenges. First, existing VLN benchmarks provide limited structured supervision for learning individual and relational social context. Second, heterogeneous social cues must be transformed into predictions that are useful for evaluating and selecting alternative navigation actions.

To address these challenges, we formulate Social Context Vision-and-Language Navigation (Social Context-VLN), a new VLN setting that extends Social-VLN by explicitly incorporating individual and relational social context into navigation decision-making. The setting requires an agent to follow a natural-language instruction while considering pedestrian motion together with the contextual factors that may affect pedestrian behavior and responses to robot actions. The objective is therefore not only to make progress toward the navigation goal, but also to select actions that support safe and socially compliant interaction with nearby pedestrians.

To support this setting, we introduce SocialBench-VLN, the first Social Context-VLN benchmark. It provides structured annotations of Pedestrian Attribute (age and gender) and Human–Robot Interaction, together with Future Motion and Semantic Destination for trajectory construction, covering over 35K episodes and 1.4K pedestrians. These annotations support privileged training and evaluation beyond collision avoidance.

To validate the effectiveness of the proposed context annotations, we introduce SocioWM, a socially conditioned world model that predicts the social consequences of candidate routes. Given a language instruction, observation history, graph structure, and candidate routes, SocioWM uses Route-Conditioned Social Reasoning to infer a candidate-specific Social Future Latent. Social Context Modulation then selectively injects the predicted social information into reconstructed candidate world states while preserving the information required for goal-directed navigation. The resulting socially enhanced candidate states are used to score alternative actions.

Experiments on SocialBench-VLN show that SocioWM improves the social success rate from 19% to 27%, while reducing collisions and improving personal-space compliance. Evaluations of existing VLN baselines further show that reasonable conventional navigation performance does not necessarily imply socially compliant behavior. Real-robot experiments verify that policies trained on SocialBench-VLN transfer to unseen indoor environments and real-world pedestrian interactions.

Our contributions are summarized as follows:

- We formulate Social Context-VLN, a new VLN setting where agents incorporate pedestrian motion, individual attributes, and human–robot interactions into decision-making for socially compliant navigation.
- We introduce SocialBench-VLN, the first Social Context-VLN benchmark, with structured annotations of pedestrian attributes (age, gender) and human–robot interaction, spanning over 35K episodes and 1.4K pedestrians.

- We introduce SocioWM, a socially conditioned world model that learns candidate-specific Social Future Latents and modulates candidate world states according to their predicted social implications.

2 Related Work

Vision-and-language navigation. Vision-and-language navigation (VLN) studies how embodied agents follow natural-language instructions in visual environments. Starting from graph-based navigation in real indoor scenes Anderson et al. (2018), subsequent benchmarks extend VLN to longer trajectories, continuous control, multilingual instructions, and object-oriented navigation Jain et al. (2019); Krantz et al. (2020); Qi et al. (2020); Ku et al. (2020); Zhu et al. (2021). A long line of methods improves instruction grounding and trajectory generation through data augmentation, pretraining, transformer-based memory, and map-based reasoning Fried et al. (2018); Tan et al. (2019); Hao et al. (2020); Chen et al. (2021; 2022); An et al. (2022), and recent vision-language models further enable video-based and legged-robot agents Zhang et al. (2024b;a); Cheng et al. (2024); Wei et al. (2025; 2026). However, these agents are optimized for instruction grounding and goal completion in essentially static environments; when pedestrians appear, they are treated as obstacles or background rather than socially situated participants whose behavior shapes navigation.

Social-VLN and human-aware navigation. Human-aware and social navigation studies collision avoidance and personal-space compliance through social-force models, reciprocal collision avoidance, and human-aware planning Helbing & Molnar (1995); Van Den Berg et al. (2011); Kruse et al. (2013); Chen et al. (2019); Biswas et al. (2022), while trajectory-prediction methods forecast pedestrian motion via social and graph-based interaction modeling Alahi et al. (2016); Gupta et al. (2018); Mangalam et al. (2020); Salzmann et al. (2020); Yuan et al. (2021); these settings largely abstract away language-conditioned goals. Closer to our setting, HA-VLN Dong et al. (2025) introduces dynamic multi-human interactions into language-guided navigation, whereas crowd- and social-navigation benchmarks run without language instructions Vuong et al. (2024); Cancelli et al. (2023); Li et al. (2021). In all cases, however, pedestrians are treated as dynamic obstacles without structured annotations of attributes or robot-induced interaction responses. SocialBench-VLN is the first benchmark that jointly provides language guidance, dynamic pedestrians, pedestrian attributes, and human–robot interaction annotations.

World models for embodied navigation. Methodologically, our framework builds on world models that learn latent environment dynamics and evaluate actions by imagining their consequences Ha & Schmidhuber (2018); Hafner et al. (2019; 2023). In language-guided navigation, candidate-conditioned world models reconstruct candidate world states to score alternative actions Yao et al. (2026), yet they predict visual rather than social consequences. SocioWM addresses this gap by predicting a candidate-specific Social Future Latent and modulating candidate states accordingly, requiring only RGB-D observations and language instructions at inference.

3 Methodology

3.1 Social Context-VLN

We formulate Social Context-VLN as a vision-and-language navigation setting that extends Social-VLN by treating social context, beyond pedestrian motion, as a key factor in navigation decision-making. An episode is defined as

$$\mathcal{E} = (\mathcal{S}, \mathcal{I}, g, \mathcal{P}), \tag{1}$$

where $\mathcal{S}$ denotes the environment, $\mathcal{I}$ the natural-language instruction, g the navigation goal, and $\mathcal{P}$ the set of pedestrians in the episode.

For each pedestrian $p \in \mathcal{P}$, we use $\mathbf{m}^p_{1:t}$ to denote the observed motion history and define the explicit social context as

$$\mathbf{c}^p_t = (\mathbf{u}^p, \boldsymbol{\xi}^p_t)\,. \tag{2}$$

Here, $\mathbf{u}^p = [a^p, g^p]$ denotes Pedestrian Attribute, where a^p and g^p represent the pedestrian's age group and gender, respectively. The variable $\boldsymbol{\xi}^p_t$ denotes Human–Robot Interaction, including pedestrian responses such as passing, slowing down, stopping, and yielding during interaction with the

robot. Together, Pedestrian Attribute and Human–Robot Interaction constitute the explicit social context considered in Social Context-VLN.

At each decision step t, the agent receives the instruction $\mathcal{I}$ and observation history $o_{1:t}$ and selects an action a_t. The objective is to reach the instructed goal without collision while respecting pedestrian personal space. Unlike Social-VLN, which focuses on pedestrian motion, Social Context-VLN additionally uses Pedestrian Attribute and Human–Robot Interaction in navigation decision-making. These annotations provide privileged supervision during training and evaluation but are unavailable to the deployed agent.

3.2 SocialBench-VLN Construction

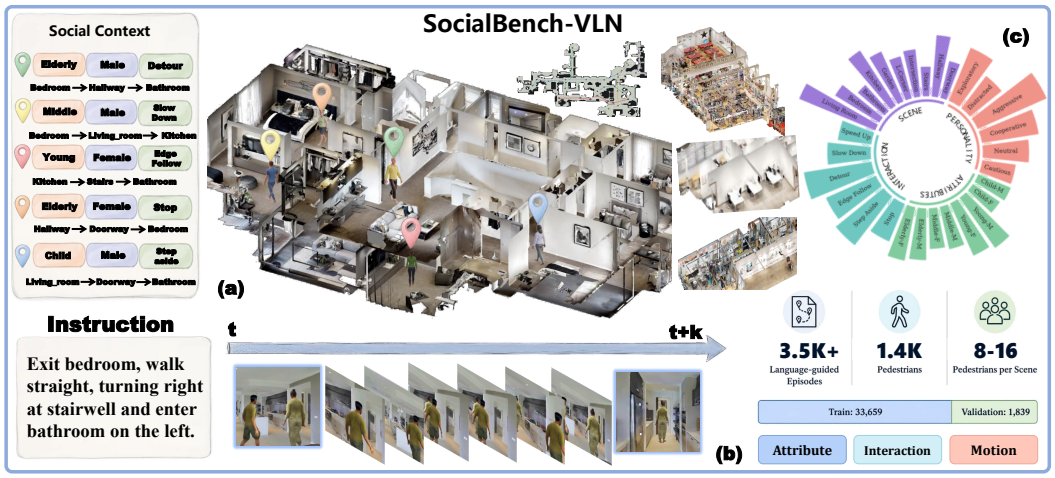

Figure 2: **Overview of SocialBench-VLN.** (a) Habitat simulation scenes with dynamic pedestrians and their associated social-context annotations, including Pedestrian Attribute and Human–Robot Interaction. (b) A language instruction and the temporally aligned RGB observations collected during navigation. (c) Dataset statistics, including annotation distributions shown by the circular charts and key statistics summarized by the cards.

To support and evaluate Social Context-VLN, as shown in Figure 2, we construct SocialBench-VLN on the Habitat simulation platform using existing indoor scenes and language-guided navigation instructions. For each episode, 8–16 pedestrians are instantiated in navigable regions and move concurrently with the navigation agent. In total, the benchmark contains over 35K language-guided episodes involving 1.4K distinct pedestrians. Pedestrian states and human–robot interactions are recorded at each simulation step and temporally aligned with the agent's RGB-D observations.

Table 1 compares representative benchmarks along four dimensions: language guidance, dynamic pedestrians, Pedestrian Attribute, and Human–Robot Interaction. Existing benchmarks support language-guided navigation or navigation among dynamic pedestrians, but lack explicit annotations of Pedestrian Attribute and Human–Robot Interaction. SocialBench-VLN is the only benchmark that covers all four dimensions, supporting context-aware VLN among dynamic pedestrians.

i) Pedestrian Attribute Assignment. Each pedestrian is assigned an age group and gender, which remain fixed throughout the episode and determine visual appearance, linear and angular velocities, and reaction distance.

ii) Human–Robot Interaction Simulation. Pedestrian behavior is updated according to the relative robot– pedestrian state and pedestrian-specific parameters. When paths intersect, pedestrians may pass, slow down, stop, or yield, and the resulting response is recorded as Human–Robot Interaction.

iii) Destination-Guided Trajectory Construction. Each pedestrian is assigned a Semantic Destination, such as a doorway, living room, kitchen, or bedroom. A route is planned to this destination, and its time-varying states are recorded as Future Motion.

Table 1: Comparison with representative VLN and social-navigation benchmarks. *Attr.*: structured Pedestrian Attribute annotations. *Inter.*: structured Human–Robot Interaction annotations.

Benchmark	Lang.	Dyn. Ped.	Attr.	Inter.
R2R-CE / RxR-CE Krantz et al. (2020); Ku et al. (2020)	✓	–	–	–
HA-VLN Dong et al. (2025)	✓	✓	–	–
AdaVLN Loh et al. (2024)	✓	✓	–	–
DPed-VLN Dai et al. (2026)	✓	✓	–	–
iGibson-SN Li et al. (2021)	–	✓	–	–
HabiCrowd Vuong et al. (2024)	–	✓	–	–
HM3D-S Cancelli et al. (2023)	–	✓	–	–
Social-HM3D / Social-MP3D Gong et al. (2025)	–	✓	–	–
SocNavBench Biswas et al. (2022)	–	✓	–	–
CrowdNav Chen et al. (2019)	–	✓	–	–
SocialBench-VLN	✓	✓	✓	✓

3.3 SocioWM: A Socially Conditioned World Model

Overview As shown in Figure 3, SocioWM is a candidate-conditioned world-model framework for Social Context-VLN. Given a language instruction and RGB-D observation history, the navigation branch first produces a base navigation state and generates candidate actions together with their corresponding routes. Route-Conditioned Social Reasoning (RCSR) organizes pedestrian information into four token groups. Age, Gender, and Human–Robot Interaction encode explicit social context, while Future Motion captures pedestrian movement. These tokens are fused into a social memory and interact bidirectionally with each candidate route to predict a candidate-specific Social Future Latent and its corresponding Social Correction. Candidate World-State Reconstruction and Social Context Modulation then inject this correction into each candidate state through a learned gate, after which the navigation policy selects the next waypoint or STOP action.

Route-Conditioned Social Reasoning RCSR first maps the RGB-D observation history $O_{1:t}$ into four semantically organized token groups:

$$\left(\{T_t^a, T_t^g, T_t^i\},\ \{T_t^m\}\right) = E_{\mathrm{s}}(O_{1:t}), \tag{3}$$

Here, the superscripts a, g, i, and m denote Age, Gender, Human–Robot Interaction, and Future Motion, respectively. Thus, T_t^a, T_t^g, and T_t^i encode explicit social context, whereas T_t^m represents pedestrian movement. During training, these representations are learned from structured pedestrian annotations and candidate-level social supervision. Future Motion is supervised against simulator-recorded future motion states through $\mathcal{L}_{\mathrm{m}}$.

A fusion layer combines the four token groups into a unified social memory:

$$M_t^s = F_{\mathrm{s}}\left(T_t^a, T_t^g, T_t^i, T_t^m\right), \tag{4}$$

where the superscript s denotes the social memory. The resulting memory jointly represents pedestrian-specific context, human–robot interaction, and motion information required for candidate-route reasoning.

For each candidate route R_t^n generated by the Candidate Planner, RCSR performs bidirectional interaction between the social memory and the route representation:

$$\begin{aligned} \widetilde{M}_t^n &= M_t^s + \mathrm{Attn}\left(Q = M_t^s, K = R_t^n, V = R_t^n\right), \\ \widetilde{R}_t^n &= R_t^n + \mathrm{Attn}\left(Q = R_t^n, K = M_t^s, V = M_t^s\right). \end{aligned} \tag{5}$$

Here, $\widetilde{M}_t^n$ denotes the social-memory tokens updated by candidate-route information, while $\widetilde{R}_t^n$ denotes the route tokens refined by pedestrian information.

Their joint representation is transformed into a candidate-specific Social Future Latent:

$$z_t^{f,n} = f_{\mathrm{tr}}\left([\widetilde{M}_t^n; \widetilde{R}_t^n]\right), \tag{6}$$

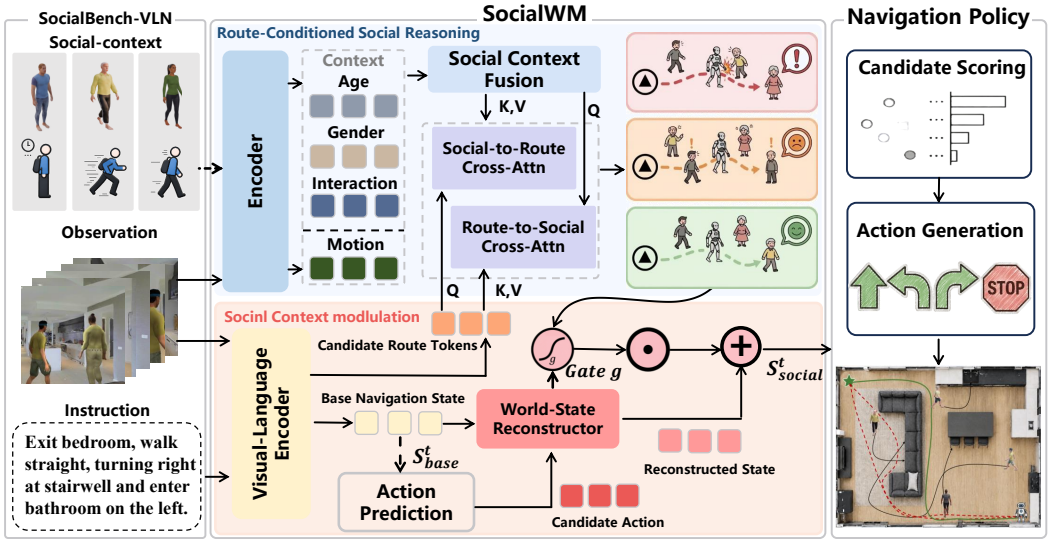

Figure 3: Overall framework of SocioWM. RCSR organizes pedestrian social context and motion information into social memory and interacts with each candidate route to predict a candidate-specific Social Future Latent. The resulting Social Correction is selectively injected into the reconstructed candidate world state through Social Context Modulation for action selection.

where the superscript f denotes the future social consequence predicted for candidate route n. A Social Projector maps this latent to a candidate-specific Social Correction:

$$\Delta_t^{s,n} = P_{\mathrm{s}}\left(z_t^{f,n}\right). \tag{7}$$

The resulting Social Correction is passed to Social Context Modulation for candidate-state adjustment.

Social Context Modulation The navigation branch processes the language instruction $\mathcal{I}$ and RGB-D observation history $O_{1:t}$ through a Visual-Language Graph Encoder to obtain the base navigation state:

$$S_t = E_{\mathrm{vlg}}\left(\mathcal{I}, O_{1:t}\right). \tag{8}$$

The Candidate Planner then generates candidate actions and their associated routes from S_t:

$$\{(a_t^n, R_t^n)\}_{n=1}^{N} = P_{\mathrm{c}}(S_t), \tag{9}$$

where P_{c} denotes the Candidate Planner.

For each candidate, two computations proceed in parallel. RCSR processes the candidate route R_t^n to produce the Social Correction $\Delta_t^{s,n}$. Meanwhile, Candidate World-State Reconstruction combines the base navigation state S_t with the candidate action representation $e(a_t^n)$:

$$\widehat{H}_t^n = f_{\mathrm{r}}\left(S_t, e(a_t^n)\right), \tag{10}$$

where $\widehat{H}_t^n$ denotes the reconstructed candidate world state.

The reconstructed candidate state and its corresponding Social Correction are combined through gated residual modulation:

$$g_t^n = \sigma\left(f_g\left([\widehat{H}_t^n; \Delta_t^{s,n}]\right)\right), \quad \widetilde{H}_t^n = \widehat{H}_t^n + g_t^n \odot \Delta_t^{s,n}, \tag{11}$$

where $\widetilde{H}_t^n$ is the socially modulated candidate state. The learned gate selectively incorporates the predicted social consequence of each candidate route while preserving the information required for goal-directed navigation.

Finally, the modulated world state is fed back into the navigation state and scored by the socially conditioned navigation policy:

$$S_t^{s,n} = F_{\text{fb}}\left(S_t, \widetilde{H}_t^n\right), \qquad \ell_t^n = \pi_{\text{s}}\left(S_t^{s,n}\right), \tag{12}$$

where $S_t^{s,n}$ denotes the socially conditioned state for candidate n, and π_{s} denotes the socially conditioned navigation policy. Invalid candidates are masked, and the highest-scoring valid candidate is selected as the next waypoint or STOP action.

4 Experiments

4.1 Implementation Details

Simulation environments and metrics. All simulation experiments are conducted in Habitat on SocialBench-VLN. The benchmark contains more than 35K language-guided episodes involving 1.4K distinct pedestrians, with 8–16 pedestrians per scene. We use scene-disjoint training and validation splits with 33,659 and 1,839 episodes. All methods use the same episode definitions, pedestrian simulation, candidate-action space, observation preprocessing, and evaluation protocol.

Social Context-VLN evaluates both navigation and human–robot interaction. We therefore use SR, PSC, Coll., and ISR as the primary metrics. Let $\mathcal{E}$ denote the evaluation episodes and $\mathcal{E}_{\text{inter}} \subseteq \mathcal{E}$ the episodes with triggered human–robot interaction. For an episode $e \in \mathcal{E}$, let $\mathbb{I}_{\text{goal}}(e)$ indicate whether the robot stops within $3\,\text{m}$ of the goal, and let $\mathbb{I}_{\text{nocoll}}(e)$ indicate collision-free navigation.

For interaction-triggered episodes, ISR uses the same collision-free goal criterion as SR:

$$\text{SR} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{I}_{\text{goal}}(e)\mathbb{I}_{\text{nocoll}}(e), \qquad \text{ISR} = \frac{1}{|\mathcal{E}_{\text{inter}}|}\sum_{e\in\mathcal{E}_{\text{inter}}} \mathbb{I}_{\text{goal}}(e)\mathbb{I}_{\text{nocoll}}(e). \tag{13}$$

PSC measures the fraction of navigation steps during which the robot remains at least $1.0\,\text{m}$ from every same-floor pedestrian, whereas Coll. is the fraction of episodes containing at least one human–robot distance below $0.6\,\text{m}$. We additionally report SPL and NE as conventional navigation metrics. Higher SR, PSC, ISR, and SPL are better, while lower Coll. and NE are better.

Training details. We initialize SocioWM from the SC^2-WM checkpoint trained for 60K iterations. Training proceeds in two stages. Stage 1 optimizes RCSR to learn pedestrian representations and candidate-level social risks using

$$\begin{aligned} \mathcal{L}_{\text{S1}} &= \mathcal{L}_{\text{col}} + \mathcal{L}_{\text{m}} + \lambda\left(\mathcal{L}_{\text{int}} + \mathcal{L}_{\text{attr}}\right), \\ \mathcal{L}_{\text{attr}} &= \text{Mean}\left(\mathcal{L}_{\text{A}}, \mathcal{L}_{\text{G}}\right). \end{aligned} \tag{14}$$

where $\mathcal{L}_{\text{col}}$ and $\mathcal{L}_{\text{int}}$ supervise candidate-level collision and human–robot interaction prediction, respectively. The motion loss $\mathcal{L}_{\text{m}}$ directly supervises future pedestrian motion prediction, while $\mathcal{L}_{\text{A}}$ and $\mathcal{L}_{\text{G}}$ denote the KL-divergence losses for age and gender distributions.

In Stage 2, the complete training objective is

$$\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{nav}} + \mathcal{L}_{\text{col}} + \mathcal{L}_{\text{m}} + \lambda\left(\mathcal{L}_{\text{int}} + \mathcal{L}_{\text{attr}}\right). \tag{15}$$

where $\mathcal{L}_{\text{nav}}$ denotes the overall navigation objective, including the candidate-action and visual-prediction objectives. We set $\lambda = 1.0$. Stage 2 retains the social supervision from Stage 1 while optimizing the model for navigation decision-making.

Stage 1 trains RCSR and its supervision heads for 40K global rollouts with a learning rate of 1×10^{-5}. Stage 2 additionally optimizes the Social Projector, fusion gate, candidate world-model components, visual predictor, feedback modules, and enhanced SAP head for 60K global rollouts. The learning rates are 1×10^{-5} for RCSR and its supervision heads and 1×10^{-6} for the candidate world-model components. The original navigation backbone and global navigation head remain frozen throughout training.

Both stages use AdamW and safe candidate-route replay. Training is distributed over six NVIDIA RTX 4090 GPUs, with two parallel Habitat environments per GPU, yielding a global batch size of 12. The final model is evaluated on all 1,839 validation episodes.

4.2 Main Results

Table 2 compares representative VLN agents on the validation split of SocialBench-VLN. Compared with SC^2-WM, SocioWM improves SR, PSC, and ISR by 8.80, 1.75, and 5.60 percentage points and reduces Coll. by 5.06 points. It also improves SPL from 13.66% to 20.78% and reduces NE from 6.60 to 6.35.

Although its SPL and NE remain below several conventional VLN methods due to the navigation capability inherited from SC^2-WM, SocioWM achieves the best SR, ISR, PSC, and Coll., demonstrating the effectiveness of candidate-conditioned social reasoning.

Table 2: Main results on SocialBench-VLN. Params denotes the number of model parameters. SR, SPL, and NE are conventional navigation metrics, while ISR, PSC, and Coll. are the primary social metrics.

Method	Params	Navigation Metrics			Social Metrics		
		SR↑	SPL↑	NE↓	ISR↑	PSC↑	Coll.↓
NaVid Zhang et al. (2024b)	7B	16.86	12.15	7.98	5.14	87.73	55.21
UniNaVid Zhang et al. (2024a)	7B	17.56	12.86	7.18	5.52	85.56	54.84
NaVILA Cheng et al. (2024)	8B	20.34	18.92	7.62	6.05	85.77	55.36
StreamVLN Wei et al. (2025)	7B	21.15	18.58	5.08	6.27	86.25	56.87
AwareVLN Gong et al. (2025)	7B	22.95	19.64	5.45	6.44	86.99	59.05
InternNav Wei et al. (2026)	7B	25.82	**22.28**	**4.61**	8.72	87.02	53.94
SC^2-WM Yao et al. (2026)	0.5B	18.27	13.66	6.60	5.60	88.60	52.31
SocioWM (Ours)	0.5B	**27.07**	20.78	6.35	**11.20**	**90.35**	**47.25**

4.3 Ablation Studies

We conduct module, annotation, and loss-weight ablations on the validation split with identical initialization, training protocol, and metrics. We report SR, PSC, Coll., and ISR for all studies.

Table 3: Module ablation.

Variant	SR↑	PSC↑	Coll.↓	ISR↑
w/o RCSR	18.27	88.60	52.31	5.60
w/o Gated Mod.	20.74	91.28	45.62	7.36
Unidir. Route–Social	24.25	88.78	49.54	10.12
Full SocioWM	27.07	90.35	47.25	11.20

Table 4: Annotation ablation.

Supervision	SR↑	PSC↑	Coll.↓	ISR↑
w/o Age	26.10	90.21	48.56	10.82
w/o Gender	26.31	90.14	48.34	10.98
w/o Attribute	25.94	90.07	48.71	10.55
w/o Interaction	24.61	89.36	49.83	7.86
w/o Motion	23.96	89.18	50.21	8.44
w/o All Structured Sup.	19.34	88.74	51.91	5.76
Full Supervision	27.07	90.35	47.25	11.20

Module ablation. As shown in Table 3, removing RCSR recovers the SC^2-WM baseline, confirming the contribution of route-conditioned social reasoning. Removing the gate raises PSC and lowers Coll., but reduces SR and ISR to 20.74% and 7.36%, indicating overly conservative avoidance at the expense of goal completion. The unidirectional variant is inferior on all four metrics, demonstrating the benefit of bidirectional route–social interaction.

Annotation ablation. Table 4 shows that the four supervision types play distinct and complementary roles. Removing Future Motion supervision by dropping $\mathcal{L}_{\mathrm{m}}$ while retaining the collision objective causes the largest SR drop and Coll. increase among the single-supervision ablations, highlighting the importance of direct motion supervision for anticipating route-level conflicts. Removing Human–Robot Interaction causes the largest ISR drop, demonstrating its importance in

interaction-triggered episodes. Age and Gender provide complementary pedestrian-specific context, as removing both produces a larger degradation than removing either one alone. When all structured supervision losses, including collision, motion, interaction, and attribute supervision, are removed, SR and ISR decrease to 19.34% and 5.76%. This variant retains the complete architecture but is trained only through the navigation objective, explaining its small advantage over *w/o RCSR*; the substantial gap to full supervision confirms that most gains arise from the structured annotations.

Loss-weight sensitivity. Table 5 reports results for different loss weights. Increasing λ from 0 to 0.5 substantially improves all four metrics. At $\lambda = 1.0$, SR, Coll., and ISR improve further, while PSC remains nearly unchanged. The best SR and ISR are achieved at $\lambda = 1.0$. Increasing λ to 1.5 further improves PSC and reduces Coll., but slightly lowers SR and ISR.

Table 5: Loss-weight sensitivity.

Loss weight	SR↑	PSC↑	Coll.↓	ISR↑
$\lambda = 0.0$	23.12	89.05	50.62	7.18
$\lambda = 0.5$	26.64	90.39	48.23	10.52
$\lambda = 1.0$	27.07	90.35	47.25	11.20
$\lambda = 1.5$	26.48	90.83	46.81	10.71

Figure 4 presents qualitative results in simulation and the real world. In simulation, SocioWM anticipates the pedestrian conflict and reaches the goal through a safe detour. In contrast, SC^2-WM prioritizes progress toward the goal but collides by ignoring the potential interaction. The real-world sequence further demonstrates navigation around a pedestrian in an unseen indoor environment.



Figure 4: **Qualitative analysis.** Top: SocioWM safely detours around the pedestrian, whereas SC^2-WM prioritizes direct goal progress and collides. Bottom: real-world navigation around a pedestrian in an unseen indoor environment.

5 Conclusion

We introduced Social Context-VLN and SocialBench-VLN, a benchmark with structured annotations of Pedestrian Attribute, Human–Robot Interaction, and Future Motion. We further proposed SocioWM, which predicts the social consequences of candidate routes and uses them to condition candidate-state transitions. On SocialBench-VLN, SocioWM improves SR from 18.27% to 27.07%, while also improving PSC and ISR, reducing Coll., and demonstrating consistent social benefits. Real-world experiments verify its transfer to unseen indoor environments with pedestrians. Despite these results, the benchmark currently relies on simulator-based pedestrian behavior and a limited set of structured context categories. Future work will extend the benchmark to more diverse human behaviors and longer-horizon interactions in real-world environments.

Ethical Statement

This work aims to improve the safety and social compliance of embodied agents navigating among pedestrians. Potential beneficial applications include assistive navigation for people with visual impairments, domestic service robots, and autonomous systems operating in crowded indoor environments. By incorporating pedestrian attributes, Human–Robot Interaction, and Future Motion into navigation decision making, our work seeks to reduce human–robot collisions, personal-space violations, and unnecessary disruption to pedestrians.

We also recognize several ethical risks. Socially aware navigation systems may be misused for unauthorized surveillance, behavioral profiling, crowd monitoring, or military and security applications. Moreover, attribute-conditioned modeling may reinforce biases or lead to inappropriate assumptions about pedestrians if used outside the controlled setting considered in this work. The Age and Gender annotations in SocialBench-VLN are introduced solely as structured training and evaluation signals for studying socially compliant navigation; they should not be used to identify, rank, or make decisions about individuals in real-world deployments. Real-world use should respect privacy, obtain appropriate consent, maintain human oversight, and include reliable safety mechanisms.

The primary motivation of this research is to support safer and more socially responsible embodied AI. We discourage the use of SocioWM or SocialBench-VLN for malicious, discriminatory, intrusive, or otherwise unethical purposes, and encourage responsible evaluation and deployment in accordance with applicable laws, institutional policies, and community guidelines.

Reproducibility

To facilitate reproducibility, we provide detailed descriptions of the Social Context-VLN task, SocialBench-VLN construction procedure, SocioWM architecture, training objectives, optimization schedule, and evaluation protocol in the main paper and appendices. We also report the data split, pedestrian simulation procedure, candidate route generation process, loss definitions, and experimental settings.

To support independent verification and future research, we will release the source code, SocialBench-VLN dataset, and pedestrian models, subject to the licenses and privacy requirements of the underlying software and environments. These resources will enable other researchers to reproduce the benchmark and build upon Social Context-VLN and SocioWM.

References

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385*, 2022.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 3674–3683. IEEE, 2018.

Abhijat Biswas, Allan Wang, Gustavo Silvera, Aaron Steinfeld, and Henny Admoni. Socnavbench: A grounded simulation testing framework for evaluating social navigation. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3):1–24, 2022.

Enrico Cancelli, Tommaso Campari, Luciano Serafini, Angel X Chang, and Lamberto Ballan. Exploiting proximity-aware tasks for embodied social navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10923–10933. IEEE, 2023.

Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *2019 international conference on robotics and automation (ICRA)*, pp. 6015–6022. IEEE, 2019.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 2021.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16516–16526. IEEE, 2022.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

Haojie Dai, Xiangyi Wang, Liuyi Wang, Kai Sheng, Zongtao He, Chengju Liu, Wei Ye, and Qijun Chen. Dped-vln: A benchmark for socially compliant vision-and-language navigation in dynamic pedestrian environments. *arXiv preprint arXiv:2609.21504*, 2026.

Yifei Dong, Fengyi Wu, Qi He, Lingdong Kong, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, et al. Ha-vln 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions. *arXiv preprint arXiv:2503.14229*, 2025.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems*, 31, 2018.

Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9122–9129. IEEE, 2025.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264. IEEE, 2018.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13134–13143. IEEE, 2020.

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

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1862–1872, 2019.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

Dillon Loh, Tomasz Bednarz, Xinxing Xia, and Frank Guan. Adavln: Towards visual language navigation in continuous indoor environments with moving humans. *arXiv preprint arXiv:2411.18539*, 2024.

Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision*, pp. 759–776. Springer, 2020.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9982–9991, 2020.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European conference on computer vision*, pp. 683–700. Springer, 2020.

Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621, 2019.

Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics research: the 14th international symposium ISRR*, pp. 3–19. Springer, 2011.

An Vuong, Toan Nguyen, Minh Nhat Vu, Baoru Huang, Huynh Thi Thanh Binh, Thieu Vo, and Anh Nguyen. Habicrowd: A high performance simulator for crowd-aware visual navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5821–5827. IEEE, 2024.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

Meng Wei, Chenyang Wan, Peng Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 12380–12396, 2026.

Xuan Yao, Yuze Zhu, Junyu Gao, Zongmeng Wang, and Changsheng Xu. Sc $\hat{}\{2\}$-wm: A self-correcting world model with closed-loop feedback for vision-and-language navigation in continuous environments. *arXiv preprint arXiv:2608.07548*, 2026.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9813–9823, 2021.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024a.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024b.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12689–12699, 2021.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this paper, large language models were used strictly to improve the clarity and quality of the manuscript. After completing the manuscript, we used GPT-5.6-Sol to refine the language and identify grammatical or stylistic issues. The model was instructed with the following prompt:

> *Please proofread the following academic manuscript for grammar, clarity, and concision. Preserve its technical meaning, terminology, equations, numerical results, and citations. Do not introduce new claims or alter the methodology or conclusions.*

The use of the language model was limited to proofreading and language refinement. It was not used to generate the research ideas, methodology, experimental results, data, figures, or conclusions.

B SOCIALBENCH-VLN CONSTRUCTION

This appendix provides additional details on the construction of SocialBench-VLN. The generation process follows a structured causal chain. Each pedestrian is first assigned a Semantic Destination, based on which a complete trajectory is annotated as a sequence of waypoints. The pedestrian's Age and Gender determine an attribute-specific visual appearance and modulate the sampling distribution of a latent personality type. The sampled personality further affects the pedestrian's interaction-behavior tendency. When the robot approaches the pedestrian, the corresponding predefined behavior is triggered and executed in the simulator. The resulting states are recorded as Future Motion and Human–Robot Interaction (HRI).

SocialBench-VLN is constructed in the Habitat simulator by extending original VLN scenes and language-guided navigation episodes with multiple goal-directed pedestrians. The benchmark contains more than 35K navigation episodes and 1.4K pedestrian agents, with 8–16 pedestrians in each scene. The training and validation sets are split at the scene level to avoid scene overlap.

B.1 PEDESTRIAN ATTRIBUTES, APPEARANCE, AND BEHAVIORAL TENDENCIES

Each pedestrian is assigned a fixed age group and gender. For each Age–Gender combination, we construct an attribute-specific pedestrian appearance, including the corresponding body shape, visual style, and avatar configuration. Thus, pedestrians with different attributes are rendered with distinguishable visual appearances rather than being represented by a shared generic human model.

Age and Gender also affect the pedestrian's motion characteristics. For each pedestrian, we sample attribute-conditioned weights for linear velocity, angular velocity, and reaction distance. These weights scale the corresponding base motion parameters, allowing pedestrians with different ages and genders to exhibit different walking speeds, turning behaviors, and reaction distances even under similar spatial conditions.

In addition, Age and Gender modulate the sampling distribution of a latent personality type. The personality type is not directly determined by either attribute; instead, its sampling probability depends on the corresponding Age–Gender combination. We use six latent personality types: *Neutral*, *Cooperative*, *Cautious*, *Exploratory*, *Aggressive*, and *Distracted*. The sampled personality determines the pedestrian's tendency to select different interaction behaviors. For example, a cautious pedestrian is more likely to slow down, yield, or stop, whereas an aggressive pedestrian is more likely to accelerate through an interaction region.

Consequently, Age and Gender influence pedestrian appearance, motion parameters, reaction distance, and the distribution of interaction behavior tendencies through two complementary mechanisms: direct attribute-conditioned parameterization and personality-conditioned behavior sampling.

These attribute-to-appearance, attribute-to-motion, and attribute-to-personality mechanisms are used only during dataset construction. They are not provided to the navigation agent as inference-time inputs, and the latent personality is not used as an explicit supervision target.

B.2 Goal-Directed Trajectory Annotation

Each pedestrian is first assigned a Semantic Destination, such as a *doorway*, *kitchen*, *living room*, or *bedroom*. We then annotate a sequence of waypoints in a top-down view exported from the Habitat scene. The waypoints define a complete route from the pedestrian's initial position to the assigned destination. Each waypoint may be associated with a semantic location, which describes the region traversed by the pedestrian and provides an organizational reference for assigning interaction behaviors.

Thus, Semantic Destination specifies where the pedestrian is intended to go, whereas the waypoint sequence specifies how the pedestrian reaches that destination. The complete annotated route is used as the initial trajectory reference before simulation execution.

B.3 Predefined Interaction Behaviors and Simulation

For designated interaction points along each trajectory, we predefine an interaction behavior. The behavior set includes speeding up, slowing down, detouring, following the edge, stepping aside, and stopping. The assigned behavior depends on the semantic location and the pedestrian's personality-conditioned behavioral tendency. For example, at a doorway, a cautious pedestrian is more likely to slow down, yield, or stop, whereas an aggressive pedestrian is more likely to speed up and pass through the interaction region.

The manually annotated route is first projected onto Habitat's walkable navigation mesh (NavMesh). It is then executed by the pedestrian simulation policy to obtain a continuous and executable trajectory. At each simulation step, the policy updates the pedestrian's position, heading, and velocity according to the current waypoint, semantic location, the Age–Gender-conditioned motion parameters, latent personality, and the relative state between the robot and the pedestrian.

When the robot enters the reaction region associated with an interaction point, the pedestrian executes the corresponding predefined behavior. The base reaction distance is set to $2\,\mathrm{m}$ and is scaled by the Age–Gender-conditioned reaction weight. The resulting positions, headings, and velocities are recorded as Future Motion. The predefined behavior triggered and executed during the robot–pedestrian interaction is recorded as the Human–Robot Interaction state.

B.4 Robot Reference-Path Correction

The initial robot reference path is obtained from the original VLN episode. After the pedestrian trajectories are inserted and executed in the simulator, we replay the robot reference path and check the spatiotemporal distance between the robot and pedestrians on the same floor. When a potential collision is detected, the system generates multiple lateral detour candidates around the conflict location, projects them onto the walkable NavMesh, and constructs alternative paths using shortest-path planning.

The candidate paths are ranked according to path length, safety clearance, and continuity with the original reference path. When automatic correction succeeds, the system selects a collision-free path with low additional cost. If the corrected path still contains a collision or an undesirable discontinuity, we manually adjust the pedestrian trajectory or the robot reference path. This process continues until the replayed pedestrian and robot reference trajectories remain collision-free throughout the episode.

This correction procedure is used only to construct valid reference trajectories and does not determine the robot's online decisions. During evaluation, the navigation agent must still select among candidate routes and handle potential social risks autonomously.

B.5 Recorded Information

Each episode stores the language instruction, the robot reference path, candidate routes, manually annotated pedestrian trajectories, Semantic Destinations, waypoint-level semantic locations, predefined interaction behaviors, Age, Gender, motion-parameter weights, Future Motion states, and Human–Robot Interaction states.

Semantic Destinations and waypoint-level semantic locations are used primarily to construct goal-directed pedestrian trajectories and assign interaction behaviors. Age, Gender, Future Motion, and Human–Robot Interaction are retained as structured information for training and analysis. The latent personality is used only during dataset construction; it is not provided to the navigation model and is not used as an explicit supervision target.

C Inference-Time Social Context Analysis

This section analyzes how social-context information affects the action distribution of SocioWM. We first compare the action probabilities before and after route-conditioned social reasoning, and then decompose the changes according to Age, Gender, and Human–Robot Interaction. Future Motion is kept unchanged and is not separately visualized in this analysis. We use a fixed trained checkpoint



Figure 5: **Action-probability analysis before and after social reasoning.** The RGB observations and top-down maps show two representative scenes. In the top-down maps, red points denote all valid candidate points generated from the GraphMap, while yellow points highlight the three candidates with the highest initial navigation probabilities, denoted as C1, C2, and C3. The curves compare the initial and final probabilities of STOP and these three candidates. In Scene 1, the highest probability changes from C2 to C1, whereas C2 remains the most probable action in Scene 2.

and perform all analyses at inference time without additional training or parameter updates. At each decision step, the policy generates a dynamically sized set of valid navigable candidate routes from the GraphMap together with a dedicated STOP action. In the top-down maps of Figure 5, red points denote all valid candidate points generated from the GraphMap. The STOP action is always retained in the analysis. Among the non-STOP candidate routes, we select the three candidates with the highest initial navigation probabilities. Their corresponding candidate points are highlighted in yellow and denoted as C1, C2, and C3. Their identities are kept fixed when computing the final probabilities after social-context processing. Therefore, the plotted action set consists of STOP and the three initially highest-scoring candidate routes.

As shown in Figure 5, the effect of social reasoning depends on the interaction context. In Scene 1, C2 is initially preferred, but the final distribution assigns the highest probability to C1. This shift indicates that the predicted social consequences of the candidate routes can alter the action decision when a potential pedestrian interaction is present. In Scene 2, C2 remains the highest-probability action before and after social reasoning, with only a small redistribution among the candidates. Thus, the social branch changes the decision when social information is decision-relevant while preserving the original navigation preference when the interaction risk is weak.

To further analyze the source of these probability changes, we perform an inference-time leave-one-attribute-out analysis. For each scene, one social-context representation is masked while all model parameters and input observations remain unchanged. For an attribute k and an action $q \in \{\text{STOP}, \text{C1}, \text{C2}, \text{C3}\}$, its contribution is measured as

$$\Delta P_k^q = P_{\text{full}}^q - P_{\text{w/o}\,k}^q, \tag{16}$$

where P_{full}^q is the probability produced by the complete model and $P_{\text{w/o}\,k}^q$ is the probability after masking attribute k. A positive value indicates that the attribute increases the probability of action q, whereas a negative value indicates that it suppresses that action. This analysis measures the dependence of the trained policy on each attribute representation and does not correspond to retraining a separate model without that attribute.

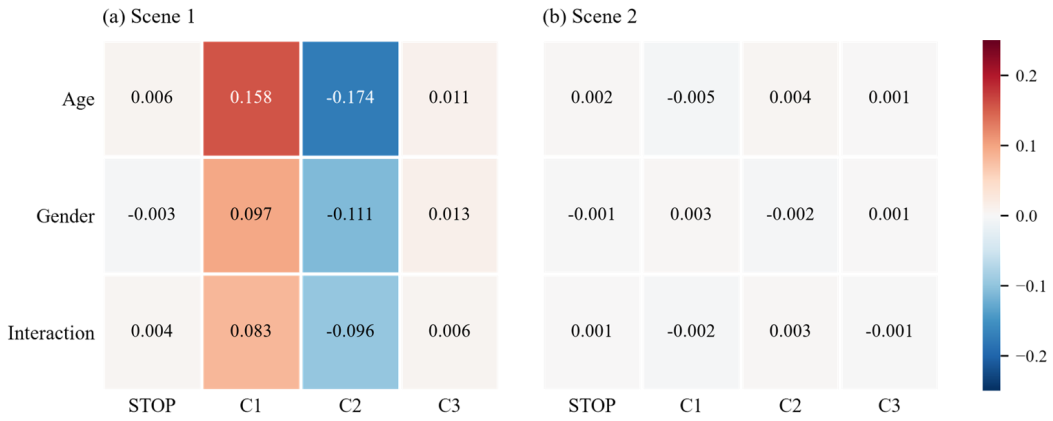

Figure 6: **Attribute-wise action-probability changes.** Each cell shows ΔP_k^q for Age, Gender, and Human–Robot Interaction over STOP and candidate routes C1–C3. In Scene 1, the social-context representations increase the probability of C1 and suppress C2, whereas the changes in Scene 2 remain small across all actions.

Figure 6 decomposes the probability changes shown in Figure 5. In Scene 1, Age, Gender, and Human–Robot Interaction all increase the probability of C1 and decrease the probability of C2, providing complementary evidence for changing the preferred route. In Scene 2, the effects of all three representations remain close to zero, which is consistent with the stable action distribution observed in Figure 5.

D Real-World Deployment Details

We further evaluate SocioWM on a Unitree G1 humanoid robot in unseen indoor environments. The robot is equipped with an Intel RealSense D435 RGB-D camera, which provides synchronized RGB and depth observations at 30 Hz. Model inference is performed on a workstation equipped with an NVIDIA RTX 5090 GPU. The robot streams its RGB-D observations to the workstation, where the high-level policy rescores candidate routes at 2–4 Hz and returns the selected navigation action to the robot. The low-level motion controller executes the action at 10–20 Hz. During deployment, the policy uses only the language instruction and online RGB-D observations; no privileged pedestrian annotations, future trajectories, or interaction labels are provided.