2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Human2Nav: Learning Crowd Navigation from Human Videos across Robots via Feasibility-Guided Flow Matching

Shenghong Zhang[1], Junjie Chen[1], Sichi Yan[1], Yutong Ban[1] and Xiao Li[1]

***Abstract*— Enabling robots to navigate safely and efficiently in dynamic, crowded environments requires learning from large-scale demonstrations, which are costly and unsafe to collect on physical platforms. While human videos offer a rich and scalable alternative, transferring these motion patterns to robots is challenged by the embodiment gap across observation and action spaces. This paper presents Human2Nav, a data-efficient framework that learns navigation policies directly from human videos via test-time feasibility-guided flow matching. Human2Nav employs a bird's-eye-view representation to align visual observations and trains a conditional flow matching model to capture nuanced human navigation patterns. Crucially, we introduce a training-free feasibility guidance mechanism that during inference steers generated trajectories to satisfy heterogeneous robot-specific kinematic and dynamic constraints without retraining. Extensive experiments in simulation and on real-world heterogeneous robotic platforms demonstrate that Human2Nav achieves superior data efficiency and navigation performance compared to model-based and learning-based baselines, while ensuring safe and executable trajectories across diverse crowd scenarios.**

I. INTRODUCTION

Enabling autonomous robots to navigate safely and efficiently in dynamic, crowded environments remains a central challenge in robotics [1]. Unlike model-based planners, learning-based approaches can capture the subtle and complex dynamics necessary for navigating dense human crowds [2], [3]. However, their success hinges on access to large-scale, diverse demonstrations. Collecting such robot-centric data is notoriously expensive and often unsafe, particularly in real-world crowded scenarios, creating a significant bottleneck that limits the scalability and generalization of learning-based navigation systems. One approach to alleviating this data scarcity is teleoperating robots in real crowds to collect task-relevant demonstrations [4]. While effective, this process is costly, time-consuming, and raises safety concerns. Alternatively, wearable sensors can record human trajectories for later transfer to robots [5]. Yet, these datasets often exhibit significant heterogeneity in both observation perspective and locomotion dynamics, making direct application to robot navigation challenging. Simulation-based reinforcement learning methods [6], [7] offer another path,

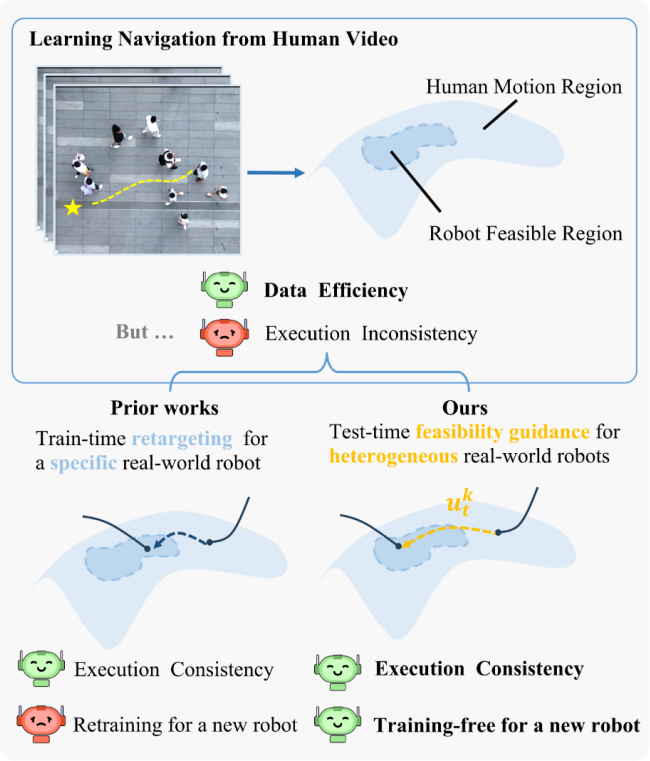


Fig. 1: Human2Nav learns a conditioned flow-matching model for crowd navigation from passive human videos, enabling data-efficient policy training. At test time, feasibility guidance term u_t^k adapts the policy to heterogeneous robot dynamics, bridging the domain gap between human motions and robot capabilities and ensuring safe, executable navigation without retraining.

but they typically suffer from sample inefficiency and difficulties in sim-to-real transfer, especially in dense, dynamic environments.

Large-scale pedestrian trajectory datasets from UAVs [8], [9] or surveillance systems [10] offer rich, low-cost observations of natural crowd behaviors. This raises a key question: *Can we transfer motion patterns from human videos to robots with different dynamics and constraints?* This is challenging due to two fundamental gaps: observation mismatch—humans and robots perceive environments differently—and action mismatch—human motions may be infeasible for robots. Bridging these gaps requires an intermediate representation or mapping that enables safe, executable navigation, but direct application of human data remains fundamentally difficult.

To this end, we present Human2Nav, a framework that enables robots to learn from large-scale, passively collected pedestrian trajectories by addressing both observation and

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4707400 and in part by the NSFC under Grant 52405029.

In this work, we used large language models (LLMs) solely as general-purpose tools. Specifically, we employed LLMs to improve the clarity and readability of the paper.

[1] School of Mechanical Engineering, Shanghai Jiao Tong University {zsh000, sjtu_cjj, sichiYan, yban, sjtu_lixiao}@sjtu.edu

action gaps as shown in Figure 1. We first employ a bird's-eye-view (BEV) representation to align aerial observations with ground robot sensors. Leveraging flow matching, Human2Nav learns a conditional vector field that generates trajectories reflecting real crowd dynamics. However, human trajectories often include maneuvers—such as sharp turns or in-place rotations—that many robots cannot execute. Directly applying such trajectories can result in infeasible or unsafe motions. To overcome this, we introduce feasibility-guided flow matching, a training-free approach that embeds robot-specific kinematic and dynamic constraints into trajectory generation. Unlike prior methods requiring retraining or post-processing, Human2Nav produces safe, executable trajectories that generalize across heterogeneous robots while remaining data-efficient and fully leveraging human motion priors.

To summarize, our main contributions are as follows:

- We propose Human2Nav, a data-efficient framework that learns crowd navigation behaviors directly from human videos, achieving scalable policy learning and superior effectiveness compared to representative baselines.
- We introduce feasibility-guided inference, a training-free approach that embeds robot-specific kinematic and dynamic constraints into trajectory generation, enabling safe and practical deployment across heterogeneous robots.
- We demonstrate real-world deployment of Human2Nav on heterogeneous robotic platforms, achieving safe and efficient crowd navigation without platform-specific modifications, highlighting its strong generalization and readiness for real-world use.

II. RELATED WORK

A. Learn from Human Videos

Leveraging human videos for robot learning poses significant challenges, as these datasets often lack direct robot actions and embodiment information. Prior work has explored several strategies, including extracting keypoint-based trajectories and transferring them to robots [11], [12], adopting object-centric representations that condition motion on scene elements or abstract actions [13], and incorporating partial real-robot data during training to guide policy learning [14]. Along similar lines, DeepMoTIon learns human-aware navigation from pedestrian surveillance data for human imitation and safe crowd navigation [15]. Despite these advances, human motions frequently differ from robot kinematics and dynamics, leading to infeasible or unsafe behaviors. To address this, existing approaches focus on cleaning or adapting human data before training, such as using privileged motion imitation policies [16], applying nonparametric regression with graph-based heuristics on paired datasets [17], spatio-temporal motion retargeting of keypoint trajectories [11], or two-stage retargeting pipelines for humanoid robots [18]. These methods share the common goal of extracting reliable motion labels from noisy human data for downstream training. In contrast, our work is, to the best of our knowledge, the first to extend human video learning to crowd navigation. Although navigation involves fewer degrees of freedom than manipulation, it still suffers from infeasible motions and new forms of observation inconsistency. We address these challenges by constructing ego-centric BEV representations to resolve observation gaps and introducing feasibility-aware guidance at inference, which adapts predicted trajectories to diverse real-world robot platforms without requiring retraining. This approach enables scalable, safe, and executable navigation policies directly from abundant third-person human video data.

B. Generative Models for Robot Navigation

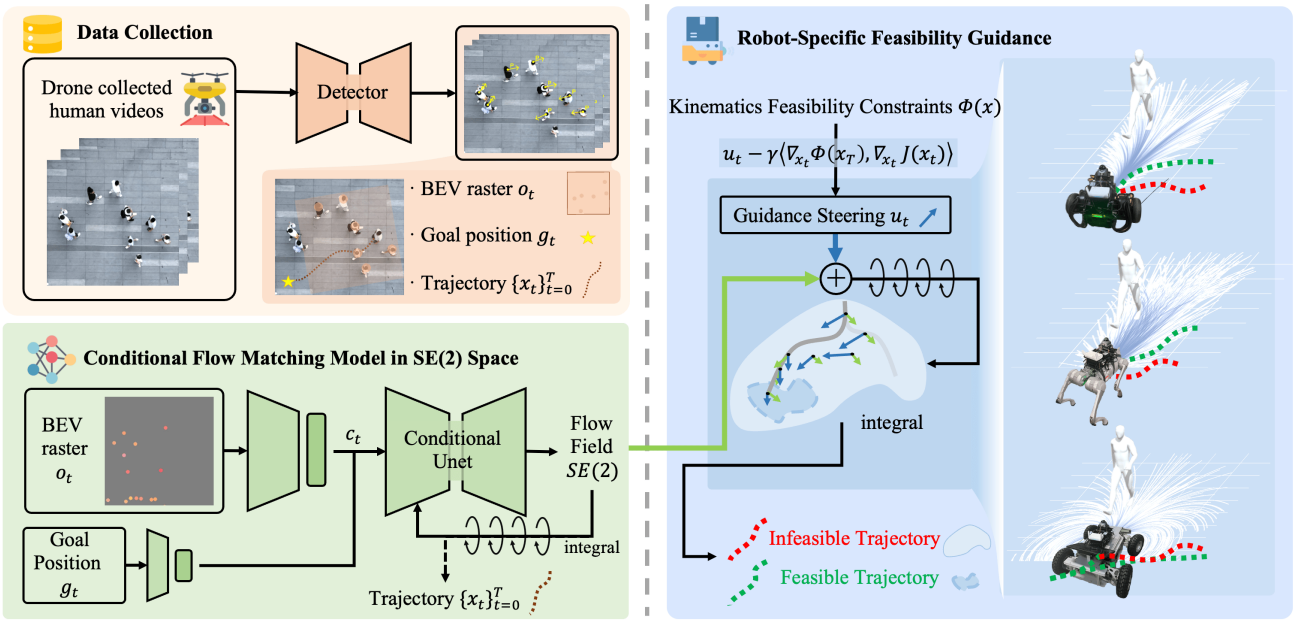
Continuous dynamics-based generative models have recently shown strong multi-modal trajectory modeling capabilities in high-dimensional continuous spaces, making them increasingly popular for robot manipulation and navigation. These methods formulate trajectory generation as a conditional process, leveraging task-relevant embeddings and observations. Diffusion models were first applied to robotics in [19], modeling visuomotor policies as a conditional denoising diffusion process over the action space. By exploiting gradient information from action score functions [20], these models enable flexible conditioning and have been extended to multi-sensor inputs and navigation tasks [21], [22]. However, their iterative denoising process is computationally expensive, limiting real-time applicability. To address this, Flow Matching [23] replaces stochastic sampling with deterministic ODE integration via direct modeling of velocity fields, improving sampling efficiency while preserving the multi-modality of diffusion models. Flow-based generative approaches have demonstrated practical advantages in continuous control [24] and navigation [25], supporting rich conditioning inputs such as outputs from visual-language models. Extensions like [26] further incorporate control barrier functions to enforce safety guarantees. Building on these advances, our work leverages flow matching to learn crowd navigation policies from heterogeneous pedestrian trajectory data, and crucially embeds robot-specific kinematic constraints into the generation process, bridging the gap between human demonstrations and real-world robot execution.

III. METHOD

This section presents Human2Nav, which learns crowd navigation from human videos with feasibility-guided flow matching. As shown in Fig. 2, it has two stages: (1) build a BEV human-trajectory dataset and train a conditional flow-matching model, and (2) apply training-free, robot-specific feasibility guidance at inference to produce executable trajectories. This design reduces data needs and bridges the gap between human demonstrations and robot execution.

A. Training Phase: Flow Matching from Human Trajectories

To capture natural human navigation behaviors, we construct a dataset of pedestrian trajectories from UAV-mounted cameras. We develop a CenterNet-like [27] oriented keypoint detector for UAV imagery that localizes each pedestrian's



(a) Training Phase (b) Inference Phase

Fig. 2: Overview of the proposed **Human2Nav** framework. (a) *Training phase*: Drone-collected human videos are processed to construct BEV rasters o_t, goal positions g_t, and pedestrian trajectories $\{x_t\}_{t=0}^T$. A conditional flow matching model in SE(2) space is trained to capture human motion patterns, where integrating the predicted flow field generates trajectory distributions. (b) *Inference phase*: At test time, feasibility-guided inference embeds robot-specific kinematic and dynamic constraints $\Phi(x)$ into the predicted flow, steering trajectories toward the feasible set and enabling safe, executable navigation across heterogeneous robots (red: infeasible, green: feasible).

head and estimates body orientation. Keypoints are trained with a focal loss, while orientation and sub-pixel offsets are regressed via L1 loss. Detections are associated across frames using the Hungarian algorithm to produce trajectories with position and heading. Assuming a flat ground plane, we compute a homography matrix $\mathbf{H}$ to project image coordinates to metric-consistent BEV coordinates:

$$\mathbf{p}_{\text{world}} = \mathbf{H} \cdot \mathbf{p}_{\text{image}} \tag{1}$$

where $\mathbf{p}_{\text{image}} = [u, v, 1]^\top$ denotes homogeneous pixel coordinates. The resulting dataset consists of timestamped pedestrian poses and inferred goals, forming training tuples (BEV raster, trajectory, goal).

We formulate navigation as an observation-conditioned imitation learning problem, modeling trajectory generation with a conditional flow field governed by

$$\dot{x}_t = u_t(x_t, c), \quad x_0 \sim p_0(x) \tag{2}$$

where c denotes conditioning inputs (BEV raster observations and goals), and u_t is the conditional vector field to be learned. We model $x_t = (x, y, \theta)$ and use trajectories of length T=3.2s with a sampling interval of 0.1s. The BEV raster has 3 channels encoding binary occupancy and velocity components (v_x, v_y). The network f_θ follows a U-Net backbone similar to [19], and conditions on c via FiLM modulation. We set $p_0(x)$ as an isotropic Gaussian in the $\mathfrak{se}(2)$ tangent space, i.e., $x_0 = \exp(\xi_0)$ with $\xi_0 \sim \mathcal{N}(0, I)$, which is mapped to SE(2) via the exponential map. Following the conditional flow matching framework [23], the training objective minimizes

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, p(x_0, x_1)} \left\| f_\theta(x_t, c, t) - u_t(x_t \mid x_0, x_1) \right\|^2 \tag{3}$$

with x_t sampled along the probability path $\mathcal{N}(x_0 \exp(t \cdot \log(x_0^{-1} x_1)), \sigma)$, and $u_t(x_1|x_0) = \log(x_t^{-1} x_1)$. We sample t from a Beta distribution during training and use uniform t at inference following [24].

To ensure geometric consistency on SE(2), we perform both interpolation and integration in the Lie group. Geodesic interpolation is given by

$$x_t = x_0 \exp\big(t \cdot \log(x_0^{-1} x_1)\big) \tag{4}$$

and trajectory integration during rollout follows

$$X_{t+\Delta t} = X_t \exp(\Delta t \cdot V_t) \tag{5}$$

where $X_t \in \text{SE}(2)$ denotes the current state and $V_t \in \mathfrak{se}(2)$ is the velocity in the Lie algebra. This formulation respects the topology of orientation space, yielding stable and geometrically consistent trajectory evolution.

B. Inference Phase: Feasibility-Aware Guidance

While the BEV representation bridges the observation gap between third-person videos and egocentric robot control, human trajectories may still be infeasible for robots with actuation and safety limits. To address this, we formulate inference as a constrained optimal control problem. Given a pre-trained conditional flow matching model that defines a vector field $f_t^p(x)$ in the trajectory space, we introduce a steering term u_t to modulate the dynamics at test time:

$$\dot{x}_t = f_t^p(x_t) + u_t \tag{6}$$

The objective is to steer the generated trajectory toward the feasible set while remaining close to the human-induced

distribution. This is achieved by minimizing a constraint-aware cost function $\Phi(x)$ subject to the robot dynamics:

$$J(u) = \alpha\, \Phi(x_T^u) + \int_0^T L(u_t)\, \mathrm{d}t \quad \text{s.t.} \quad \dot{x}_t = h_t(x_t, u_t), \quad x_0 = x_{\text{init}} \tag{7}$$

Instead of retraining the flow matching model, we adopt a training-free guidance strategy inspired by [28], [29] and Pontryagin's Maximum Principle (PMP). We propagate co-states backward and iteratively refine the steering term u_t using the Extended Method of Successive Approximations (E-MSA), effectively nudging flow-generated trajectories toward feasibility while staying close to the learned human distribution. Physical feasibility is enforced via soft margin constraints on key motion quantities, including maximum velocity, acceleration, curvature, and angular velocity:

$$\mathcal{C}(x, \text{limit}) = \mathbb{E}_t\Big[\tfrac{1}{\gamma} \cdot \text{softplus}\big(\gamma\big(\tfrac{|x_t|}{\text{limit}} - 1 + \delta\big)\big)\Big], \tag{8}$$

where γ controls sharpness and δ provides a tolerance margin. The overall constraint-aware penalty is then aggregated as:

$$\begin{aligned}\Phi(x_T) = \lambda_v \mathcal{C}(v_T, v_{\max}) + \lambda_a \mathcal{C}(a_T, a_{\max}) \\ + \lambda_\omega \mathcal{C}(\omega_T, \omega_{\max}) + \lambda_\kappa \mathcal{C}(\kappa_T, \kappa_{\max})\end{aligned} \tag{9}$$

The complete optimization process is summarized in Algorithm 1, where guidance terms are iteratively updated to project trajectories onto the feasible set without retraining the flow model. Theoretically, the smooth nature of the constraint penalty guarantees stable gradient-based convergence, while penalizing the guidance effort acts as a minimal intervention that inherently preserves the learned human motion distribution. This inference-only procedure enables cross-robot generalization and practical deployment, producing safe and executable navigation behaviors.

Algorithm 1 Feasibility-Guided Inference

Given: Pre-trained flow matching model f^p
Inputs: Condition c; number of integration steps N; time steps $\{t_i\}_{i=0}^N$; optimization steps M; learning rate γ; momentum β
Initialize: Initial state $x_0 \sim \mathcal{N}(0, I)$; control inputs $u_{t_i} \leftarrow \mathbf{0}$
for $n = 1$ to M **do**
  Forward pass: Integrate trajectory with control:
  $$x_{t_{i+1}} = x_{t_i} + f^p(x_{t_i}, t_i, c) + u_{t_i}, \quad \forall i \in [0, N-1]$$
  Terminal gradient: $\mu_T \leftarrow \texttt{clip}(\nabla_{x_T} \Phi(x_T))$
  for $i = N-1$ to 0 **do** ▷ Adjoint backward pass
    $J(x_{t_i}) \leftarrow x_{t_i} + u_{t_i} + f^p(x_{t_i} + u_{t_i}, t_i, c)$
    $\mu_{t_i} \leftarrow \langle \mu_{t_{i+1}}, \nabla_{x_{t_i}} J(x_{t_i}) \rangle$
    $\mu_{t_i} \leftarrow \texttt{clip}(\mu_{t_i})$
  end for
  Control update: $u_{t_i} \leftarrow \beta \cdot u_{t_i} - \gamma \cdot \mu_{t_i}, \quad \forall i$
end for
Final rollout: Integrate x_{t_i} with optimized u_{t_i}
return Trajectory $\{x_{t_i}\}_{i=0}^N$

IV. Experiments

We evaluate Human2Nav with respect to effectiveness, data efficiency, and real-world deployability in simulation (a customized SocNavBench-like environment [30]) and on physical robots. Real-world platforms include: (1) Diablo: a two-wheel differential drive robot, (2) Unitree GO2: a quadruped robot with omnidirectional control. Our experiments answer the following three key questions:

- **Q1**: Can Human2Nav learn effective navigation policies from passive human videos, thereby achieving higher data efficiency and outperforming representative model-based and RL approaches?
- **Q2**: Can Human2Nav incorporate robot-specific motion constraints to ensure dynamically feasible trajectories?
- **Q3**: Can Human2Nav be directly transferred to real-world robotic platforms and still perform safe and efficient navigation in dynamic crowded environments?

Experiment Setup. We design four representative scenarios (Fig. 3): co-flow, where the robot follows, overtakes, or weaves through agents moving in similar directions; cross-flow, which requires traversing perpendicularly or diagonally through laterally moving pedestrians; counter-flow, involving navigation against incoming or static pedestrians; and multi-direction flow, where the robot must navigate through agents moving in diverse directions without any dominant flow pattern.

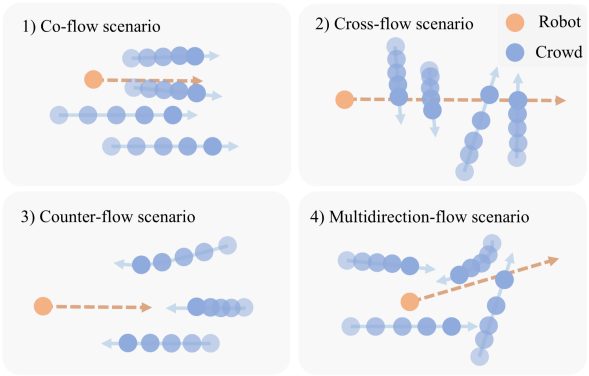


Fig. 3: Visualization of the four representative navigation scenarios: Co-Flow, Cross-Flow, Counter-Flow, and Multidirection-Flow.

In simulation experiments, each scenario contains 50 episodes sampled from held-out real-world pedestrian datasets, selected to cover diverse motion patterns and replayed in a nonreactive manner. Each episode lasts 5 seconds, with local crowd densities ranging from 1 to 15 agents.

In real-world experiments, each robot is subject to distinct kinematic and dynamic constraints, including maximum velocity, acceleration, turning radius, and angular velocity limits. These constraints are incorporated into Human2Nav via feasibility-guided inference to ensure dynamically feasible trajectory generation. All platforms share the same sensor suite (Livox Mid360 LiDAR, RealSense D435i camera,

Fig. 4: Visualization of navigation behaviors across four crowd scenarios. The top-left corner of each subfigure shows the corresponding BEV observation. The green star and blue star denote the start and goal positions, respectively. Humans are depicted in gray with historical trajectory points, and arrows indicate their current headings. The ego agent is shown in red, and the planned trajectory is highlighted in orange.

NVIDIA ORIN). Onboard perception uses CenterPoint 3D detection [31] with a LIO-based odometry system. Low-level trajectory execution is handled by PD controllers tuned for each platform. Real-world experiments are conducted in a controlled indoor environment, where human participants follow pre-defined directional trajectories corresponding to those used in simulation, ensuring consistency between virtual and physical interactions.

Baseline. To address **Q1**, we compare Human2Nav against two widely used baselines. 1) ORCA [32]: a model-based collision avoidance method implemented with the RVO2 library. 2) SARL [33]: a reinforcement learning approach that jointly models human–robot and human–human interactions. SARL uses a discrete action space of 80 actions (5 speeds and 16 headings) and is trained in two stages: pretraining for 50 epochs on 5,000 ORCA-generated demonstrations, followed by fine-tuning with 20,000 RL episodes in a circular environment using temporal-difference learning. This setup establishes a representative comparison between rule-based navigation and RL methods.

Metrics. We evaluate navigation performance using two primary metrics. The first is **Success Rate (SR)**, defined as the fraction of episodes in which the robot reaches the goal within the time budget and without collision; a failure in either aspect is counted as unsuccessful. To address **Q2**, we further introduce the **Trajectory Feasibility Score (TFS)**, which measures the proportion of planned waypoints that satisfy platform-specific dynamic constraints (e.g., velocity, acceleration, turning rate). We compute TFS over the full planning horizon and the executed trajectory segments, and report the minimum of the two to reflect both plan and execution.

V. Results and Discussion

A. Data Efficiency

Towards understanding **Q1**, we begin by qualitatively inspecting the navigation behaviors learned by Human2Nav. Figure 4 illustrates trajectory snapshots across four challenging crowd scenarios, accompanied by their BEV representations. These visualizations highlight Human2Nav's ability to generate trajectories that are smooth, proactive, and crowd-aware. Specifically, the policy anticipates pedestrian motions to exploit emerging gaps in the crowd, rather than reacting only to immediate collisions. This provides preliminary evidence that our framework effectively transfers nuanced human–human interaction patterns from video data into feasible and intelligent robot navigation behaviors.

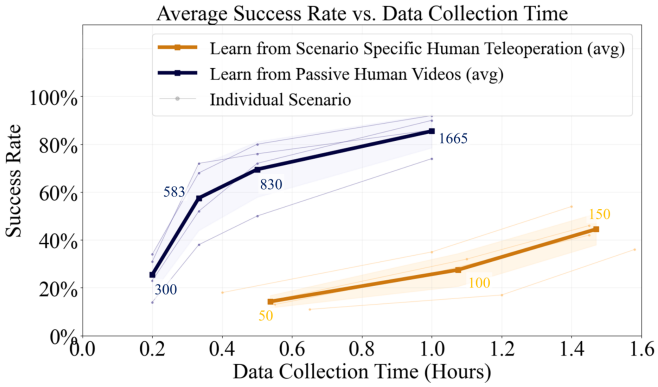


Fig. 5: Visualization of success rate in simulation versus data collection time from different sources: human videos (blue) and teleoperation (orange). Scenario-specific results are shown as faint lines, while cross-scenario averages appear as bold lines with error shading. Results show that learning from passive human videos scales more efficiently with data collection time, yielding more sample-efficient navigation policies and stronger cross-scenario generalization than teleoperation-based demonstrations.

We further evaluate the data efficiency and scalability of Human2Nav by measuring how policy performance varies with the amount of training data. To enable a direct comparison, we developed a simulated human teleoperation pipeline: at each planning step, the simulation is paused to allow a human operator to specify the next waypoint based on the robot's current observation. Notably, teleoperation data were collected in a scenario-specific manner (one policy per scenario), whereas the human videos constitute a passive, cross-scenario dataset. All models, including those trained on teleoperation data, used identical architecture and training settings. As depicted in Figure 5, the success rate of Human2Nav improves consistently as more passive human video demonstrations are incorporated. This scaling trend underscores the capability of our flow matching framework in distilling a coherent and generalizable navigation policy from large-scale, passively collected human videos. Importantly, Human2Nav achieves superior data efficiency without recourse to expensive, scenario-specific teleoperation, confirming its advantages in terms of both scalability and cross-scenario generalization.

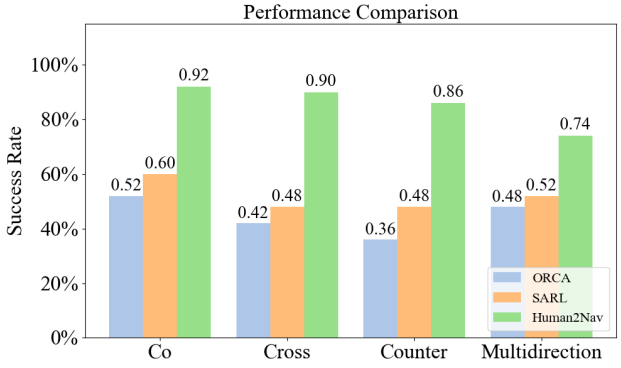


Fig. 6: Average closed-loop success rates over 50 episodes for four test scenarios, showing that Human2Nav consistently outperforms the baselines in crowd navigation.

To quantitatively compare against established baselines, we report closed-loop navigation performance across multiple scenarios in Figure 6. Human2Nav consistently outperforms both ORCA and SARL by a substantial margin. ORCA, as a classical reactive method, often exhibits conservative behavior such as freezing in dense crowds, which frequently results in timeout failures. Although SARL improves upon model-based techniques through learning, it remains constrained by its training regime: it relies on suboptimal ORCA-generated pretraining data and is fine-tuned in a simplified circular environment. Consequently, it fails to generalize to more realistic crowd settings. In contrast, Human2Nav leverages real human video data to capture rich social cues and predictive navigation strategies, resulting in more robust and human-like navigation performance.

B. Feasibility-Aware Guidance

With the proposed feasibility-guided inference, **Q2** aims to understand whether our method can effectively incorporate heterogeneous robot motion constraints to ensure dynamically feasible trajectories. We conduct an ablation study in simulation, evaluating TFS and SR metrics for three robotic platforms (differential-drive, Ackermann, and quadruped). Comparisons include three methods: our proposed Human2Nav with the feasibility guidance module added (Human2Nav w/ guidance), the same model without guidance (Human2Nav w/o guidance), and the ORCA baseline. Results in Fig. 7 indicate that ORCA attains a high no-collision rate but the lowest Trajectory Feasibility Score, as it does not account for kinematic constraints. This often leads to dynamically infeasible actions such as instantaneous turns or spinning in place, which limit its practical applicability. Both versions of Human2Nav maintain high success rates by leveraging navigation policies learned from human video data. However, the version without guidance frequently violates platform-specific constraints such as maximum velocity or turning rate. In contrast, adding the feasibility guidance module significantly improves trajectory feasibility, achieving nearly perfect TFS without reducing success rate and without requiring model retraining. These results confirm that our approach effectively incorporates motion constraints directly into the trajectory generation process, facilitating safe and practical deployment across different robot platforms.

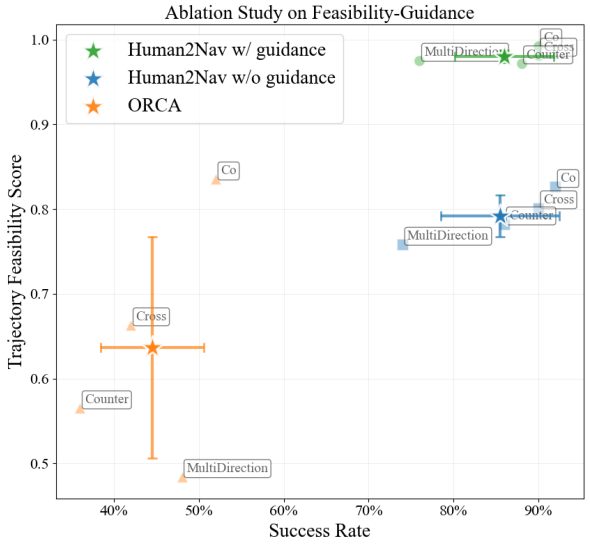


Fig. 7: Ablation study on feasibility-guided inference: Trajectory Feasibility Score (TFS) and Success Rate (SR) for Human2Nav (with/without guidance) and ORCA. Guidance significantly improves feasibility without compromising performance, enabling cross-platform deployment.

C. Real-World Evaluation

We explore the real-world deployment of Human2Nav across two heterogeneous robot platforms to understand **Q3**. Figure 8 presents navigation performance across four crowd scenarios. Despite their distinct kinematic and dynamic constraints, both the differential-drive mobile robot and the quadruped successfully executed smooth, collision-free, and goal-directed navigation behaviors. In the cross-flow scenario, the planner anticipated pedestrian trajectories and adjusted its speed to merge safely. In multi-direction flow, it identified and navigated through emerging gaps while maintaining smooth and feasible motion. The co-flow scenario shows Human2Nav's ability to maintain appropriate following distances, while in counter-flow it strategically decided when to yield or proceed based on pedestrian density. These behaviors emerged naturally from learning human navigation patterns, without explicit social rule encoding. These results affirmatively answer **Q3**, demonstrating Human2Nav's direct deployability across diverse platforms for safe and efficient navigation in dynamic crowds while respecting platform constraints.

VI. Limitations

Our method focuses on crowd navigation in open, obstacle-sparse spaces and does not explicitly model semantic elements such as traffic lights or pedestrian crossings. Future work will incorporate semantic scene cues as additional inputs or conditioning signals, and explore broader data sources such as YouTube or other public videos to extend human demonstration coverage.

Fig. 8: Real-world evaluation on two platforms (differential-drive and quadrupedal robots) across four crowd scenarios. Blue and red stars indicate start and goal positions. Green and yellow trajectories show robot and human motion, respectively. Each subfigure shows three temporal snapshots (left to right), with transparency indicating time progression (lighter: earlier). Bottom row displays LiDAR point cloud and planned trajectory (orange) in Rviz2.

VII. Conclusion

In this work, we proposed Human2Nav, a data-efficient framework that learns navigation policies directly from human videos, bypassing the need for costly robot-collected demonstrations. Our approach addresses the embodiment gap through a novel bird's-eye-view representation combined with test-time feasibility-guided flow matching. The introduced training-free feasibility guidance mechanism dynam-

ically incorporates kinematic constraints during inference, enabling practical deployment across heterogeneous robotic platforms without retraining. Extensive experiments demonstrate that Human2Nav achieves superior data efficiency and navigation performance compared to both model-based and learning-based baselines. The successful real-world deployment on diverse robots confirms our framework's capability for safe and efficient crowd navigation while maintaining strong generalization capabilities across different platforms.

REFERENCES

[1] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra et al., "Principles and guidelines for evaluating social robot navigation algorithms," ACM Transactions on Human-Robot Interaction, vol. 14, no. 2, pp. 1–65, 2025.

[2] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," The International Journal of Robotics Research, vol. 35, no. 11, pp. 1289–1307, 2016.

[3] 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 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 16 330–16 337.

[4] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 11 807–11 814, 2022.

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

[6] W. Wang, R. Wang, L. Mao, and B.-C. Min, "Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning," in 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023, pp. 11 348–11 355.

[7] Y. F. Chen, M. Liu, M. Everett, and J. P. How, "Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning," in 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017, pp. 285–292.

[8] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in 2009 IEEE 12th international conference on computer vision. IEEE, 2009, pp. 261–268.

[9] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer, 2016, pp. 549–565.

[10] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5030–5039.

[11] T. Yoon, D. Kang, S. Kim, J. Cheng, M. Ahn, S. Coros, and S. Choi, "Spatio-temporal motion retargeting for quadruped robots," IEEE Transactions on Robotics, 2025.

[12] G. Chen, M. Wang, T. Cui, Y. Mu, H. Lu, Z. Peng, M. Hu, T. Zhou, M. Fu, Y. Yang et al., "Fmimic: Foundation models are fine-grained action learners from human videos," arXiv preprint arXiv:2507.20622, 2025.

[13] T. Ma, J. Zheng, Z. Wang, Z. Gao, J. Zhou, and J. Liang, "Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation," arXiv preprint arXiv:2505.11865, 2025.

[14] H. Kim, J. Kang, H. Kang, M. Cho, S. J. Kim, and Y. Lee, "Uniskill: Imitating human videos via cross-embodiment skill representations," arXiv preprint arXiv:2505.08787, 2025.

[15] M. Hamandi, M. D'Arcy, and P. Fazli, "Deepmotion: Learning to navigate like humans," in 2019 28th IEEE international conference on robot and human interactive communication (RO-MAN). IEEE, 2019, pp. 1–7.

[16] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 8944–8951.

[17] S. Choi, M. K. Pan, and J. Kim, "Nonparametric motion retargeting for humanoid robots on shared latent space." in Robotics: science and systems, 2020.

[18] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," arXiv preprint arXiv:2410.11792, 2024.

[19] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," The International Journal of Robotics Research, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[20] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," Advances in neural information processing systems, vol. 32, 2019.

[21] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 63–70.

[22] B. Yang, H. Su, N. Gkanatsios, T.-W. Ke, A. Jain, J. Schneider, and K. Fragkiadaki, "Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following," arXiv preprint arXiv:2402.06559, 2024.

[23] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," arXiv preprint arXiv:2210.02747, 2022.

[24] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164

[25] S. Gode, A. Nayak, D. N. P. Oliveira, M. Krawez, C. Schmid, and W. Burgard, "Flownav: Combining flow matching and depth priors for efficient navigation," 2025. [Online]. Available: https://arxiv.org/abs/2411.09524

[26] X. Dai, Z. Yang, D. Yu, S. Zhang, H. Sadeghian, S. Haddadin, and S. Hirche, "Safe flow matching: Robot motion planning with control barrier functions," 2025. [Online]. Available: https://arxiv.org/abs/2504.08661

[27] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6569–6578.

[28] M. Du and S. Song, "Dynaguide: Steering diffusion polices with active dynamic guidance," arXiv preprint arXiv:2506.13922, 2025.

[29] L. Wang, C. Cheng, Y. Liao, Y. Qu, and G. Liu, "Training free guided flow matching with optimal control," 2025. [Online]. Available: https://arxiv.org/abs/2410.18070

[30] A. Biswas, A. Wang, G. Silvera, A. Steinfeld, and H. Admoni, "Socnavbench: A grounded simulation testing framework for evaluating social navigation," ACM Transactions on Human-Robot Interaction (THRI), vol. 11, no. 3, pp. 1–24, 2022.

[31] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11 784–11 793.

[32] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in 2008 IEEE international conference on robotics and automation. Ieee, 2008, pp. 1928–1935.

[33] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, "Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning," in 2019 international conference on robotics and automation (ICRA). IEEE, 2019, pp. 6015–6022.