MAction-SocialNav: Multi-Action Socially Compliant Navigation via Reasoning-enhanced Prompt Tuning

Zishuo Wang[1], Xinyu Zhang[1], Zhuonan Liu[1], Tomohito Kawabata[1], Daeun Song[2], Xuesu Xiao[3], and Ling Xiao[1*], Senior Member, IEEE

***Abstract*—Socially compliant navigation requires robots to move safely and appropriately in human-centered environments by respecting social norms. However, social norms are often ambiguous, and in a single scenario, multiple actions may be equally acceptable. Most existing methods simplify this problem by assuming a single "correct" action, which limits their ability to handle real-world social uncertainty. In this work, we propose MAction-SocialNav, an efficient vision language model for socially compliant navigation that explicitly addresses action ambiguity, enabling generating multiple plausible actions within one scenario. To enhance the model's reasoning capability, we introduce a novel meta-cognitive prompt (MCP) method. Furthermore, to evaluate the proposed method, we curate a multi-action socially compliant navigation dataset that accounts for diverse conditions, including crowd density, indoor and outdoor environments, and dual human annotations. The dataset contains 789 samples, each with three-turn conversation, split into 710 training samples and 79 test samples through random selection. We also design five evaluation metrics to assess high-level decision precision, safety, and diversity. Extensive experiments demonstrate that the proposed MAction-SocialNav achieves strong social reasoning performance while maintaining high efficiency, highlighting its potential for real-world human robot navigation. Compared with zero-shot GPT-4o (CoT) and Claude (CoT), our model achieves cleaner action sets (higher APG: 0.595 vs 0.013/0.013), more reliable ranking (higher MAA: 3.571 vs 0.057/0.095), and lower error rates (0.264 vs 0.692/0.649), while maintaining real-time efficiency (1.524 FPS, over 3× faster).**

***Index Terms*—Socially Compliant Navigation, Prompt Tuning, Vision Language Models.**

I. Introduction

THE proliferation of mobile robots in unstructured, human centric environments, ranging from service robots in retail hubs to delivery platforms on urban sidewalks, has elevated the importance of socially compliant navigation [1]. While conventional industrial navigation focuses primarily on path efficiency and obstacle avoidance, robots operating in shared spaces with human must exhibit social compliance. This requires navigating beyond mere geometric collision avoidance to adhere to social norms, such as respecting personal space, yielding right-of-way, aligning with human traffic flows, and avoiding walking on grassed areas [2], [3]. Therefore, the fundamental challenge in social navigation is to formulate a decision-making framework that prioritizes socially compliant actions over purely feasible ones.

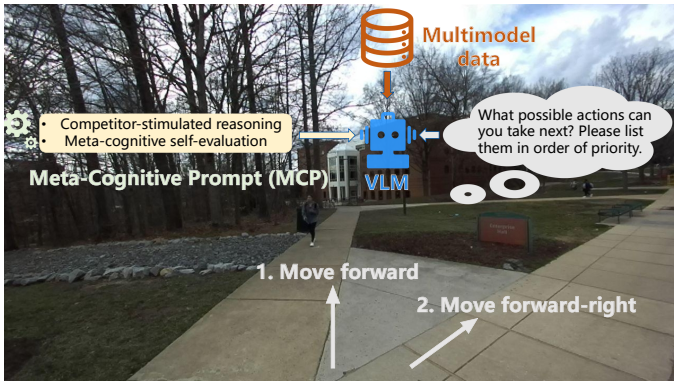

Fig. 1. **Task formulation and high-level concept of multi-action social navigation.** Given multimodal observations, the agent is required to reason about the scene and generate multiple feasible navigation actions ranked by priority, rather than a single deterministic action. A meta-cognitive prompt (MCP) is proposed to stimulate structured reasoning and self-evaluation in the VLM, enabling socially compliant and interpretable decision-making.

Existing approaches [4], [5] predominantly formulate social navigation as a deterministic decision-making problem. However, in real-world social environments, pedestrians often exhibit multiple equally compliant behaviors within the same scene; for example, both stopping and bypassing an obstacle may be feasible choices in certain situations. By enforcing a single ground truth action, existing methods [6], [7] overlook this inherent diversity in human decision-making and penalize valid alternative behaviors. Such deterministic supervision restricts the model to a narrow solution space, preventing it from capturing the full distribution of socially acceptable interactions.

Furthermore, most existing approaches [8], [9] rely on large scale vision language models (VLMs), which suffer from high inference latency, significantly restricting their deployment in real-world robotic systems. Small language models (SLMs) have recently emerged as resource-efficient alternatives to computationally expensive large language models (LLMs) in various tasks. However, their potential has not yet been explored for socially compliant navigation, where efficiency,

Manuscript received: December 25, 2025; Revised: March 8, 2026; Accepted: May 31, 2026.

This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers comments. This work was supported by JSPS KAKENHI Grant No. 24K20787 and NVIDIA Academic Grant Program.

[1]Zishuo Wang, Xinyu Zhang, Zhuonan Liu, Tomohito Kawabata, and Ling Xiao are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan {zishuo.wang.x3, xinyu.zhang.y3, kawabata.tomohito.o8}@elms.hokudai.ac.jp; furisuto1210@gmail.com; ling@ist.hokudai.ac.jp

[2]Daeun Song is with the Department of Artificial Intelligence, Ewha Womans University, Seoul, Republic of Korea songd@ewha.ac.kr

[3]Xuesu Xiao is with the Department of Computer Science, George Mason University, Fairfax, VA, USA xiao@gmu.edu

*Corresponding author: Ling Xiao.

Digital Object Identifier (DOI): see top of this page.

responsiveness, and reliable social reasoning are all critical.

To address the above-mentioned challenges, firstly, we propose MAction-SocialNav as shown in Fig. 1, a socially compliant navigation model capable of predicting multiple plausible actions in a given social environment. To enhance social reasoning performance, we introduce a novel meta-cognitive prompt (MCP) method. Moreover, we construct a multi-action socially compliant navigation dataset and design five evaluation metrics to assess high-level decision precision, safety, and diversity of the proposed model.

The main contributions of this work are summarized as follows:

- We propose a socially compliant navigation model that can predict multiple socially acceptable actions in a given environment. To the best of our knowledge, this is the first work to explicitly model and evaluate action ambiguity in socially compliant navigation.
- We introduce a novel MCP method, which comprises two components: meta-cognitive self-evaluation (MCSE) and competitor-stimulated reasoning (CSR). We also construct a multi-action dataset for evaluating the performance of proposed model. Moreover, we design five evaluation metrics to assess high-level decision precision, safety, and diversity.
- Comprehensive experiments demonstrate that our method significantly outperforms typical zero-shot large scale VLMs in generating safe and socially appropriate navigation actions in complex human–robot interaction scenarios.

II. RELATED WORK

A. Social Robot Navigation

Research on social robot navigation spans from classical hand-engineered models to modern learning-based frameworks. Traditional approaches, such as the social force model [10] and proxemics-based methods [11], [12], explicitly encode human motion patterns and interpersonal spatial constraints. While these methods are interpretable, their limited adaptability constrains performance in dynamic and diverse social environments.

To address these limitations, recent studies increasingly adopt data-driven approaches, including imitation learning for socially compliant trajectory generation [13], [14], deep reinforcement learning for optimizing long-horizon interactions [15]–[17], and trajectory-prediction models for inferring human intent [18], [19]. Graph neural networks (GNNs) further enhance multi-agent reasoning by explicitly modeling relational dependencies among humans and robots [20]. Despite these advances, most learning-based methods focus primarily on generating feasible behaviors and often struggle to capture latent social norms, group dynamics, and nuanced human–robot relationships. As a result, achieving truly socially aware navigation remains an open challenge.

In parallel, recent progress in vision language navigation (VLN) demonstrates that language can serve as an effective interface for embodied navigation, enabling agents to ground high-level intent in perception and to generalize beyond hand-crafted cost functions [21], [22]. Building on this paradigm, several recent works extend VLMs to social navigation settings [4], [23], [24]. Notably, Social-LLaVA [25] introduces the SNEI dataset and formulates social navigation as a multi-stage process involving perception, prediction, chain-of-thought reasoning, action selection, and explanation. Models trained on SNEI, such as LLaVA-v1.5-7B, demonstrate promising capabilities in understanding social cues, anticipating pedestrian behavior, and generating interpretable navigation decisions.

However, three limitations remain. First, real-world social navigation often permits multiple socially acceptable actions, whereas SNEI provides only a single annotated response per scenario. Second, its limited scale (325 multi-turn dialogues) restricts scenario diversity and model generalization. Third, existing work largely overlooks VLM efficiency, despite low-latency inference being essential for real-world robotic deployment.

B. VLM Reasoning.

LLMs have emerged as powerful tools across a wide range of domains due to their strong capabilities in understanding and generating human-like text. However, since vanilla LLMs are primarily trained for generic natural language processing objectives, they often perform poorly on tasks that require structured reasoning or domain-specific decision-making [26].

To address this limitation, Chain-of-Thought (CoT) prompting has been proposed to explicitly elicit intermediate reasoning steps, significantly improving the reasoning performance of LLMs by encouraging step-by-step deliberation. Building upon the linear reasoning paradigm of CoT, Tree-of-Thoughts (ToT) [27] introduces a branching reasoning structure, where LLMs are instructed to simulate discussions among multiple experts and explore diverse reasoning paths before reaching a consensus. These prompting strategies substantially enhance reasoning accuracy, but they also incur considerable inference latency and computational overhead, which limits their applicability in real-time or resource-constrained scenarios.

Motivated by these challenges, SLMs have gained increasing attention as a promising alternative. SLMs are generally defined as models with substantially fewer parameters than large scale LLMs, typically not exceeding 3 billion parameters. Owing to their compact size, SLMs can be deployed on end-user devices such as personal computers and smartphones, even without GPU acceleration. Representative examples include the Phi series [28], TinyLLaMA [29], and NVILA [30]. However, their potential for socially compliant robot navigation remains largely underexplored.

III. METHOD

A. Overview

This paper proposes an efficient and effective VLM for multi-action socially compliant navigation. Fig. 2 illustrates the overall pipeline of the proposed MAction-SocialNav framework. We adopt the NVILA [30] architecture as the backbone and fine-tune it for multi-action social navigation.

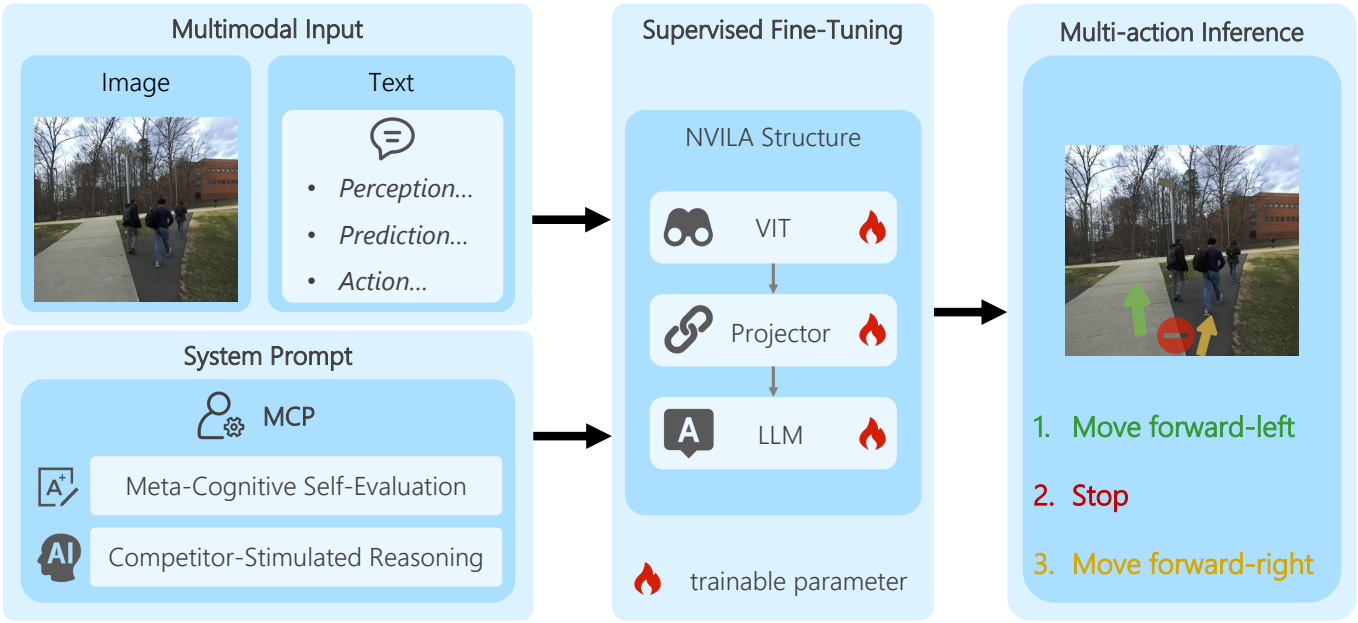

Fig. 2. **Overview of MAction-SocialNav Framework.** We formulate socially compliant navigation as a multi-turn dialogue process. Given a scene observation I and a designed MCP (S_{MCP}) as the system prompt, the model sequentially performs perception and prediction through intermediate queries, and finally generates a ranked set of executable actions.

Specifically, we first reformulate supervision as multi-action ranking over a grounded discrete action space $\mathcal{A}$, enabling the model to learn preferences among plausible actions instead of imitating a single action. Then, we propose a novel meta-cognitive prompt (MCP) method to enhance reasoning capability under SLMs constraints. Finally, we curate a multi-action socially compliant navigation dataset and design five evaluation metrics to assess high-level decision precision, safety, and diversity.

Given a visual observation I and a multi-turn conversation T, the proposed MCP is used as the system prompt and concatenated with the multi-turn conversation T. The goal is to generate a textual response Y that specifies the robot's next behavior.

Formally, the model is trained via full parameter supervised fine-tuning (SFT). The optimization objective is the response sequence $Y = \{y_1, y_2, \ldots, y_L\}$ (L is the response length) via the auto-regressive next-token prediction loss, conditioned on the injected system prompt:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(y_t \mid I, S_{MCP}, T_{<t}, y_{<t}). \tag{1}$$

This formulation ensures that the learned policy p intrinsically incorporates the high-level constraints defined in S_{MCP} without requiring runtime search or external verifiers. Instead of predicting a single deterministic action, we model decision-making as ranking a set of discrete actions, which better reflects the inherent ambiguity of real-world social navigation. By leveraging the auto-regressive dependency on previously generated tokens $y_{<t}$, the model implicitly enforces a hierarchical structure over actions. Conditioning each prediction on prior outputs allows the model to establish a preference ordering, producing a ranked set of socially compliant actions rather than an isolated decision. Details of the discrete action space and ranking protocol are introduced in Subsection III-C.

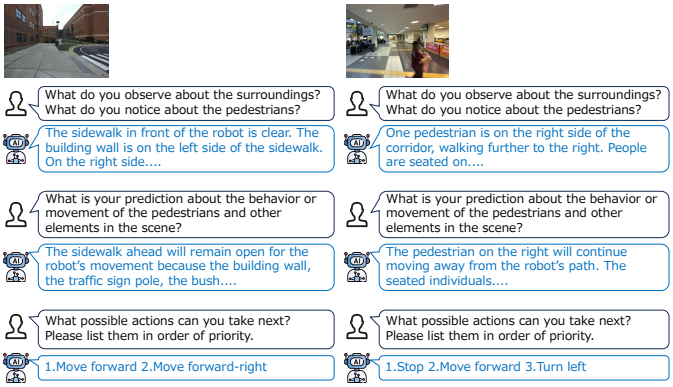

Fig. 3. Multi-turn conversation dataset with action ranking. Each training sample consists of a visual observation paired with a multi-turn dialogue. All assistant responses are annotated by two human annotators. To support multi-action supervision, we introduce a hierarchical action ranking protocol. Specifically, we generate a ranked set of candidate actions based on a priority hierarchy: Feasibility→ Social norms→ Efficiency.

B. Meta-Cognitive Prompt (MCP)

We propose a novel meta-cognitive prompt method which consists of a **Meta-Cognitive Self-Evaluation** (MCSE) and **Competitor-Stimulated Reasoning (CSR)** (as shown in Table I).

a) MCSE. The proposed MCSE, denoted as S_{meta}, enforces an implicit recursive optimization loop. By explicitly introducing a scoring signal and a decision threshold, the model

is encouraged to internally reallocate attention toward safety-critical and socially relevant factors (e.g., pedestrian proximity and trajectory conflicts) before committing to the final action decision.

b) CSR. Prior studies on LLMs suggest that defining a competitor and reference standard can influence a model's effort allocation and reasoning depth [31]. Motivated by this observation, we design a competitor-based system prompt S_{com} to examine their effects in the context of socially compliant robotic navigation. Specifically, S_{com} consists of three modes:

1) Competing against humans (S_{com}^{human}): Encouraging human-like intuition and socially grounded navigation behaviors.

2) Competing against other AI models (S_{com}^{AI}): This establishes a reference based on general intelligence and commonly observed AI-level performance.

3) Competing against self (S_{com}^{self}): Encouraging deeper internal search and self-improvement.

The final prompt S_{MCP} is constructed by concatenating S_{meta} and S_{com}:

$$S_{\mathrm{MCP}} = \{S_{\mathrm{meta}} \,\|\, S_{\mathrm{com}}\}\,. \tag{2}$$

C. Curated Multi-Action Dataset

Existing high-level social navigation datasets typically annotate each scene with a single ground-truth action accompanied by an explanation, which is not suitable for evaluation in our multi-action setting. To address this limitation, we curate a dataset consisting of **789 samples** covering a wide range of real-world navigation environments, including libraries, malls, plazas, streets, campuses, and parks. The dataset also includes scenes with varying human densities, where high-density scenarios naturally require more avoidance, yielding, and re-planning behaviors.

For each image, we manually annotate a ranked subset of a discrete action space $\mathcal{A}$ consisting of six motion primitives: {*Move forward*, *Move forward-left*, *Move forward-right*, *Turn left*, *Turn right*, and *Stop*}. This design captures key interaction intents in social navigation, including progression, local avoidance, yielding, and directional re-planning, and enables seamless integration with downstream low-level motion planners.

To support multi-action supervision, we adopt a strict hierarchical annotation protocol that shifts the learning objective from imitating a single action to modeling a preference distribution over plausible and socially appropriate behaviors. The protocol is defined as follows:

1) Feasibility. Actions that would cause collisions (e.g., with pedestrians or obstacles) or traverse non-drivable regions are removed.

2) Social norms. Among feasible actions, those that maintain comfortable interpersonal distances are ranked higher.

3) Efficiency. Among socially safe actions, those that maximize forward progress are preferred. Specifically, *Move forward* is ranked above *Move forward-left/right*, which are ranked above turning actions. *Stop* is included only when necessary or as a fallback.

Finally, each sample follows a three-stage interaction protocol: (1) *Scene Observation*, (2) *Motion Prediction*, and (3) *Ranked Action Generation*, as shown in Fig. 3. Importantly, we maintain the same input–output format and inference procedure during both training and inference, ensuring a consistent training–inference pipeline.

D. Proposed Evaluation Metrics

We design five evaluation metrics to assess high-level decision precision, safety, and diversity.

Pred@1 evaluates whether the top-ranked predicted action belongs to the ground truth (GT) acceptable action set and is defined as

$$\mathrm{Pred@1} = \mathbb{I}[\hat{a}_1 \in \mathbf{a}], \tag{3}$$

where $\hat{a}_1$ denotes the first action in the predicted action set, $\mathbf{a}$ represents the GT action set.

Pred@n measures prediction precision by comparing each predicted action against the ground truth action set and is defined as

$$\mathrm{Pred@}n = \frac{1}{|\hat{\mathbf{a}}|}\sum_{i=1}^{|\hat{\mathbf{a}}|}\left(\mathbb{I}[\hat{a}_i \in \mathbf{a}] - \mathbb{I}[\hat{a}_i \notin \mathbf{a}]\right), \tag{4}$$

where $|\hat{\mathbf{a}}|$ denotes the number of action of predicted action set. This metric ranges from -1 to 1, where higher values indicate cleaner predictions with fewer extraneous actions.

All-Pred-in-GT (APG) evaluates whether all predicted actions are valid according to the ground truth set and is defined as

$$\mathrm{APG} = \mathbb{I}\left[\forall \hat{a}_i \in \hat{\mathbf{a}},\ \hat{a}_i \in \mathbf{a}\right]. \tag{5}$$

This metric enforces a strict precision constraint and equals 1 only when the predicted action set is a subset of the ground truth.

Multi-action accuracy (MAA) captures ordering sensitivity in multi-step predictions by assigning higher importance to earlier actions. Let $w = [6, 5, 4, 3, 2, 1]$ denote a predefined weight vector. The metric is defined as

$$\mathrm{MAA} = \frac{1}{|\mathbf{a}|}\sum_{i=1}^{\min(|\hat{\mathbf{a}}|,6)} w_i \cdot \mathbb{I}[\hat{a}_i \in \mathbf{a}], \tag{6}$$

and is set to 0 when $|\hat{\mathbf{a}}| > |\mathbf{a}|$. where $|\mathbf{a}|$ denotes the number of action of ground truth set. This metric emphasizes early correct decisions, which are critical in sequential decision-making tasks.

Error rate (ER) measures the proportion of predicted actions that fall outside the annotated acceptable set and is defined as

$$\mathrm{ER} = \frac{1}{|\hat{\mathbf{a}}|}\sum_{i=1}^{|\hat{\mathbf{a}}|}\mathbb{I}[\hat{a}_i \notin \mathbf{a}]. \tag{7}$$

This metric ranges from 0 to 1, where higher values indicate a stronger tendency to generate invalid actions.

TABLE I
DETAILS OF THE PROPOSED MCP METHOD, WHICH IS A COMBINATION OF MCSE AND CSR.

Prompt Type	**Prompt Description**
MCSE	Implement a silent, recursive self-evaluation loop. Before answering, internally generate a draft and score it based on strict safety and social adherence standards. Set 90 as the minimum passing threshold, but do NOT cap the score at 100. If a solution is exceptionally robust or you are highly confident, you are encouraged to assign a score exceeding 100. If the score is below 90, you must critically analyze the flaws, refine your logic, and simulate the outcome again. Repeat this internal iteration until the solution meets or exceeds the 90-point threshold. Output ONLY the final, optimized response without revealing the intermediate thinking steps.
CSR	You are an intelligent assistant specializing in socially compliant robot navigation. You must understand human behaviors, infer intentions, and plan safe, smooth, and socially appropriate paths. You should perform competitively against {humans, other AI models, other AI models like you.}.

TABLE II
COMPARISON OF MACTION-SOCIALNAV WITH AND WITHOUT MCP, AND WITH CAUSAL REASONING.

Method	Pred@1↑	Pred@n↑	APG↑	MAA↑	ER↓
w/o MCP	0.734	0.389	0.532	3.048	0.306
w/ Causal Reasoning	0.747	0.351	0.532	3.202	0.325
w/ MCP	**0.760**	**0.473**	**0.595**	**3.571**	**0.264**

TABLE III
CROSS-VALIDATION RESULTS ACROSS SIX RANDOM DATASET SPLITS.

Method	**Pred@1↑**	**Pred@n↑**	**APG↑**	**MAA↑**	**ER↓**
w/o MCP	0.823±0.059	0.385±0.119	0.502±0.101	2.998±0.464	0.307±0.059
w/ MCP	0.831±0.059	0.426±0.059	0.536±0.059	3.163±0.275	0.287±0.039

TABLE IV
COMPARISON WITH ZERO-SHOT LARGE VLMS.

Method	Pred@1↑	Pred@n↑	APG↑	MAA↑	ER↓
GPT-4o (Raw)	0.570	-0.349	0.025	0.230	0.668
Claude (Raw)	0.405	-0.424	0.000	0.059	0.642
GPT-4o (CoT)	0.620	-0.384	0.013	0.057	0.692
Claude (CoT)	**0.760**	-0.298	0.013	0.095	0.649
GPT-4o(Ruled)	0.646	-0.273	0.013	0.149	0.637
Claude(Ruled)	0.443	-0.658	0.000	0.000	0.829
MAction-SocialNav	**0.760**	**0.473**	**0.595**	**3.571**	**0.264**

IV. EXPERIMENT

A. Experimental Setup

We implement our framework using the NVILA-Lite-2B [30] architecture. We fine-tune the vision projector and LoRA on 4 NVIDIA RTX 8000 GPUs (48GB memory each) with a learning rate of 1×10^{-4}. We set 2 as batch size of each GPU and apply gradient accumulation with 4 steps, resulting in an effective global batch size of 64. Training is conducted for 10 epochs using DeepSpeed ZeRO-2 and $BF16$ precision.

B. Main Results

Table II shows the comparison of MAction-SocialNav with and without MCP, and with causal reasoning. The *w/o MCP* follows a standard reasoning pipeline consisting of three stages: Perception → Prediction → Action. The **w/ Causal Reasoning** variant extends this pipeline by introducing an additional reasoning stage, resulting in Perception → Prediction → Reasoning → Action. The **w/ Causal Reasoning** variant extends this pipeline by introducing an additional reasoning stage, resulting in Perception → Prediction → Reasoning → Action. The causal reasoning baseline inserts one reasoning turn before action generation with the prompt: "Please explain your reasoning process for determining the next action, based on the observed scene and predicted movements." The model then provides action-wise reasons, explaining why each candidate action is assigned a certain priority according to the observed scene and predicted dynamics. For example, "because there are pedestrians directly ahead, the highest priority is to stop." In contrast, our method introduces the MCP as a system-level prompt that guides the model to generate and evaluate multiple candidate actions. Importantly, MCP does not change the model architecture or require additional fine-tuning signals; it only modifies the prompting strategy within the same Perception → Prediction → Action pipeline. As shown in Table II, introducing explicit reasoning does not consistently improve performance and, in several cases, even leads to degradation. In contrast, MAction-SocialNav with MCP consistently outperforms causal reasoning, indicating that meta-cognitive guidance via system prompting is more effective than explicit reasoning for socially compliant multi-action navigation.

We further conduct cross-evaluation using six independent random dataset splits (Table III). Across all metrics, MCP consistently outperforms w/o MCP, while achieving lower or comparable variance, particularly on Pred@1 and ER. These results demonstrate that the performance gains are robust and independent of specific data partitions.

C. Comparison with Zero-Shot Large VLMs

To examine the performance of general-purpose large VLMs in socially compliant navigation, we compare our method with two representative closed-source models, GPT-4o and Claude. Both models are evaluated in a zero-shot setting and are not fine-tuned on the navigation task. Specifically, we experimented with raw prompt, rule-guided prompt, and CoT prompt (as shown in Fig. 4).

As shown in Table IV, zero-shot large VLMs show limited performance under the multi-action setting. GPT-4o (Raw) achieves a Pred@1 of 0.570, while Claude (Raw) performs worse (0.405), indicating weak top-ranked action precision. Adding rule-based constraints brings only limited and inconsistent improvement. GPT-4o (Ruled) slightly improves

Raw Prompt

Given the current observation scenario, as a social robot, select all executable actions from the following six actions: *Move forward, Move forward-left, Move forward-right, Turn left, Turn right, Stop*. Rank the selected actions in descending priority according to: (1) Social Safety, (2) Efficiency. Output exactly one line using the following format: `1.<action> 2.<action> ...`

Rule-Guided Prompt

Given the current observation scenario, as a social robot, first prune infeasible actions while respecting the following navigation constraints:

- Do not enter vehicle lanes.
- Do not step on grass.
- Do not traverse flower beds or landscaped areas.
- Cross roads using zebra crossings when available.
- Avoid socially crowded areas such as restaurant seating areas.

Then select all executable actions from the following six actions: *Move forward, Move forward-left, Move forward-right, Turn left, Turn right, Stop.* Rank the selected actions in descending priority according to: (1) Social Safety, (2) Efficiency. Output EXACTLY one line containing only the ranked actions. Do NOT include any explanation, reasoning, or additional text. `1.<action> 2.<action> ...`

CoT Prompt

Step 1: Briefly describe the environment focusing on pedestrians, obstacles, and free space (max 80 words).
Step 2: Briefly predict the likely movement or behavior of pedestrians and dynamic elements (max 80 words).
Step 3: Select and rank feasible actions from: *Move forward, Move forward-left, Move forward-right, Turn left, Turn right, Stop*.
Rank according to: (1) Social Safety, (2) Efficiency.
Output exactly one line: `1.<action> 2.<action> ...`

Fig. 4. Three prompt configurations used for querying GPT-4o and Claude: Raw, Rule-Guided, and Constrained CoT.

TABLE V
RECALL PERFORMANCE AND AVERAGE PREDICTED ACTION SET SIZE.

Model	R	Avg. Pred.	GT Avg
GPT-4o (Raw)	0.634	3.79	1.81
Claude (Raw)	0.769	4.06	1.81
GPT-4o (CoT)	0.973	5.22	1.81
Claude (CoT)	0.921	4.10	1.81
GPT-4o (ruled)	0.532	2.67	1.81
Claude (ruled)	0.874	3.76	1.81
MAction-SocialNav	0.726	1.53	1.81

Pred@1 (0.570 → 0.646), while Claude (Ruled) shows degraded performance in Pred@n, APG, MAA. Moreover, GPT-4o occasionally produces refusal responses, failing to generate executable actions in 31 of 79 cases (39%), which constitutes task failure in navigation. CoT prompting slightly improves Pred@1 (e.g., GPT-4o: 0.570 → 0.620; Claude: 0.405 → 0.760), but the generated action sets remain noisy, as reflected by negative Pred@n, near-zero APG, and high ER. In contrast, our **MAction-SocialNav** framework significantly improves all metrics, producing cleaner action sets (higher Pred@n and APG), more reliable ranking (higher MAA), and substantially lower error rates.

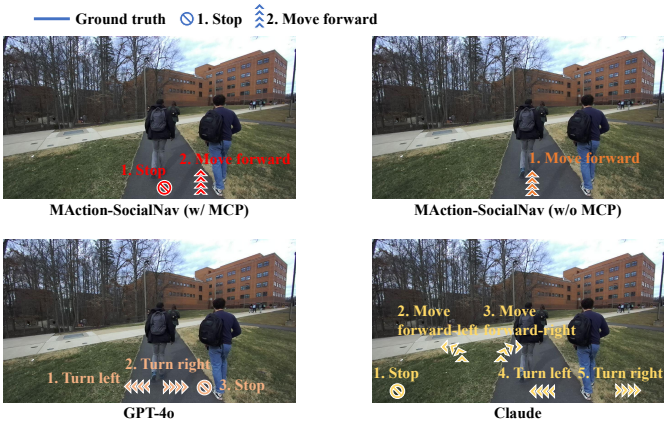

Fig. 5. Visual comparisons. MAction-SocialNav with MCP predicts exactly the same set of feasible actions as the ground truth.

For relatively higher ER, Eq. 7 metric measures the proportion of predicted actions that fall outside the annotated acceptable set. Under this definition, ER can appear relatively high because the metric evaluates each predicted action independently. In our setting, the model generates a set of candidate actions, and ER penalizes every predicted action that does not belong to the annotated acceptable set. Consequently, even if the model successfully includes valid actions in the prediction set, additional exploratory or redundant actions will still increase the ER value.

We also conducted p-value analysis. Comparing our method with Claude and GPT-4o, the differences are highly significant across all metrics, with extremely small p-values (range from 4.0×10^{-20} to 1.4×10^{-3}), confirming that ours substantially outperforms these models in the multi-action social navigation task.

Considering the efficiency, Claude (CoT) and GPT-4o (CoT) achieve inference speeds of 0.452 FPS and 0.314 FPS, respectively. In contrast, our method achieves the highest inference speed of 1.524 FPS, demonstrating both its effectiveness and efficiency for socially compliant navigation.

Furthermore, we provide Table V to report the recall results. While large VLMs (e.g., GPT-4o, Claude) achieve high Recall by over-generating candidates (avg. 2.67–5.22 vs. 1.81 GT), this inflates irrelevant actions, resulting in poor precision metrics (negative Pred@n, near-zero APG). In contrast, MAction-SocialNav generates concise sets (1.53) closely matching the ground-truth distribution, yielding cleaner predictions and superior performance across Pred@n, APG, MAA, and ER.

Fig. 5 shows the visual comparison with GPT-4o, Claude, MAction-SocialNav with MCP, and without MCP. Overall, both GPT-4o and Claude perform poorly on this task, even when explicitly constrained by detailed prompts that encode feasibility and safety norms. Furthermore, Fig. 6 shows that GPT-4o and Claude tend to generate substantially larger candidate action sets than the ground truth. While this increases the chance that correct actions appear in the predictions, it also introduces many unreasonable alternatives. These examples suggest that large multimodal models often fail to capture key social navigation norms encoded in our dataset.

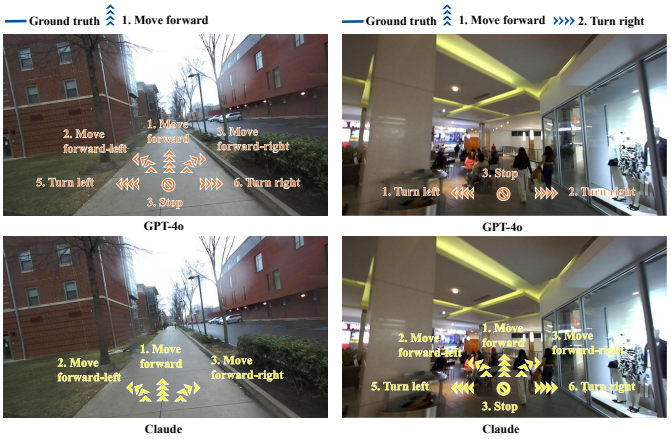

Fig. 6. Visualization examples of GPT-4o and Claude.

TABLE VI
ABLATIONAL STUDIES. INCORPORATING BOTH MCSE AND CSR CAN CONSISTENTLY IMPROVES PERFORMANCE. COMBINING MCSE AND CSR WITH "COMPETING AGAINST OTHER AI MODELS" (S^{AI}_{com}) ACHIEVES THE BEST OVERALL RESULTS.

MCSE	MCP CSR S^{human}_{com}	MCP CSR S^{self}_{com}	MCP CSR S^{AI}_{com}	Pred@1↑	Pred@n↑	APG↑	MAA↑	ER↓
–	–	–	–	0.734	0.389	0.532	3.048	0.306
✓	–	–	–	**0.785**	0.395	0.544	3.137	0.303
–	✓	–	–	**0.785**	0.356	0.468	2.772	0.322
–	–	✓	–	0.760	0.367	0.468	3.052	0.317
–	–	–	✓	0.760	0.392	0.494	3.135	0.304
✓	✓	–	–	0.722	0.392	0.494	3.103	0.304
✓	–	✓	–	0.747	0.378	0.506	3.012	0.311
✓	–	–	✓	0.760	**0.473**	**0.595**	**3.571**	**0.264**

D. Ablational Studies

To validate the efficacy of the proposed MCP, we conducted a comprehensive ablation study to analyze the individual and combined contributions of its core components: MCSE and CSR. As reported in Table VI, the baseline (Row 1) yields a Pred@1 of 0.734 and an MAA of 3.048. Incorporating MCSE alone (Row 2) results in a substantial gain in Pred@1 ($0.734 \rightarrow 0.785$), demonstrating the importance of the self-evaluation mechanism. Similarly, introducing CSR in isolation (Rows 3-5) improves performance, with the $S^{\text{human}}_{\text{com}}$ variant matching the peak Pred@1 score. However, single-module improvements often lack consistency across all metrics (e.g., Pred@n and APG).The results further reveal that the full potential of MCP is unlocked through the coupling of these modules. While combining self-evaluation with $S^{\text{human}}_{\text{com}}$ or $S^{\text{self}}_{\text{com}}$ yields moderate gains, the integration of MCSE with CSR against other AI models ($S^{\text{AI}}_{\text{com}}$) achieves superior performance (Row 8). This configuration outperforms all other settings across comprehensive metrics, boosting Pred@n to 0.473 and MAA to 3.571, while significantly reducing the ER to 0.264. This empirical evidence confirms the synergistic effect of combining self-evaluation with competitive pressure from diverse AI counterparts, leading to more robust reasoning capabilities.

Fig. 7. Examples of the test set across different difficulty levels.

TABLE VII
COMPARISON OF MACTION-SOCIALNAV ACROSS DIFFERENT SCENARIO DIFFICULTY LEVELS. THE PROPOSED MCP EFFECTIVELY ENHANCES THE PERFORMANCE OF THE MODEL IN COMPLEX SOCIAL NAVIGATION ENVIRONMENTS.

Method	Difficulty	Pred@1↑	Pred@n↑	APG↑	MAA↑	ER↓
w/o MCP	Easy	0.966	0.897	0.931	5.690	0.052
	Medium	0.621	0.193	0.414	1.724	0.403
	Difficult	0.571	-0.041	0.143	1.226	0.521
w/ MCP	Easy	0.931	0.897	0.931	5.483	0.052
	Medium	0.621	0.276	0.483	2.702	0.362
	Difficult	0.714	0.159	0.286	2.131	0.421

E. Performance Across Scenario Difficulty Levels

To further evaluate the proposed action ranking framework, we analyze performance across three difficulty levels: *Easy*, *Medium*, and *Difficult*. Scenarios are categorized using a rule-based protocol (Fig. 7) based on (1) road-network complexity (e.g., whether multiple route options are available), (2) pedestrian complexity (e.g., whether a large number of pedestrians influence action selection), and (3) environmental complexity (e.g., cluttered surroundings or static obstacles).

Table VII summarizes performance across three difficulty levels. For models without MCP, evaluation metrics consistently degrade as scenario difficulty increases, reflecting the growing ambiguity in action selection. Medium and Difficult scenarios require finer-grained preference reasoning among multiple feasible actions, which unguided models struggle to handle reliably.

In contrast, MCP significantly improves robustness in moderate and complex settings. Notably, Pred@1 in Difficult scenarios even surpasses that in Medium ones, suggesting that MCP introduces stronger behavioral priors through self-validation and competitive reasoning. In easy scenarios, decision ambiguity is minimal and baseline performance is already near saturation; therefore, MCP mainly adjusts the ranking among valid actions, leading to only minor fluctuations in Pred@1 and MAA.

F. Limitations and Future Work

The current dataset contains 789 samples, which is relatively modest in scale. In future work, we plan to further enhance the dataset by increasing its diversity, improving fine-grained annotations, and expanding its overall size to support the development of more robust socially compliant navigation models. Furthermore, our model focuses on high-level semantic reasoning and action selection. In practical deployment, the ranked actions can serve as high-level commands or semantic action priors, which are then passed to a conventional local planner for trajectory selection or execution. As future work,

we plan to integrate the proposed framework with conventional local motion planners to enable real-world robotic execution.

V. Conclusion

In this paper, we propose MAction-SocialNav, a socially compliant navigation model that explicitly addresses action ambiguity in dynamic social environments. To enhance the reasoning capability of the proposed MAction-SocialNav, we introduce a novel meta-cognitive prompt (MCP). We further curate a dedicated multi-action social navigation dataset and design five evaluation metrics to assess high-level decision-making performance in terms of precision, safety, and diversity. Experimental results demonstrate that MAction-SocialNav substantially outperforms zero-shot large multimodal baselines, offering a practical and efficient pathway toward socially compliant robot navigation in complex real-world environments.

References

[1] A. H. Raj, Z. Hu, H. Karnan, R. Chandra, A. Payandeh, L. Mao, P. Stone, J. Biswas, and X. Xiao, "Rethinking social robot navigation: Leveraging the best of two worlds," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 330–16 337.

[2] 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*, vol. 43, no. 10, pp. 1533–1572, 2024.

[3] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation—a survey," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–36, 2024.

[4] 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*, 2024.

[5] 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 (IROS)*, 2024, pp. 13 837–13 844.

[6] T. Kawabata, X. Zhang, and L. Xiao, "Socialnav-moe: A mixture-of-experts vision language model for socially compliant navigation with reinforcement fine-tuning," *arXiv preprint arXiv:2512.14757*, 2025.

[7] W. Zu, W. Song, R. Chen, Z. Guo, F. Sun, Z. Tian, W. Pan, and J. Wang, "Language and sketching: An llm-driven interactive multimodal multitask robot navigation framework," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 1019–1025.

[8] 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 (ICRA)*, 2025, pp. 7044–7051.

[9] Z. Xu, H.-T. L. Chiang, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah *et al.*, "Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs," in *Proceedings of the 8th Annual Conference on Robot Learning (CoRL)*, 2024.

[10] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[11] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 5495–5502.

[12] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th International Conference on Human-robot Interaction (HRI)*, 2011, pp. 331–338.

[13] L. Tai, J. Zhang, M. Liu, and W. Burgard, "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1111–1117.

[14] B. Ling, Y. Lyu, D. Li, G. Gao, Y. Shi, X. Xu, and W. Wu, "Socialgail: Faithful crowd simulation for social robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 873–16 880.

[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 (IROS)*, 2017, pp. 1343–1350.

[16] 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 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 6015–6022.

[17] T. Kathuria, K. Liu, J. Jang, X. J. Yang, and M. Ghaffari, "Learning implicit social navigation behavior using deep inverse reinforcement learning," *IEEE Robotics and Automation Letters*, 2025.

[18] T. Salzmann, H.-T. L. Chiang, M. Ryll, D. Sadigh, C. Parada, and A. Bewley, "Robots that can see: Leveraging human pose for trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7090–7097, 2023.

[19] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971.

[20] P. Bachiller, D. Rodriguez-Criado, R. R. Jorvekar, P. Bustos, D. R. Faria, and L. J. Manso, "A graph neural network to model disruption in human-aware robot navigation," *Multimedia tools and applications*, vol. 81, no. 3, pp. 3277–3295, 2022.

[21] W. Wu, T. Chang, X. Li, Q. Yin, and Y. Hu, "Vision-language navigation: a survey and taxonomy," *Neural Computing and Applications*, vol. 36, no. 7, pp. 3291–3316, 2024.

[22] K. Chen, D. An, Y. Huang, R. Xu, Y. Su, Y. Ling, I. Reid, and L. Wang, "Constraint-aware zero-shot vision-language navigation in continuous environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[23] D. Song, J. Liang, X. Xiao, and D. Manocha, "Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments," *IEEE Robotics and Automation Letters*, 2025.

[24] 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 (ICRA)*, 2025, pp. 9130–9137.

[25] A. Payandeh, D. Song, M. Nazeri, J. Liang, P. Mukherjee, A. H. Raj, Y. Kong, D. Manocha, and X. Xiao, "Social-llava: Enhancing robot navigation through human-language reasoning in social spaces," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[26] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2021.

[27] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 11 809–11 822.

[28] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi *et al.*, "Textbooks are all you need," *arXiv preprint arXiv:2306.11644*, 2023.

[29] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, and L. Huang, "Tinyllava: A framework of small-scale large multimodal models," *CoRR*, 2024.

[30] Z. Liu, L. Zhu, B. Shi, Z. Zhang, Y. Lou, S. Yang, H. Xi, S. Cao, Y. Gu, D. Li *et al.*, "Nvila: Efficient frontier visual language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 4122–4134.

[31] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2022.