E-SOCIALNAV: EFFICIENT SOCIALLY COMPLIANT NAVIGATION WITH LANGUAGE MODELS

Ling Xiao[1]*, Daeun Song*[2]*, Xuesu Xiao*[2]*, Toshihiko Yamasaki*[3]

[1]Hokkaido University, Sapporo, Japan
[2]George Mason University, Virginia, United States
[3]The University of Tokyo, Tokyo, Japan

ABSTRACT

Language models (LMs) are increasingly applied to robotic navigation; however, existing benchmarks primarily emphasize navigation success rates while paying little attention to social compliance. Moreover, relying on large-scale LMs can raise efficiency concerns, as their heavy computational overhead leads to slower response and higher energy consumption, making them impractical for real-time deployment on resource-constrained robotic platforms. In this work, we evaluate the social compliance of GPT-4o and Claude in robotic navigation and propose E-SocialNav, an efficient LM designed for socially compliant navigation. Despite being trained on a relatively small dataset, E-SocialNav consistently outperforms zero-shot baselines in generating socially compliant behaviors. By employing a two-stage training pipeline consisting of supervised fine-tuning followed by direct preference optimization, E-SocialNav achieves strong performance in socially aware and efficient navigation. The source code will be released upon acceptance.

Index Terms— Motion and Path Planning, Task and Motion Planning, Small Language Models

1. INTRODUCTION

Mobile robots fulfill a wide range of functions, from assisting in healthcare and eldercare to providing delivery and logistics services, and supporting security and surveillance tasks. These roles often require robots to interact effectively with humans and to navigate seamlessly through public spaces shared with pedestrians. In such dynamic environments, it becomes crucial for robots to demonstrate socially compliant behaviors in both interaction and navigation, ensuring safety, efficiency, and user acceptance [1].

The primary challenges of this task lie in understanding and predicting human intentions, managing uncertainty in dynamic and cluttered environments, and balancing efficiency with safety and comfort. To achieve this, robots need to integrate perception, prediction, and planning modules capable of producing socially compliant trajectories that adapt to diverse interaction scenarios.

Existing methods include imitation learning (IL)-based [2], reinforcement learning (RL)-based [3], and large language model (LLM)-based approaches [4, 5]. Among these, LLM-based methods are particularly promising because LLMs provide strong contextual understanding and commonsense reasoning, which align well with the requirements of socially compliant navigation.

Despite recent progress, relying on LLMs may introduce efficiency challenges. For example, VLM-Social-Nav [5] employs GPT-4v to generate navigation instructions; however, due to its large parameter size and the inability to leverage GPU acceleration, this results in significant inference latency. In addition, there has been no systematic evaluation of the zero-shot capabilities of existing LLMs (such as GPT-4 and Claude) for socially aware navigation. Understanding how well off-the-shelf models perform without task-specific training is essential for assessing their readiness for real-world deployment. Building on these insights, it is also critical to design a trainable model that is GPU-accelerated and efficient. Nevertheless, fine-tuning LLMs for this task faces a practical obstacle: high-quality, large-scale datasets are scarce, making it imperative to explore how limited data can be leveraged effectively.

This paper addresses the above-mentioned issues. First, we conduct a comprehensive zero-shot evaluation of GPT-4o and Claude for socially compliant navigation. Second, we proposed E-SocialNav, an efficient LM designed for socially compliant navigation under small-data settings. The main contributions are summarized as below:

- Evaluated GPT-4o and Claude, and developed E-SocialNav for efficient navigation under small-data settings.

- Built a multi-dialog SFT dataset and a single-dialog DPO dataset for socially compliant navigation.

- Identified suitable Small Language Models (SLMs) and Vision Towers (VTs) for this task.

Thanks to JSPS KAKENHI (Grant Number 24K20787) for funding.

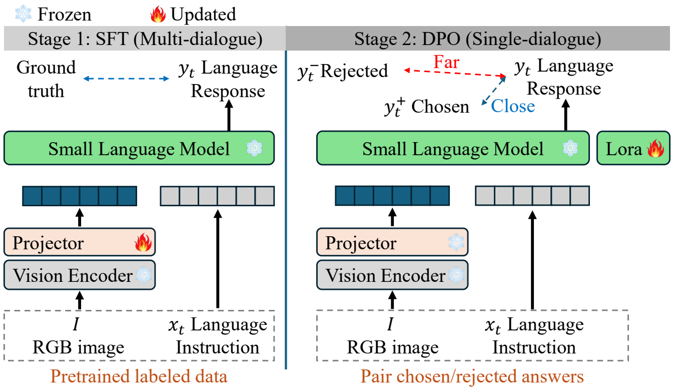
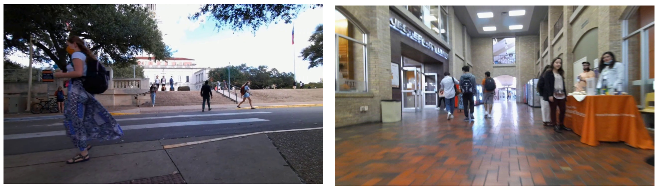


Fig. 1: The detailed structure of E-SocialNav. E-SocialNav is trained in two phases: SFT on multi-turn dialogues, followed by DPO on single-turn pairs. During SFT, only the projector is updated; during DPO, only the LoRA adapter is updated.



Human: "<image> What should the robot do?"
Chosen: "The robot should stop, wait for clear path."
Rejected: "The robot should continue straight."

Human: "<image> What should the robot do?"
Chosen: "The robot should continue straight at a moderate speed."
Rejected: "The robot should stop and wait."

Fig. 2: Visualization of constructed DPO training pairs. The chosen response is annotated by humans, whereas the rejected response is generated by modifying certain facts in the chosen response.

2. RELATED WORK

2.1. Social Robot Navigation

For social robot navigation, safety is paramount [6]. Classical methods enforce collision constraints or fuse multi-sensor data (2-D LiDAR, depth cameras) for smooth avoidance [6].

Safety alone, however, is insufficient in human-populated spaces. Robots must also respect social norms (such as personal space, group dynamics, cultural conventions) to be perceived as acceptable and trustworthy. Traditional methods often ignore these, reducing pedestrians to moving obstacles.

Learning-based approaches seek to encode both safety and social awareness. Demonstration-driven motion learning [7] and RL in simulators [6] show promise but depend on large datasets or highly realistic human simulations, which rarely capture nuanced interactions, yielding policies with poor generalization.

Recently, Large Multimodal Models (LMMs) have opened new directions. LMMs generate high-level actions [1], evaluate trajectories [8], and predict directions [5]. Datasets such as SCAND [9] and MuSoHu [10] further enable socially compliant, human-like navigation. However, research on small language models (SLMs) for this task remains limited.

2.2. Small Language Models

Large Language Models (LLMs) have shown strong abilities in reasoning, planning, and multimodal understanding. While frontier models (e.g., GPT-4, Claude) achieve state-of-the-art performance, their substantial computational demands hinder deployment in robotics and edge devices due to higher inference latency and greater computational consumption.

Recent work therefore emphasizes small language models (SLMs) [11]. Three main directions have emerged: (1) Efficient pretraining and distillation: transferring knowledge from large teachers via distillation or pruning [12] to retain reasoning capacity with lower cost; (2) Parameter-efficient fine-tuning: methods such as LoRA [13] and prompt-tuning enable task specialization with minimal overhead; (3) Architectural and training innovations: lightweight models (e.g., TinyLLaMA [14]) and data-efficient recipes build compact yet capable SLMs.

This reflects a shift from pure scaling toward deployability. By aligning efficiency with contextual reasoning, SLMs offer a practical path to bring language models into real-world interactive systems where resources, cost, and latency are critical.

3. METHODS

Socially compliant navigation aims to generate trajectories that are not only efficient and collision-free but also consistent with human social norms. Conceptually, this can be viewed as optimizing a composite objective that balances three factors: (i) progress toward the goal, (ii) safety in avoiding collisions and maintaining appropriate distances from obstacles, and (iii) adherence to socially compliant behaviors.

The overall framework of E-SocialNav is illustrated in Figure 1. E-SocialNav consists of two training phases:
Supervised Fine-tuning (SFT): We optimize only the projector. To enhance robust multimodal understanding, we employ multi-dialog datasets in which each training sample contains multi-turn conversations paired with corresponding images. This design enables the model to learn not only accurate perception but also context-aware reasoning across dialogue turns.

Formally, given an image I and T dialogue turns $\{(x_t, y_t)\}_{t=1}^{T}$, we encode I with a vision tower (VT) and a projector to obtain visual tokens $v(I)$, and form the multimodal context $x_t = [c_t; v(I)]$ by concatenating textual context c_t and $v(I)$. Let $y_{t,1:N_t}$ be the tokenized assistant response and $y_{t,<n} = (y_{t,1}, \dots, y_{t,n-1})$.

Table 1: Experimental results comparing off-the-shelf models and variants of the proposed method. SFT(X) means the components X are trainable in Stage I (supervised fine-tuning); DPO(Y) means Y are trainable in Stage II (direct preference optimization). Components not listed are frozen. Best performance is bolded.

	Model	VT	SLM	BERTScore-P↑	BERTScore-R↑	BERTScore-F1↑	SBERT-cos↑	SMS↑	FPS↑
Off-the-shelf	Claude	-	-	-0.233	0.387	0.059	0.664	0.641	0.087
	GPT-4o	-	-	0.076	0.443	0.254	0.672	0.651	0.212
E-SocialNav (Ours)	SFT(projector+lora+vision)	SigLIP [15]	Phi-2-2.7B [16]	0.585	0.434	0.509	0.744	0.802	0.978
	SFT(projector+lora)	SigLIP [15]	Phi-2-2.7B [16]	0.640	0.545	0.592	0.756	0.813	1.553
	SFT(projector)	SigLIP [15]	Phi-2-2.7B [16]	0.551	0.658	0.604	0.780	0.828	1.828
	SFT(projector) + DPO(lora)	SigLIP [15]	Phi-2-2.7B [16]	**0.706**	**0.671**	**0.688**	**0.814**	**0.846**	**2.354**

User: “What do you perceive from the image?”
E-SocialNav: “The robot is moving forward on a sidewalk with a very low crowd density. There is one human, a person wearing a black hat, moving from the northwest to the southwest at a slow speed.”
Ground truth: “There is a human on the left side of the sidewalk moving from the northwest to the southwest at a far distance, wearing a white shirt.”

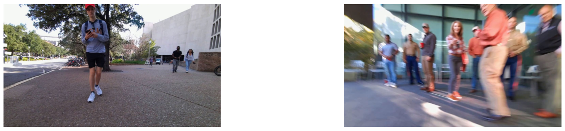
User: “What do you predict these humans will do next?”
E-SocialNav: “The person with a black hat walking from the northwest to the southwest will continue moving at a moderate distance.”
Ground truth: “The human will continue walking from the northwest to the southwest.”

User: “What should the robot do?”
E-SocialNav: “The robot should continue moving forward at a moderate speed.”
Ground truth: “The robot should continue moving forward at a moderate speed.”

Fig. 3: Visualizations: E-SocialNav accurately captures social-compliance cues from the image.

Table 2: Experimental results across different VT and SLM configurations. Best performance is bolded.

	VT	SLM	BERTScore-P↑	BERTScore-R↑	BERTScore-F1↑	SBERT-cos↑	SMS↑
E-SocialNav (Ours)	CLIP [17]	Phi-2-2.7B [16]	0.555	0.658	0.605	0.803	0.768
	Dino [18]	Phi-2-2.7B [16]	0.692	0.663	0.677	0.801	0.833
	SigLIP [15]	TinyLlama-1.1B-Chat-v1.0 [14]	0.473	0.576	0.523	0.733	0.789
	SigLIP [15]	stablelm-2-zephyr-1_6b [19]	0.700	0.622	0.661	0.788	0.837
	SigLIP [15]	Phi-2-2.7B [16]	**0.706**	**0.671**	**0.688**	**0.814**	**0.846**



User: “What should the robot do?”
Gt: “The robot should turn left at a slow speed.”
E-SocialNav: “The robot should stop.”

User: “What should the robot do?”
Gt: “The robot should turn slightly to left.”
E-SocialNav: “The robot should continue moving forward at a moderate speed. It should maintain a safe distance from the group of people, keeping a slight turning radius.”

Fig. 4: Visualization of failure cases. Gt: Ground truth.

The SFT objective averages the next-token negative log-likelihood over response tokens:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \frac{1}{\sum_{t=1}^{T} N_t} \sum_{t=1}^{T} \sum_{n=1}^{N_t} \big[-\log \pi_\theta\big(y_{t,n} \mid x_t,\, y_{t,<n}\big)\big],$$

where π_θ denotes the conditional probability distribution defined by the model parameters θ. Loss is computed only on assistant responses; prompts and image tokens are excluded.

Direct Preference Optimization (DPO): For each input, two candidate responses are provided. The *chosen response* is the human-annotated ground-truth answer, considered the most reliable. The *rejected response* is constructed by modifying the ground-truth answer with localized errors. Some examples are give in Figure 2:

Formally, for each input context x_t (including visual tokens from I), we pair a *chosen* response y_t^+ and a *rejected* response y_t^-. The sequence log-likelihoods are computed by summing token log-probabilities over supervised positions:

$$\ell_\theta^+(t) = \sum_n \log \pi_\theta\big(y_{t,n}^+ \mid x_t,\, y_{t,<n}^+\big), \tag{1}$$

$$\ell_\theta^-(t) = \sum_n \log \pi_\theta\big(y_{t,n}^- \mid x_t,\, y_{t,<n}^-\big). \tag{2}$$

Let us define the log-likelihood advantage

$$\Delta_\theta(t) = \ell_\theta^+(t) - \ell_\theta^-(t).$$

The DPO objective is the average binary logistic loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log \sigma\big(\beta\, \Delta_\theta(t)\big), \tag{3}$$

where $\sigma(\cdot)$ is the logistic sigmoid and $\beta > 0$ is an inverse-temperature hyperparameter controlling the sharpness of preference learning. In practice, we set $\beta = 0.1$, which provides stable gradients without over-amplifying preference margins.

4. EXPERIMENTS

4.1. Experimental Settings

The projector is a two-layer MLP. For evaluation, we use BERTScore, SBERT-cosine, and Sentence Mover's Similarity (SMS), as they emphasize semantic similarity; metrics such as BLEU and ROUGE mainly capture word overlap and are less suitable for our task.

Based on the SNEI dataset [1], which is derived from SCAND [9] and MuSoHu [10], we construct a multi-dialog dataset comprising 325 egocentric video-derived samples, each paired with five conversations. Among these, 60 samples are randomly selected for testing, while the remaining 265 are used for training. We also derive a DPO dataset (see Section 3 **Direct Preference Optimization (DPO)**). Training follows a two-stage schedule on four A100 GPUs and finishes in under one hour. Stage I updates the projector for 20 epochs with learning rate 5×10^{-5} and warm-up ratio 0.03 using FlashAttention-2. Stage II applies DPO for 5 epochs with the same settings.

4.2. Experimental Results

Accuracy. With GPT-4v deprecated, GPT-4o serves as the GPT baseline. From Table 1, both Claude and GPT-4o exhibit limited social compliance, while E-SocialNav aligns more closely with human annotations, achieving higher semantic-similarity scores. In the low-data regime (265 images for training), SFT performs best when the backbone is frozen and only the projector is trained (Stage I). Adding Stage II DPO fine-tuning further improves performance.

Efficiency. E-SocialNav builds on the compact 2.7B Phi-2 backbone, with Stage I updates only the projector and Stage II applies lightweight DPO. These choices keep training compute modest and reduce inference memory and latency, enabling deployment on resource-constrained hardware.

Visualizations. As can be seen from Figure 3, the proposed E-SocialNav produces responses that closely align with human annotations, reflecting both higher semantic fidelity and stronger social compliance.

Variations on VT and SLM. We conduct experiments by varying both the VTs and SLMs. The VTs evaluated include CLIP [17], DINO [18], and SigLIP [15], while the SLMs considered are Phi-2-2.7B [16], TinyLlama-1.1B-Chat-v1.0 [14], and StableLM-2-Zephyr-1.6B [19]. Among all combinations, SigLIP paired with Phi-2-2.7B consistently achieves the best performance across all evaluation metrics (Table 2).

Failure Analysis and Future Works. As shown in Figure 4, E-SocialNav recommends "stop", whereas the ground-truth annotation prescribes "turn left at slow speed". This divergence reflects the inherently conservative bias often adopted during human annotation and underscores the difficulty of establishing a universally valid social standard for navigation. Moving forward, we plan to (i) conduct subjective user studies and (ii) construct a fine-grained, large-scale benchmark dataset that captures diverse cultural norms and situational contexts. Such efforts aim to provide a more balanced foundation to mitigate annotation bias and advance the development of socially compliant navigation models that are universally adaptable.

5. CONCLUSIONS

In this paper, we first examined the effectiveness of off-the-shelf LLMs, including Claude and GPT-4o, and found that they exhibit limited social compliance in navigation tasks. To address this limitation, we proposed E-SocialNav, a lightweight model designed for socially compliant navigation under small-data settings. By adopting a two stage training pipeline consisting of SFT and DPO, E-SocialNav demonstrates substantially improved social compliance compared to zero-shot baselines. Notably, our framework leverages an SLM, which, owing to its relatively small size, enables faster response times, reduced energy consumption, and more practical deployment.

Acknowledgments

This research was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 24K20787.

6. REFERENCES

[1] Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dinesh Manocha, and Xuesu Xiao, "Social-llava: Enhancing robot navigation through human-language reasoning in social spaces," *arXiv preprint arXiv:2501.09024*, 2024.

[2] Catie Cuan, Tsang-Wei Edward Lee, Emre Fisher, Anthony Francis, Leila Takayama, Tingnan Zhang, Alexander Toshev, and Sören Pirk, "Gesture2path: Imitation learning for gesture-aware navigation," in *International Conference on Social Robotics (ICSR)*, 2024, pp. 264–279.

[3] Tribhi Kathuria, Ke Liu, Junwoo Jang, X Jessie Yang, and Maani Ghaffari, "Learning implicit social navigation behavior using deep inverse reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 10, no. 5, pp. 5146–5153, 2025.

[4] Ling Xiao and Toshihiko Yamasaki, "Llm-advisor: An llm benchmark for cost-efficient path planning across multiple terrains," *arXiv preprint arXiv:2503.01236*, 2025.

[5] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha, "Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, 2024.

[6] Jing Liang, Utsav Patel, Adarsh Jagan Sathyamoorthy, and Dinesh Manocha, "Crowd-steer: Realtime smooth and collision-free robot navigation in densely crowded scenarios trained using high-fidelity simulation," in *29th International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021, pp. 4221–4228.

[7] Huihui Sun, Weijie Zhang, Runxiang Yu, and Yujie Zhang, "Motion planning for mobile robots—focusing on deep reinforcement learning: A systematic review," *IEEE Access*, vol. 9, pp. 69061–69081, 2021.

[8] Siddarth Narasimhan, Aaron Hao Tan, Daniel Choi, and Goldie Nejat, "Olivia-nav: An online lifelong vision language approach for mobile robot social navigation," *arXiv preprint arXiv:2409.13675*, 2024.

[9] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11807–11814, 2022.

[10] Duc M Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 7442–7447.

[11] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang, "Tinyllava: A framework of small-scale large multimodal models," *arXiv preprint arXiv:2402.14289*, 2024.

[12] Savitha Viswanadh Kandala, Pramuka Medaranga, and Ambuj Varshney, "Tinyllm: A framework for training and deploying language models at the edge computers," *arXiv preprint arXiv:2412.15304*, 2024.

[13] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022, vol. 1(2), p. 3.

[14] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu, "Tinyllama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

[15] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer, "Sigmoid loss for language image pre-training," in *the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11975–11986.

[16] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al., "Phi-2: The surprising power of small language models," *Microsoft Research Blog*, vol. 1, no. 3, pp. 3, 2023.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al., "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, 2024.

[19] Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al., "Stable lm 2 1.6 b technical report," *arXiv preprint arXiv:2402.17834*, 2024.