# TopoUrbanBench: Group-Structural Compliance Evaluation for Urban Robot Navigation

**Haun Jo**[1,2*] **Hakbin Kim**[2] **Mooyeol Baek**[2] **Hwiyeon Yoo**[2] **Jaeuk Shin**[2]

**Chanmi You**[2] **Seunghyun Lee**[2] **Seong-heum Kim**[1] **Sunwook Choi**[2†]

[1]Department of Intelligent Semiconductors, Soongsil University, Seoul, Republic of Korea
[2]NAVER LABS, Seongnam, Republic of Korea

**Abstract:** Evaluation practice in social robot navigation rests largely on geometric metrics that treat pedestrians as independent obstacles, assuming social compliance is well-approximated by spatial separation from individuals. As robot navigation transitions to complex urban scenarios, the limits of this assumption become apparent: many pedestrians move within social formations rather than as independent individuals, so a trajectory that keeps safe distance from every individual may still bisect a queue or cut through a gathering's interactional core. To make this gap measurable, we present TopoUrbanBench, an evaluation framework for urban robot navigation. First, we define social structure as an intention-driven graph at generation time and measure trajectory intersection with this graph through three group-structural compliance metrics, one per archetype. Second, built on CARLA, we procedurally generate pedestrian formations anchored to urban semantic features under three archetypes—stationary gatherings, goal-anchored queues, and frame-mobile cohesive walking groups. Third, we apply this framework to evaluate vision foundation models and online RL agents, diagnosing where geometric proficiency and structural compliance diverge. Evaluations on a real-world social navigation dataset highlight the persistent challenges of group-structural compliance, exposing the limitations of current models in physical environments. The full generation pipeline is released to enable extension beyond our predefined scenarios. Code and data are available at: https://github.com/haunjo/topourbanbench.

## 1 Introduction

Social robot navigation has been evaluated primarily through geometric criteria: success rate (SR), collision rate (CR), and path efficiency (SPL) [1], with more recent work extending these to social-zone evaluation for person-to-person interactions [2]. This family of metrics has driven substantial progress in collision-avoidance algorithms and motion planners. However, as robot navigation moves toward dense and structurally complex urban environments, this reliance reveals a limitation: social acceptability cannot be inferred from each pedestrian's proximity alone.

This assumption, however, leaves a measurable gap in urban environments. A substantial share of urban pedestrians move within social groups rather than alone [3, 4], so the unit of social space to respect is often not each individual's personal buffer but the larger, invisible region a group occupies. Within such formations, individual separation and structural preservation diverge: a trajectory

*Work done during an internship at NAVER LABS
†Corresponding author

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

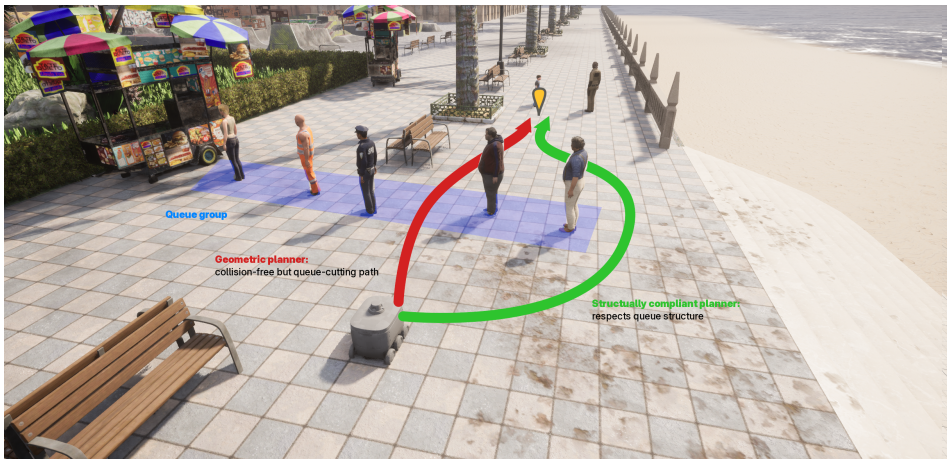



Figure 1: **Geometric sufficiency does not imply group-structural compliance.** A robot navigates toward a goal(yellow), negotiating a queue in front of a food truck in CARLA. A standard geometric planner (red path) successfully exploits a physical gap between pedestrians to produce a collision-free path. However, is a collision-free path always socially acceptable? While technically safe, this trajectory aggressively bisects the queue, disrupting the shared social structure. A structurally compliant trajectory (green path) recognizes the queue as an invisible boundary and routes around it. TopoUrbanBench introduces metrics like Queue cutting rate (QCR) to distinguish geometrically safe but socially disruptive behaviors (QCR: 1) from truly compliant navigation (QCR: 0), extending geometric evaluation with structural compliance metrics grounded in group behavioral intent.

may keep safe distance from every queue member while cutting through the line connecting them (Figure 1). We refer to this as the illusion of social compliance, and observe it systematically: when three vision foundation models [5, 6, 7] are fine-tuned on unstructured urban crowds, their geometric success rate improves while structural violations such as queue cutting persist or even worsen (Section 5).

Closing this gap requires knowing, during evaluation and training, which pedestrians form a group and what space they collectively occupy. This is fundamentally difficult to recover after the fact: group structure in urban crowds is defined by shared behavioral intent, yet post-hoc annotation can observe only geometry. Proximity-based recovery captures spatial co-presence but not the intent that distinguishes a queue from a passing cluster or a gathering from incidental adjacency [8, 9], and it further degrades when members briefly disperse or formations overlap [10, 11]. The alternative does not exist at scale for outdoor urban environments [9]. Without reliable group structure as ground truth, compliance cannot be precisely measured, and the training signal for improving it cannot be reliably derived.

To make this gap measurable, we present **TopoUrbanBench**, an evaluation framework that operationalizes group-structural compliance for urban robot navigation. We represent social structure as an intention-driven graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defined at generation time and measure compliance through three archetype-specific metrics — gathering disruption score (GDS), queue cutting rate (QCR), and cohesive proxemic intrusion (CPI). Built on CARLA [12], the same pipeline procedurally generates three urban formation archetypes anchored to semantic features and supports both offline imitation-learning and online reinforcement learning.

**Contributions.**

- **Topological group modeling and compliance metrics** (Section 3). We represent social structure as an intention-driven graph defined at generation time, and derive three archetype-specific metrics — GDS, QCR, and CPI — that measure trajectory intersection with group structure, extending geometric evaluation with precise compliance measurement beyond individual obstacle avoidance.

- **A procedural benchmark and training dataset** (Section 4). Built on CARLA, TopoUrbanBench procedurally synthesizes three urban formation archetypes (gatherings, queues, cohesive groups) anchored to semantic features, producing a large-scale IL dataset and an online RL environment with compliance reward signals.

- **Comprehensive baseline evaluation in simulation and a real-world case study** (Section 5). We evaluate three vision foundation models and an RL agent, revealing that geometric success and structural compliance diverge systematically — the *illusion of social compliance*. Open-loop qualitative analysis on SCAND [13] further suggests that compliance signals learned from our synthetic benchmark transfer to real sensor streams.

## 2 Related Work

### 2.1 Urban Social Navigation Simulators

Early work evaluated social navigation in 2D kinematic environments with reactive crowd models (SFM [14], ORCA [15]), enabling fast RL training but with limited visual fidelity [16, 17]. Photorealistic indoor simulators added visual realism via replayed or SFM-driven crowds in 3D scenes [18, 19, 20], but their narrow layouts miss the scale and structural diversity of outdoor navigation. Large-scale urban simulators — CARLA [12], MetaUrban [21], SimWorld [22] — expanded scene complexity for outdoor research, yet do not assign group membership or behavioral intent to their crowds, making group-structural compliance evaluation infeasible without additional annotation.

### 2.2 Group Modelling for Social Navigation

F-formation theory [23] laid the conceptual foundation for treating social groups as structured spatial entities, defining the interactional space through shared orientation, gaze, and movement rather than proximity alone, and establishing three nested regions: O-space, P-space, and R-space. Building on proxemic principles [24], a growing body of work has translated this theory into navigation systems that model social groups explicitly — through joint trajectory prediction, group-centric reward design, and boundary-reactive steering [10, 25, 26, 11, 27, 28, 29]. These approaches established that group-aware navigation is both meaningful and tractable, and motivate the present work.

At the metric level, however, existing approaches typically steer around a single radial buffer fitted to the group centroid [11] applying the same circular approximation regardless of formation type, or evaluate social zone compliance in small-scale indoor settings [2]. For group annotation, JRDB-Social [9] contributes rich multi-level annotations of social groups, offering a valuable resource for VLM-based social understanding. Comparable resources at urban street scale remain scarce, and group structure is typically recovered post-hoc through proximity clustering [10, 11] — capturing spatial co-presence but not the behavioral intent that differentiates structurally distinct formations.

### 2.3 Learning-Based Navigation Models

Learning-based navigation has increasingly adopted the paradigm of training end-to-end policies on large, diverse collections of expert demonstrations, enabling generalization across environments and robot platforms [30, 31]. ViNT [5] and NoMaD [6] extend this direction as vision foundation models trained through imitation learning on real-world datasets spanning indoor and outdoor environments; CityWalker [7] follows a similar paradigm using web-scale urban video data. Their pre-training mixtures include social navigation demonstrations such as SCAND [13], which implicitly contains group interactions, yet whether the resulting behavior is group-structurally compliant has not been examined.

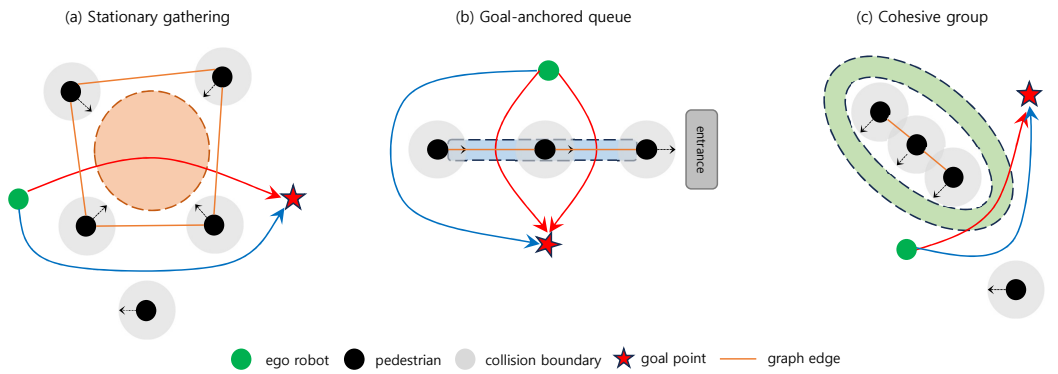



Figure 2: **Three urban formation archetypes and their compliance-relevant regions.** Each panel distinguishes two types of region: individual collision boundaries around each pedestrian (grey), and the group's structural primitive (coloured shaded region) that the robot's trajectory may violate without contacting any individual. **(a) Stationary gathering**: O-space (jointly attended convex region, orange). **(b) Goal-anchored queue**: P-space (line segments between consecutive members, blue), shown with the external entrance target. **(c) Cohesive group**: R-space (buffer co-moving with the group, green). Intention vectors (black dashed) define group membership; graph edges (orange) connect synchronized members. The red trajectory violates the structural primitive while staying clear of every collision boundary; the blue trajectory routes around both.

## 3 Topological Modeling of Human Social Groups

We define social structure at generation time as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ anchored to shared intent, and measure compliance through trajectory intersection with this graph.

### 3.1 Intention-Driven Group Anchoring

Each agent $v_i \in \mathcal{V}$ carries an intention vector $I_{v_i}$ encoding its behavioral target, and an intention-driven group $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ takes $\mathcal{V}$ as the agents sharing one intention with $\mathcal{E}$ connecting them ($e_{ij} \in \mathcal{E}$ iff $I_{v_i} = I_{v_j}$). The graph is fixed for the episode while its geometric realization deforms with member motion (Figure 2). This makes membership invariant to transient dispersion: $\mathcal{E}$ persists when members briefly separate, so crossings are flagged rather than missed. Edge intersection is also deterministic, serving as both evaluation criterion and reward signal.

### 3.2 Archetypes and Compliance Metrics

**From graph to spatial primitive.** Each formation is an intention-driven graph $G_g = (V_g, E_g)$, and compliance is measured by how the robot trajectory intersects its *geometric realization*. At time $t$, we map each graph $G_g$ to a spatial primitive $S_g(t) = \Phi(G_g, t) \subset \mathbb{R}^2$ and define a single per-formation violation indicator

$$v_g(t) = \mathbb{I}[\,\tau_R(t) \in S_g(t)\,], \tag{1}$$

where $\tau_R(t) \in \mathbb{R}^2$ is the robot's ground-plane position and $T$ the episode length. All three metrics are aggregations of $v_g$: gatherings and cohesive groups accumulate dwell time, while queues take a discrete count of formations violated at least once (pseudocode in Appendix C.2).

**Three urban formation archetypes.** F-formation theory [23] organizes a focused encounter into three nested regions — O-space, P-space, and R-space. We assign one urban archetype per region: stationary gatherings (O-space), goal-anchored queues (P-space), and frame-mobile cohesive groups (R-space), reporting the triple $(\mathrm{QCR}, \mathrm{GDS}, \mathrm{CPI})$.

**Gathering Disruption Score (GDS) — O-space.** For a stationary gathering, the edges $E_g$ connect members jointly attending an empty interactional core that they do not physically occupy; a robot entering this core disrupts the shared focal space even without contact. Here $\Phi$ realizes the O-space [23, 8] as a disk centred at the members' centroid $c_m = \frac{1}{|V_m|}\sum_{i\in V_m} p_i$ with radius $r_m = \min_{i\in V_m}\|p_i - c_m\|$, capturing the region enclosed by the members rather than the area they occupy. GDS is the dwell time inside any of the $M$ gatherings:

$$\mathrm{GDS} = \int_0^T \max_{m\in[M]} v_m(t)\, dt. \tag{2}$$

**Queue Cutting Rate (QCR) — P-space.** For a queue (formed relative to an external target such as a transit stop), $\Phi$ realizes the P-space directly as the chain of line segments joining consecutive members. Crossing any segment inserts the robot into the queue's ordering, a structural violation. Since one crossing already breaks the formation, QCR counts how many of the $K$ queues are cut at least once:

$$\mathrm{QCR} = \sum_{k\in[K]} \mathbb{I}[\,\exists\, t \in [0,T] : \, v_k(t) = 1\,] \;\in\; \{0,\dots,K\}. \tag{3}$$

A crossing is detected per edge: $v_k(t) = 1$ when $\tau_R$ changes side relative to a member-to-member segment, i.e. the sign of its cross product with that segment flips between consecutive timesteps.

**Cohesive Proxemic Intrusion (CPI) — R-space.** For a cohesive walking group (whose primitive translates with its motion), $\Phi$ realizes the R-space as a buffer of width $D_{\text{buffer}}$=1.2 m immediately outside the convex hull of members (their P-space), co-moving with the group. This width corresponds to the boundary between Hall's personal and social distance zones [24] and a standard personal-space threshold in social robot navigation [32]. Entering the buffer breaks the group's shared movement frame even without contact. CPI is the dwell time inside the R-space of any cohesive group:

$$\mathrm{CPI} = \int_0^T \max_{c} v_c(t)\, dt. \tag{4}$$

## 4 The TopoUrbanBench Framework

**Map data extraction.** The pipeline takes CARLA's multi-modal map data (RGB BEV, semantic segmentation, walkable region) as input and identifies functional locations — food carts, bus stops, plazas — as semantic anchors where urban formations naturally arise (Figure 3(a)).

**Semantic anchoring and group placement.** Urban formations are spatially tied to function: queues form at service points, gatherings at focal objects. At each anchor, the pipeline instantiates the archetype-specific graph $G_g$ (Section 3.2) with membership and intent fixed before trajectory synthesis. For each scene, one to three expert ego-robot paths are selected whose unmodified routes cross at least one formation boundary, creating the structural dilemma the benchmark is designed to measure.

**Dynamic crowd and expert trajectory.** Group-assigned agents follow archetype-specific formation constraints (Appendix B.1), while unaffiliated pedestrians are synthesized independently via CrowdES [33], a data-driven crowd engine trained on real pedestrian datasets (ETH/UCY [34, 35], SDD [36], EDIN [37]). The result is a dynamic crowd layer in which structured formations and naturalistic individual flows coexist (Figure 3(b)). Expert ego-robot paths are then planned with A* on the walkable map and corrected per archetype to bypass the formations they would otherwise cross, without re-simulating the crowd (Appendix B.2). These paths are replayed in CARLA to collect synchronized RGB observations and ego-pose at 4 Hz for Track 1 imitation learning (Figure 3(c)), while the same group graph $\mathcal{G}$ serves as a per-step reward signal for Track 2 online learning.

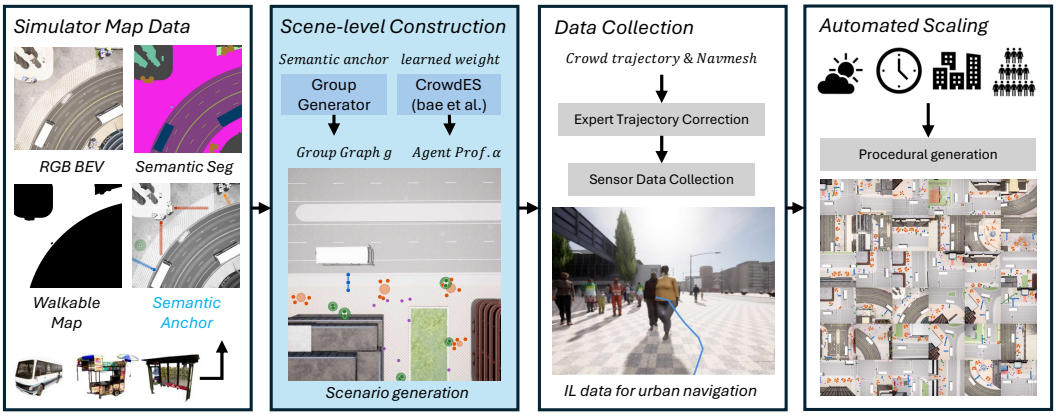



Figure 3: **Overview of the TopoUrbanBench generation pipeline. (a)** Map data extracted from CARLA, including semantic anchor candidates. **(b)** A generated scene fusing intention-driven group graphs with CrowdES [33] individual pedestrians. **(c)** Collected IL data: expert-corrected trajectories with synchronized sensor recordings. **(d)** Dataset scaled procedurally across towns, weather, and times of day.

Table 1: Offline imitation-learning dataset scale. Both subsets share the same base scenes (≈165 across Town03/05/10) with 5-condition weather augmentation; the Topo-FT subset additionally contains structured group formations. Q/G/C denote Queue, Gathering, and Cohesive walking formations. Each scene records synchronized front-facing RGB observations (640×480, 4 Hz) and ego-pose annotations.

| Subset | Episodes | Total Duration | Avg. Expert Path (m) | Avg. Groups/Scene (Q / G / C) |
|---|---|---|---|---|
| Standard-FT | 574 | 6.4 h | $37.7 \pm 13.0$ | 0 |
| Topo-FT | 566 | 6.4 h | $37.3 \pm 14.7$ | 8.2 (1.0 / 2.0 / 5.2) |

**Benchmark scale.** Applied across Town03, Town05, and Town10 under diverse weather conditions (Figure 3(d)), the pipeline produces an offline IL dataset for Track 1 (Table 1) and a 100-episode closed-loop evaluation suite shared by both tracks.

## 5 Validation and Baselines

We evaluate TopoUrbanBench along two tracks: *Track 1* validates the offline dataset by fine-tuning vision foundation models, and *Track 2* validates the online environment by training an RL agent with our compliance metrics as reward signals.

### 5.1 Experimental Setup and Baselines

We evaluate across three diverse CARLA maps (Town03, Town05, Town10), generating 100 evaluation episodes capped at 2 minutes each (Appendix B.5). These scenarios are generated separately from the training scenes: although both use the same towns, crowd and group placements and start–goal pairs differ.

**Track 1: Validating Offline Dataset Utility (Vision Foundation Models).** We evaluate **ViNT** [5], **NoMaD** [6], and **CityWalker** [7] under three regimes: *Zero-Shot* (pre-trained weights), *Standard-FT* (IL on unstructured CARLA crowds), and *Topo-FT* (IL on our topological dataset).[3]

[3]We additionally report off-sidewalk rate (OSR), the fraction of episode duration spent outside CARLA's walkable region $\mathcal{W}$, as a validity check: routing around formations through non-walkable areas yields trivially low compliance, so results with elevated OSR are invalid for comparison. Table entries aggregate OSR over all episodes; episode-level filtering (OSR $> 0.5$) is reported in Appendix D.2.

**Track 2: Validating Online Environment Utility (Reinforcement Learning).** To show our topological metrics can serve as training signals, we train a continuous-control **Vanilla PPO** [38] agent, comparing *Standard-PPO* (geometric collision/goal rewards only) against *Topo-PPO* (with our GDS, QCR, and CPI added as negative reward signals).

### 5.2 Evaluating Vision Foundation Models (Track 1)

Table 2 shows that geometric performance and structural compliance do not move together. Standard-FT lifts SR far above Zero-Shot across all three models, yet structural violations stay high or even rise (NoMaD QCR and GDS both increase). Under Topo-FT the geometric metrics then diverge: NoMaD improves SR, CityWalker trades SR for compliance, and ViNT lowers QCR and GDS as CR rises. No single geometric score predicts compliance — the gap conventional evaluation cannot surface, and the reason a dedicated compliance axis is needed. The Zero-Shot rows confirm this from the other side: their low violations are artifacts of failure (freezing or sidewalk-leaving; SR as low as 6.9%, OSR up to 32%), meaningful only alongside SR and OSR.

Table 2: Track 1. Evaluation of Vision Foundation Models. Geometric metrics (SR, SPL, CR) and topological metrics (QCR, GDS, CPI) do not move together: Topo-FT mostly reduces structural violations relative to Standard-FT, but the accompanying changes in SR and CR differ across models.

| | | Geometric Metrics | | | | Topological Metrics (Ours) | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Training Regime** | **SR (%) ↑** | **SPL ↑** | **CR (%) ↓** | **OSR (%) ↓** | **QCR ↓** | **GDS (s) ↓** | **CPI (s) ↓** |
| **ViNT** [5] | Zero-Shot | 15.0 | 0.14 | **35.6** | 32.1 | 0.52 | 5.0 | **4.8** |
| | Standard-FT | 78.2 | 0.72 | 37.9 | 2.0 | 0.62 | 5.2 | 6.6 |
| | **Topo-FT (Ours)** | **78.3** | **0.75** | 41.7 | **0.0** | **0.48** | **4.9** | 8.3 |
| **NoMaD** [6] | Zero-Shot | 6.9 | 0.06 | **34.5** | 19.5 | **0.33** | **3.1** | **4.5** |
| | Standard-FT | 77.0 | 0.50 | 58.6 | **0.0** | 0.82 | 6.7 | 11.1 |
| | **Topo-FT (Ours)** | **87.3** | **0.62** | 57.5 | **0.0** | 0.68 | 4.0 | 8.5 |
| **CityWalker** [7] | Zero-Shot | 58.6 | 0.55 | **34.5** | **8.0** | 0.77 | 4.1 | 8.2 |
| | Standard-FT | **79.3** | **0.75** | 42.5 | 9.2 | 0.84 | 4.0 | **7.6** |
| | **Topo-FT (Ours)** | 73.6 | 0.70 | 37.9 | 9.2 | **0.68** | **3.0** | 8.4 |

### 5.3 Validating the Environment with Reinforcement Learning (Track 2)

Table 3 confirms that our metrics function as effective reward signals: Topo-PPO sharply reduces GDS and CPI at no SR cost. QCR, however, remains unchanged across both variants, suggesting that discrete queue-cutting events pose a harder credit-assignment problem for RL than the continuous accumulation of GDS and CPI.

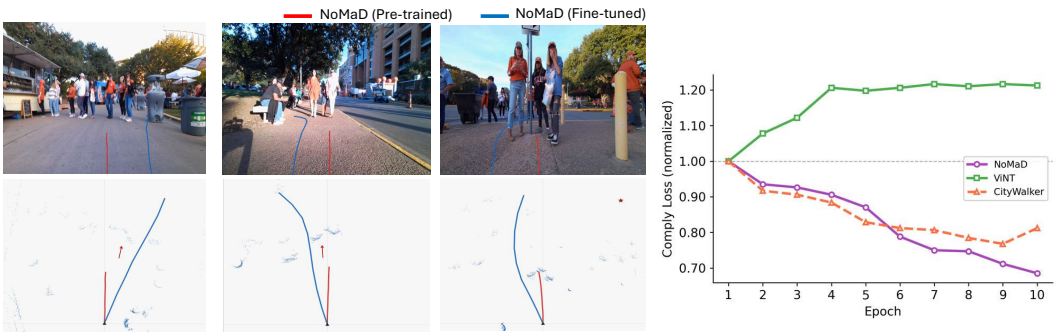
Table 3: Track 2. Online Reinforcement Learning. Topo-PPO reduces continuous violations (GDS, CPI) at no SR cost, while QCR remains unchanged — a credit-assignment difficulty discussed in Section 5.3.

| | Geometric Metrics | | | | Topological Metrics (Ours) | | |
|---|---|---|---|---|---|---|---|
| **Method** | **SR (%) ↑** | **SPL ↑** | **CR (%) ↓** | **OSR (%) ↓** | **QCR ↓** | **GDS (s) ↓** | **CPI (s) ↓** |
| Standard-PPO | **66.7** | 0.62 | 18.5 | 18.5 | **0.22** | 2.6 | 6.2 |
| **Topo-PPO (Ours)** | **66.7** | **0.64** | **14.9** | **14.8** | **0.22** | **1.5** | **3.8** |

### 5.4 Qualitative Sim-to-Real Case Study on SCAND

As a qualitative open-loop sim-to-real case study, we analyze NoMaD predictions on SCAND [13] — a dataset heavily represented in the GNM [30] pre-training mixture, so the model is already visually acclimated to the evaluation environments.

**Compliance learning across architectures.** To examine how well each architecture can learn a group-structural signal, we augment each baseline's MSE imitation objective with a compliance loss




(a) Group Avoidance Behavior on SCAND

(b) Comply Loss During Fine-Tuning

Figure 4: **Sim-to-real case study on SCAND. (a) Group avoidance on SCAND frames**: comparison of pretrained NoMaD and Topo-FT NoMaD (with the compliance loss) predictions, shown in first-person view (top) and bird's-eye view (bottom). **(b) Compliance loss when added to Topo-FT fine-tuning (normalized to epoch 1)**: ViNT fails to learn the objective, CityWalker partially converges, and NoMaD reduces the loss consistently.

that penalizes waypoints intersecting formation primitives. Figure 4(b) compares this loss across architectures. ViNT (regression) fails to learn the objective, with its loss rising and plateauing above the initial value; CityWalker (regression) partially converges; only NoMaD (diffusion) reduces the loss consistently. One plausible explanation is the multi-modal nature of group avoidance: a single formation typically admits several valid bypass paths (left or right around a queue, either side of a gathering), and the compliance loss penalizes intersection with the spatial primitive without prescribing which side to pass. Regression policies, which map an observation to a single waypoint sequence, may average over these modes — often producing a compromise trajectory that splits the difference and intersects the primitive. Diffusion policies sample from a learned distribution over trajectories, so different modes can be represented as distinct samples rather than collapsed into a mean. This is consistent with group-structural compliance being a multi-modal objective that distributional policies may represent more naturally; since the three models also differ in pre-training and action representation, a controlled comparison is needed to isolate the role of distributional output. We therefore focus the open-loop SCAND evaluation on NoMaD.

**Group avoidance on real-world scenes.** Figure 4(a) shows NoMaD predictions on three SCAND frames near the food-truck queue, in both first-person and bird's-eye views. The pretrained model, despite its extensive visual prior on SCAND, routinely generates trajectories that head into or through the group. After Topo-FT in CARLA with the compliance loss — with the visual encoder frozen so only the action decoder adapts — the same model routes around the group. This suggests that the structural signal learned in our synthetic benchmark is decoder-level and transfers to real sensor streams.

## 6 Conclusion and Future Work

TopoUrbanBench operationalizes group-structural compliance through intention-driven graphs and three archetype-specific metrics, distinguishing geometric proficiency from compliance and supporting both diagnostic evaluation and online policy learning. Across our experiments, geometric and compliance metrics consistently fail to predict each other — the gap conventional evaluation cannot surface — and our qualitative SCAND case study suggests that compliance learned in synthetic environments can transfer to real sensor streams at the decoder level, most clearly for the diffusion-based policy in our comparison. We release the full pipeline to support extension beyond our predefined scenarios.

**Limitations and Future Work.** We organize the limitations by our three contributions. *Group modeling and metrics.* The three archetypes cover common formations but not composite ones, and the spatial primitives assume proxemic distances that vary across cultures [24]. The generation pipeline already allows these distances and group densities to be adjusted, and a new archetype can be added by defining its edge topology, spatial primitive, and correction rule. *Benchmark design.* TopoUrbanBench is designed to test where group-structural compliance fails, not to reproduce real urban crowds; the placement, density, and dynamics of its groups are not matched to real-world data. Pedestrians follow pre-generated trajectories and do not react to the ego robot, so our metrics measure intrusion into a given group structure rather than disruption of a live social interaction; adding pedestrians that react to the robot would allow such interactions to be evaluated. Rendering and pedestrian animation are also simplified; animating group interactions (e.g., members facing and gesturing to each other) and ablating these cues would be an interesting direction, revealing how much they help models respect group structure. *Evaluation.* The Standard-FT versus Topo-FT comparison is not fully controlled: both subsets share base scenes, crowd density, and the MSE imitation objective, but adding group formations also changes the appearance and difficulty of the scenes, so the gains cannot be attributed to group-detouring supervision alone. An ablation that fixes the Topo-FT scenes and varies only expert path correction would separate this effect. Our architecture comparison suggests that policies without distributional output may have more difficulty learning compliance, but this remains to be confirmed by a controlled study; building social structure into the architecture as an explicit constraint is one possible direction. Finally, our SCAND case study (Section 5.4) is qualitative and open-loop; quantitative evaluation with group annotations on SCAND and closed-loop tests on a physical robot are left for future work.

**Acknowledgments**

This work was supported by the G-LAMP Program of the NRF grant funded by the MOE (No. RS-2025-25441317), and by the MSIT and NIPA through the Advanced GPU Utilization Support Program (02-26-01-0499), Republic of Korea.

## References

[1] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[2] Z. Fang, A. Xiao, D. Hsu, and G. H. Lee. From obstacles to etiquette: Robot social navigation with vlm-informed path selection. *IEEE Robotics and Automation Letters*, 11(4):3947–3954, 2026. doi:10.1109/LRA.2026.3662586.

[3] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[4] A. Nicolas and F. H. Hassan. Social groups in pedestrian crowds: Review of their influence on the dynamics and their modelling. *arXiv preprint arXiv:2107.13293*, 2021.

[5] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. Vint: A foundation model for visual navigation. In *Conference on Robot Learning (CoRL)*, 2023.

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

[7] X. Liu, J. Li, Y. Jiang, N. Sujay, Z. Yang, J. Zhang, J. Abanes, J. Zhang, and C. Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6875–6885, 2025.

[8] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PLoS ONE*, 10(5):e0123783, 2015.

[9] S. Jahangard, Z. Cai, S. Wen, and H. Rezatofighi. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22087–22097, 2024.

[10] R. Bhaskara, M. Chiu, and A. Bera. Sg-lstm: Social group lstm for robot navigation through dense crowds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3835–3840, 2023.

[11] U. K. Roy and S. Rahman. Taga: A tangent-based reactive approach for socially compliant robot navigation around human groups. *arXiv preprint arXiv:2503.21168*, 2025.

[12] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, pages 1–16, 2017.

[13] 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*, 7(4):11807–11814, 2022.

[14] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51 (5):4282, 1995.

[15] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha. Reciprocal n-body collision avoidance. In *Robotics research: the 14th international symposium ISRR*, pages 3–19, 2011.

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

[17] 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)*, pages 6015–6022, 2019.

[18] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal. Thör: Human-robot navigation data collection and accurate motion trajectories dataset. *IEEE Robotics and Automation Letters*, 5(2):676–682, 2020.

[19] 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)*, 11(3):1–24, 2022.

[20] N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters*, 7(4):11047–11054, 2022.

[21] W. Wu, H. He, J. He, Y. Wang, C. Duan, Z. Liu, Q. Li, and B. Zhou. Metaurban: An embodied ai simulation platform for urban micromobility. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[22] X. Ye, J. Ren, Y. Zhuang, X. He, Y. Liang, Y. Yang, M. Dogra, X. Zhong, E. Liu, K. Benavente, R. M. Nagaraju, D. V. Sharma, Z. Ma, T. Shu, Z. Hu, and L. Qin. Simworld: An open-ended simulator for agents in physical and social worlds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[23] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[24] E. T. Hall. *The hidden dimension*. Doubleday, 1966.

[25] N. T. Do, T. D. Pham, N. H. Son, T. D. Ngo, and X. T. Truong. Deep reinforcement learning based socially aware mobile robot navigation framework. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 226–231, 2020. doi:10.1109/NICS51282.2020.9335911.

[26] K. Katyal, Y. Gao, J. Markowitz, S. Pohland, C. Rivera, I.-J. Wang, and C.-M. Huang. Learning a group-aware policy for robot navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11328–11335, 2022.

[27] V. Schmuck and O. Çeliktutan. Sang: Socially aware navigation between groups. *International Journal of Social Robotics*, 17(6):989–1002, 2025.

[28] A. Wang, C. Mavrogiannis, and A. Steinfeld. Group-based motion prediction for navigation in crowded environments. In *Conference on Robot Learning (CoRL)*, pages 871–882, 2021.

[29] X. Lu, A. Faragasso, Y. Wang, A. Yamashita, and H. Asama. Group-aware robot navigation in crowds using spatio-temporal graph attention network with deep reinforcement learning. *IEEE Robotics and Automation Letters*, 10(4):4140–4147, 2025. doi:10.1109/LRA.2025.3549663.

[30] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. Gnm: A general navigation model to drive any robot. In *International Conference on Robotics and Automation (ICRA)*, 2023. URL https://arxiv.org/abs/2210.03370.

[31] R. Alyassi, C. Cadena, R. Riener, and D. Paez-Granados. Social robot navigation: a review and benchmarking of learning-based methods. *Frontiers in Robotics and AI*, 12:1658643, 2025.

[32] 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 (THRI)*, 14(2):1–65, 2025.

[33] I. Bae, J. Lee, and H.-G. Jeon. Continuous locomotive crowd behavior generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[34] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE International Conference on Computer Vision (ICCV)*, pages 261–268, 2009.

[35] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3):655–664, 2007.

[36] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision (ECCV)*, 2016.

[37] B. Majecka. Statistical models of pedestrian behaviour in the forum, 2009. M.Sc. thesis, School of Informatics, University of Edinburgh.

[38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

# A Comparison of Robot Navigation Simulators

Table 4 situates TopoUrbanBench within the broader landscape of robot navigation simulators along three axes relevant to group-aware evaluation.

Table 4: Comparison of robot navigation simulators. *Group Behavior*: group social behavior (formation, cohesion, role) is explicitly modeled and simulatable, not merely emergent from individual-agent clustering. *Group Annotation*: group membership is available as ground-truth annotation rather than inferred from proximity. *Compliance Metrics*: evaluation extends beyond individual-distance measures.

| Simulator | Rendering | Setting | Crowd Engine | Group Behavior | Group Annotation | Compliance Metrics |
|---|---|---|---|---|---|---|
| *Indoor and small-scale settings* | | | | | | |
| CrowdNav [17] | 2D | Abstract | ORCA / SFM | × | × | Individual |
| SocNavBench [19] | 3D | Indoor | Real Recorded | × | × | Individual |
| SEAN 2.0 [20] | 3D | Indoor | SFM | ✓ | × | Individual |
| *Urban outdoor settings* | | | | | | |
| MetaUrban [21] | 3D | Urban | SFM / Rule-based | × | × | Individual |
| SimWorld [22] | 3D | Urban | Rule-based | × | × | Individual |
| **TopoUrbanBench (Ours)** | **3D** | **Urban** | **Data-driven (CrowdES)** | ✓ | ✓ | **Structural** |

Prior platforms drive crowd dynamics with physics-based reactive models (SFM, ORCA) and lack generation-time group membership; TopoUrbanBench fills both gaps via CrowdES [33].

# B Detailed Specifications of the TopoUrbanBench

TopoUrbanBench is structurally divided into two components: an offline imitation-learning dataset for Track 1 and an online closed-loop environment for Track 2. Table 5 details the offline dataset statistics, Table 7 the closed-loop evaluation suite and the ego-robot sensor configuration.

Table 5: Offline imitation-learning dataset statistics (4 Hz, ≈165 base scenes across Town03/05/10, 5-condition weather-augmented). Q/G/C denote Queue, Gathering, and Cohesive walking formations.

| Subset | Episodes | Total | Avg. Duration (s) | Avg. Length (m) | Groups/Scene (Q/G/C) |
|---|---|---|---|---|---|
| Standard-FT | 574 | 6.4 h | $40.5 \pm 13.6$ | $37.7 \pm 13.0$ | 0 |
| Topo-FT | 566 | 6.4 h | $40.8 \pm 15.6$ | $37.3 \pm 14.7$ | 8.2 (1.0 / 2.0 / 5.2) |

## B.1 Generation Pipeline Architecture

Figure 5 illustrates the three-layer architecture of the TopoUrbanBench generation pipeline.

**Layer 1: Semantic Extraction.** The pipeline begins by processing CARLA's semantic segmentation into spatial maps and semantic anchors. The *Walkable Map* identifies navigable sidewalk regions and serves as the navigation mesh for downstream trajectory pathfinding. The *semantic anchors*—bus stops and food carts for queues, open plazas and building facades for gatherings, and unobstructed sidewalk corridors for cohesive groups—constrain where each archetype may be instantiated.

**Layer 2: Group Graph Construction.** Given the semantic anchors, the *Formation Allocator* determines the archetype, size, and anchor location for each group. The *Group Graph Builder* assigns agents from the crowd pool to each formation and constructs the archetype-specific edge topology: a directed path $P_n$ for queues, a cycle $C_n$ for gatherings, and a complete graph $K_n$ for cohesive groups. Each group is tagged with archetype metadata—goal direction for queues, focus center for gatherings, shared motion direction for cohesive groups—fixed at generation time and used by the compliance metrics throughout the episode. The resulting graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encodes group membership and behavioral intent with behavior state $\mathcal{S} \in \{\text{queue}, \text{gathering}, \text{cohesive}\}$ per group.

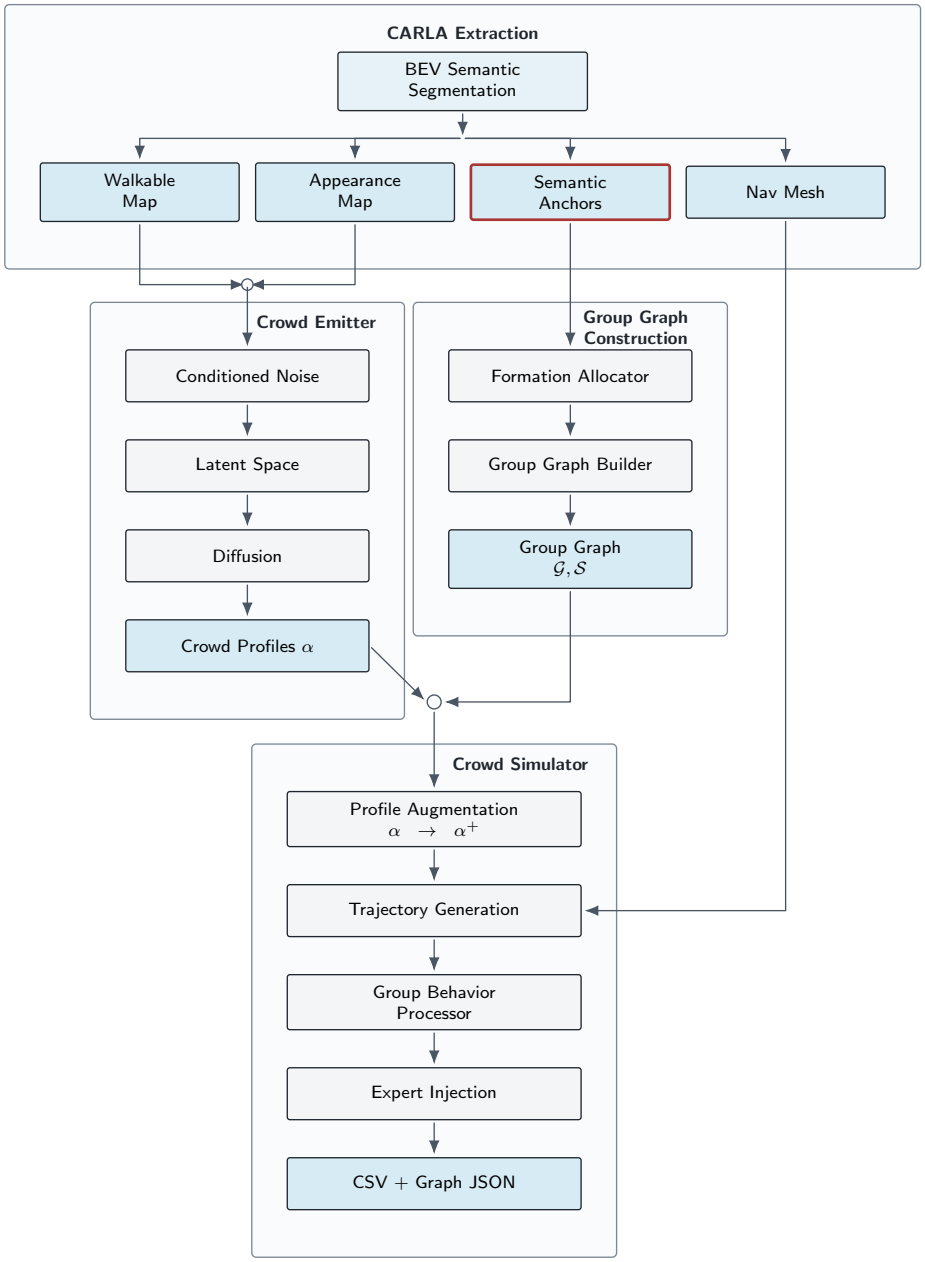



Figure 5: System architecture of TopoUrbanBench. **CARLA Extraction**: semantic segmentation produces a Walkable Map, an Appearance Map, and Semantic Anchors (bus stops, food carts, plazas); the Nav Mesh is derived in parallel. **Crowd Emitter**: a conditioned diffusion model generates kinematic profiles $\alpha = \{\kappa, v, T_s, T_d, c_s, c_d\}$ per agent. **Group Graph Construction**: the Formation Allocator assigns agents to archetype-specific formations near semantic anchors; the Group Graph Builder instantiates edge topologies ($P_n$, $C_n$, $K_n$) to produce the group graph $\mathcal{G}$. **Crowd Simulator**: profiles are augmented to $\alpha^+$ under topological constraints, trajectories are synthesized with a Group Behavior Processor enforcing formation structure, expert trajectories are injected, and scenes are exported as CSV with group graph JSON.

**Layer 3: Trajectory Synthesis and Augmentation.** The *Crowd Emitter* (CrowdES [33]) generates a baseline kinematic profile for each agent via a conditioned diffusion process. Each profile is defined as $\alpha = \{\kappa, v, T_s, T_d, c_s, c_d\}$, where $\kappa$ is the agent type, $v$ the preferred walking speed, $T_s, T_d$ the spawn and end times, and $c_s, c_d$ the spawn and goal coordinates. For group-assigned agents, the profile is augmented to $\alpha^+$ by fusing topological constraints from $\mathcal{G}$. Unaffiliated agents retain their original profiles $\alpha$ and move independently.

### B.2 Expert Trajectory Generation Pipeline

Each generated crowd scene receives one to three compliant expert ego-robot trajectories, generated post-hoc on the recorded crowd without re-simulating it.

**Candidate selection.** A non-group agent whose path traverses the scene is selected, preferring agents that spawn near a formation and would violate it along their original route.

**A* planning and correction.** The path is planned with A* on the walkable map extracted from CARLA's semantic segmentation (Appendix B.1), treating static formations (queues and gatherings) as high-cost regions. Residual violations are then corrected per archetype: queue crossings are rerouted around the tail of the queue, gathering intrusions are bypassed around the outside of the O-space, and intrusions into cohesive groups, which move and cannot be avoided at planning time, are pushed out by repulsion from the group.

**Smoothing and replay.** The corrected trajectory is smoothed under kinematic constraints and replayed in CARLA to collect synchronized sensor observations at 4 Hz. For Standard-FT, which contains no group graph, the same A*-based procedure is used while avoiding dense crowd regions.

### B.3 Group Generation Examples

**Semantic anchors as spatial context.** Queues form directed linear chains extending away from a service point (bus stop, food cart), with members facing toward the anchor. Gatherings are placed in detected open areas, with each member oriented toward the group centroid.

**Anchor types across towns.** Bus stop anchors (`vehicle.mitsubishi.fusorosa`, `static.prop.busstoplb`) are present in all three towns. Food-cart anchors (`SM_Cartel_Add`, `BP_FoodCart`) appear exclusively in Town10. Agent appearance (clothing, body shape) is randomized per instance.

### B.4 Training Protocols

**Track 1: Vision Foundation Model Fine-Tuning.** We fine-tune ViNT [5], NoMaD [6], and CityWalker [7] on the Standard-FT and Topo-FT subsets (Table 5) using imitation learning with MSE loss between predicted and expert waypoints. All models are trained with AdamW and a cosine annealing schedule; hyperparameters are listed in Table 6. The observation encoder is frozen for all models; CityWalker updates only its attention decoder, while ViNT and NoMaD fine-tune the remaining parameters end-to-end.

Table 6: Track 1 fine-tuning hyperparameters.

| | **CityWalker** | **ViNT** | **NoMaD** |
|---|---|---|---|
| Learning rate | $5\times10^{-5}$ | $1\times10^{-4}$ | $1\times10^{-4}$ |
| Batch size | 16 | 64 | 64 |
| Epochs | 10 | 10 | 10 |
| Frozen params | obs. encoder | obs. encoder | full backbone |

**GPS Adaptation for Point-Goal Navigation.** ViNT and NoMaD are *image-goal* models. For point-goal evaluation, we follow the GPS adaptation of [5]: the goal image encoder is replaced by a




Figure 6: **Group generation examples rendered in CARLA.** Each panel shows one social group instance (queue or gathering) captured from a free-viewpoint camera 8 m above the group centroid. Queue members face toward the service anchor (bus stop or food cart); gathering members orient toward the group centroid. Group size and spatial extent vary across instances; agent appearance and scene geometry differ by town and random seed.

lightweight two-layer MLP that takes a normalized goal displacement vector as input and produces the model's goal token. The GPS MLP is trained in two stages—first on Standard-FT data, then continued on Topo-FT—with the observation encoder kept frozen throughout. The Standard-FT warm start provides an established goal-following policy before compliance is learned on Topo-FT data; all other parameters are re-initialized from the pretrained backbone in both regimes.

**Track 2: PPO Training.** We train Vanilla PPO [38] over a continuous action space of linear and angular velocity. Standard-PPO uses a goal-reaching reward ($+100$), a collision penalty ($-5$), a step penalty ($-0.01$), and a shaped progress reward. Topo-PPO adds three formation-specific penalty terms:

$$\begin{aligned} r_{\text{topo}} = &- \lambda_{\text{Q}} \cdot \mathbb{I}[\text{queue P-space crossed}] \\ &- \lambda_{\text{G}} \cdot \mathbb{I}[\text{inside gathering O-space}] \\ &- \lambda_{\text{C}} \cdot \mathbb{I}[\text{inside cohesive R-space}], \end{aligned}$$

with $\lambda_{\text{Q}} = \lambda_{\text{G}} = \lambda_{\text{C}} = 0.3$. Training follows a three-phase curriculum over 2M steps: empty scenes (500k), unstructured crowds (500k), and social formations (1M).

Table 7: Closed-loop evaluation suite specification and ego-robot sensor configuration. **(a)** Scene-level statistics (20 Hz; towns T03 / T05 / T10: 33 / 31 / 36 scenarios; Expert Path = compliant trajectory length used as SPL denominator). **(b)** Group formation breakdown; Size = members per group instance. **(c)** Ego-robot sensors; mount offset in metres $(x,\ y,\ z)$ relative to base link.

| *Evaluation Suite* | | | |
|---|---|---|---|
| Scenarios | 100 | | |
| Max Horizon | 120 s | | |
| Pedestrians / Scene | $60 \pm 16$ | | |
| Expert Path (m) | $46.1 \pm 8.8$ | | |
| Groups / Scene | $9.3 \pm 3.3$ | | |
| *Group Formation Breakdown* | | | |
| Archetype | Groups / Scene | Avg. Group Size | Size Range |
| Queue | $1.6 \pm 1.7$ | $5.4 \pm 1.8$ | 2–7 |
| Gathering | $2.1 \pm 1.6$ | $7.1 \pm 1.9$ | 4–12 |
| Cohesive | $5.6 \pm 3.2$ | $2.2 \pm 0.5$ | 1–4 |
| Total | $9.3 \pm 3.3$ | $35.6 \pm 10.4$ agents / scene | |
| *Ego-Robot Sensor Suite* | | | |
| Sensor | Resolution / Channels | FOV / Range | Track |
| RGB Camera (front) | $640 \times 480$ px | $90^\circ$ | 1 |
| LiDAR Raycast (1D ring) | 1 ch, 360 rays | 20 m | 2 |

### B.5 Closed-Loop Evaluation Suite

Table 7 summarises the closed-loop evaluation suite. This section provides supplementary detail on each component.

**Scene diversity.** The 100 scenarios are drawn from three CARLA maps: Town03 (urban grid, 33 scenarios), Town05 (suburban grid with roundabouts, 31 scenarios), and Town10 (downtown complex with plazas and mixed pedestrian infrastructure, 36 scenarios). Each scenario fixes the full pedestrian crowd configuration—agent positions, group assignments, and trajectories—so that all evaluated models face an identical scene and results are directly comparable across models.

**Group configuration.** The average of $9.3 \pm 3.3$ social groups per scene reflects a deliberate mix of all three archetypes. Queue groups ($1.6 \pm 1.7$ per scene, 2–7 members) form at bus stops and

food carts present across all three towns. Gathering groups ($2.1 \pm 1.6$ per scene, 4–12 members) occupy open plazas and building facades, which are most prevalent in Town10. Cohesive groups ($5.6 \pm 3.2$ per scene, 1–4 members) represent small walking parties distributed across sidewalk corridors in all towns; their higher count reflects the ubiquity of dyads and triads in urban pedestrian flow. Together, the 100 scenarios contain a total of $35.6 \pm 10.4$ group-affiliated agents per scene, representing a substantial fraction of the $60 \pm 16$ pedestrians present.

**Expert path and SPL denominator.** The Expert Path ($46.1 \pm 8.8$ m) is the length of the compliant trajectory synthesised by our pipeline (Appendix B.2)—the shortest path from start to goal that incurs no topological violations. This is used as the SPL denominator in place of the Euclidean goal distance, so that SPL penalises detours relative to a violation-free reference rather than relative to a straight line that may pass through group formations.

**Sensor configuration and track separation.** Track 1 (vision-based IL models) operates from a single front-facing RGB camera at $640 \times 480$ resolution; no depth, LiDAR, or map information is provided. Track 2 (PPO) operates from a 1D LiDAR ring returning 360 range readings at 20 m range; no camera image is available. The two tracks therefore represent fundamentally different observation modalities and cannot be directly compared on navigation efficiency metrics.

**Episode termination and metric computation.** Episodes terminate upon goal arrival (success), physical collision with any agent or static object (failure), or 120 s timeout. QCR, GDS, and CPI are computed per episode from the fixed group graph $\mathcal{G}$ generated at scene creation time. Geometric metrics (SR, SPL, CR) follow standard definitions [1]. We do not aggregate the compliance triple into a scalar; instead, all three are reported separately to preserve interpretability across archetype types.

**Validity filtering.** Episodes with Off-Sidewalk Rate OSR $> 0.5$ are flagged in the compliance tables but retained for transparency. The compliance triple should not be compared across methods when OSR differs substantially, since trajectories that route through non-walkable areas bypass the formations the metrics measure against.

# C Metric Derivations

## C.1 F-Formation Intuition

When people engage in focused interaction, they arrange themselves so each participant can readily attend to a jointly held focus. Kendon [23] formalizes this as the *F-formation*: a spatial configuration comprising (i) the *O-space* — the open, jointly attended region in the center of the group; (ii) each participant's *P-space* — their personal interaction space; and (iii) the *R-space* — the surrounding activity region of the whole group. Our three compliance metrics each target the spatially critical zone of one archetype. Figure 7 shows one CARLA Bird's Eye View (BEV) example per archetype with zones annotated; the dashed red line illustrates a violation trajectory.

**Queue — P-space cutting (QCR).** A queue is a goal-oriented F-formation in which members arrange in a directed chain toward a shared service point. Adjacent members' P-spaces overlap to form a continuous physical barrier along the line. A robot that steps between two consecutive queue members crosses this P-space boundary — the spatial equivalent of cutting in line. Each such crossing event raises both disruption and collision risk, because the inter-member gap is physically occupied. QCR counts crossing events per episode; each queue is counted at most once.

**Gathering — O-space disruption (GDS).** A stationary gathering is a focused F-formation in which members orient inward toward a shared O-space — the empty interactional core. This space is geometrically unoccupied (members stand at its periphery, not inside it), yet it is socially attended:

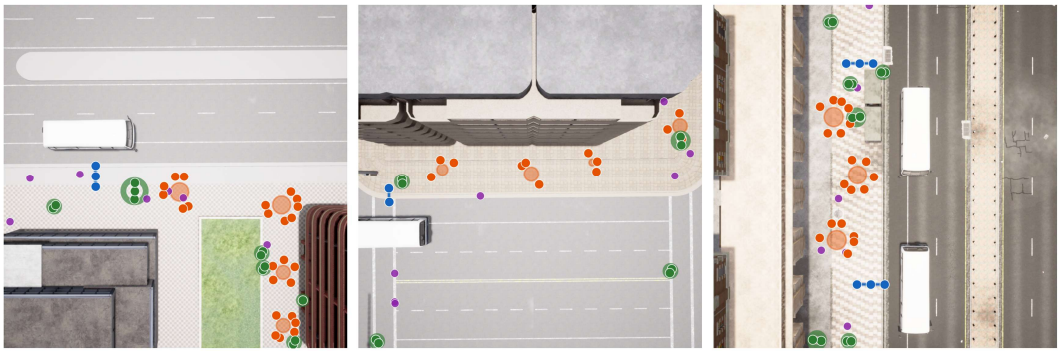


Figure 7: CARLA Bird's Eye View of three generated crowd scenes. Coloured dots indicate group-affiliated agents: **blue** = queue, **orange** = gathering, **green** = cohesive; **purple** = unaffiliated pedestrians. Each scene simultaneously contains all three formation archetypes, naturally situated near CARLA's pedestrian infrastructure (buses, bus stops, and plazas).

every participant faces it simultaneously. A robot traversing the O-space disrupts the focused encounter of all members at once, even without physical contact. GDS measures the total time the robot spends inside the O-space, accumulated over all gathering groups in the scene.

**Cohesive group — R-space intrusion (CPI).** A cohesive walking group is a mobile F-formation whose members move jointly in a shared direction. The group's R-space is an annular buffer around the convex hull of member positions that translates with the group over time. Unlike the queue P-space or gathering O-space, this buffer is dynamically occupied by members in motion — a robot that enters it embeds itself inside a moving social structure, simultaneously elevating collision probability and fragmenting the group's formation. CPI counts the total time the robot spends inside the R-space buffer.

### C.2 Metric Computation from Ego Trajectory

Each metric takes as input the ego trajectory $\tau_R$ (robot positions sampled at 1 Hz) and the group graph $\mathcal{G}$ (member positions at each frame), and returns one scalar per episode.

**Algorithm 1** Queue Cutting Rate (QCR)

**Require:** $\tau_R$; queues $\{Q_k\}$ with ordered member positions per frame
**Ensure:** QCR $\geq 0$ *(number of queues cut)*

QCR $\leftarrow 0$
**for** each queue $Q_k$ **do**
  **for** each frame $t$ **do**
    **for** each adjacent member pair $(v_i,\ v_{i+1})$ in $Q_k$ **do**
      **if** robot crossed the segment $v_i \rightarrow v_{i+1}$ between $t-1$ and $t$ **then**
        QCR $\leftarrow$ QCR $+ 1$; **goto next queue**
      **end if**
    **end for**
  **end for**
**end for**
**return** QCR *(each queue counted at most once per episode)*

**Algorithm 2** Gathering Disruption Score (GDS)

**Require:** $\tau_R$; gatherings $\{M_m\}$ with member positions per frame; $\Delta t = 1\,\text{s}$
**Ensure:** GDS $\geq 0$ *(total seconds inside any O-space)*

GDS $\leftarrow 0$
**for** each frame $t$ **do**
  **for** each gathering $M_m$ **do**
    compute O-space: disk centred at member centroid, radius = nearest member distance
    **if** robot is inside O-space of $M_m$ **then**
      GDS $\leftarrow$ GDS $+ \Delta t$; **goto next frame**
    **end if**
  **end for**
**end for**
**return** GDS *(each frame counted at most once)*

**Algorithm 3** Cohesive Proxemic Intrusion (CPI)

**Require:** $\tau_R$; cohesive groups $\{C_c\}$ with member positions per frame; $D_{\text{buf}} = 1.2\,\text{m}$; $\Delta t = 1\,\text{s}$
**Ensure:** CPI $\geq 0$ *(total seconds inside any R-space)*

CPI $\leftarrow 0$
**for** each frame $t$ **do**
  **for** each cohesive group $C_c$ **do**
    compute R-space: 1.2 m buffer outside convex hull of member positions
    **if** robot is inside R-space of $C_c$ **then**
      CPI $\leftarrow$ CPI $+ \Delta t$; **goto next frame**
    **end if**
  **end for**
**end for**
**return** CPI *(each frame counted at most once)*

# D Extended Analysis: Vision Foundation Models

The analyses in this section use an additional run on the Track 1 evaluation episodes under the same protocol, with per-episode logging. Since pedestrian appearance and ego spawn positions are randomized in each episode, values differ slightly from Table 2. Three analyses complement the Track 1 results: archetype density and difficulty scaling (Section D.1), compliance on successful trajectories (Section D.2), and topology violations in collision-free episodes (Section D.3). Analyses cover seven Track 1 models: CityWalker (Zero-Shot, Std-FT, Topo-FT), ViNT (Std, Topo), and NoMaD (Std, Topo).

## D.1 Archetype Density and Difficulty Scaling

A well-calibrated benchmark should produce proportionally harder compliance as social structure becomes denser. We test this by partitioning CityWalker (Topo-FT) episodes by the number of formations of each archetype type present in the scene, and comparing compliance violations and success rates across easy and hard subsets. Difficulty tiers are defined per archetype: Queue **Easy** (1–2 queues) vs. **Hard** ($\geq$4); Gathering **Easy** (1–2) vs. **Hard** ($\geq$3); Cohesive **Easy** (1–3 groups) vs. **Hard** ($\geq$8).

Table 8 shows that compliance violations and SR both scale with density for Gathering and Cohesive archetypes. In the hardest Gathering scenes (avg. 3.5 gatherings), GDS is 2.4$\times$ higher and SR drops from 91.7% to 59.6% compared to easy scenes. Cohesive groups show a similar pattern: CPI grows 7.5$\times$ and SR drops 16 pp from easy to hard. Queue scaling is less pronounced in SR (the hardest queue scenes are actually easier to navigate), but QCR roughly doubles. These results confirm that denser archetype configurations impose substantially harder compliance burdens, validating that the benchmark's topological structure exercises reasoning beyond individual obstacle avoidance.

Table 8: Difficulty scaling by archetype density: CityWalker (Topo-FT) on Easy and Hard subsets. $n$: number of episodes in the subset. **Avg. Count**: mean number of formations of that archetype type present in the scene. **Metric**: the archetype-specific compliance metric for that row (QCR for Queue, GDS [s] for Gathering, CPI [s] for Cohesive).

| **Archetype** | **Difficulty** | $n$ | **Avg. Count** | **SR (%)** ↑ | **Metric** ↓ |
|---|---|---|---|---|---|
| Queue (QCR) | Easy (1–2 queues) | 36 | 1.6 | 66.7 | 0.25 |
| | Hard (≥4 queues) | 20 | 4.4 | 85.0 | 0.55 |
| Gathering (GDS) | Easy (1–2 gatherings) | 24 | 1.8 | 91.7 | 2.30 |
| | Hard (≥3 gatherings) | 47 | 3.5 | 59.6 | 5.46 |
| Cohesive (CPI) | Easy (1–3 groups) | 28 | 2.4 | 78.6 | 0.59 |
| | Hard (≥8 groups) | 24 | 10.5 | 62.5 | 4.41 |

### D.2 Compliance on Successful Trajectories

Aggregate compliance scores mix two distinct failure modes: robots that fail navigation entirely (and therefore encounter fewer formations) and robots that reach the goal but violate social norms along the way. To isolate the latter, Table 9 re-computes the compliance triple for (i) episodes where the robot reached the goal, and (ii) those episodes that additionally pass the off-sidewalk filter (OSR $\leq 0.5$), which removes trajectories that bypass formations by routing through non-walkable space.

Table 9: Topological compliance conditioned on navigation success. *Success* filters to episodes where the robot reached the goal; *Success + OSR≤0.5* additionally excludes episodes with >50% time off-sidewalk. $n$: number of episodes in the subset (higher reflects better SR ↑). QCR: mean queue-cutting events per episode. GDS: mean seconds inside O-space per episode. CPI: mean seconds in R-space buffer per episode. All three compliance metrics are lower-is-better (↓).

| | **Success only** | | | | **Success + OSR ≤ 0.5** | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | $n$ ↑ | **QCR** ↓ | **GDS** ↓ | **CPI** ↓ | $n$ ↑ | **QCR** ↓ | **GDS** ↓ | **CPI** ↓ |
| CityWalker Zero-Shot | 58 | 0.24 | 2.74 | 2.85 | 53 | 0.26 | 2.98 | 2.98 |
| CityWalker Std-FT | 74 | 0.26 | 4.26 | 2.01 | 67 | 0.28 | 4.64 | 2.21 |
| CityWalker Topo-FT | 67 | 0.27 | 2.77 | 1.19 | 61 | 0.30 | 2.99 | 1.24 |
| ViNT Std | 69 | 0.35 | 1.94 | 1.43 | 65 | 0.35 | 2.02 | 1.52 |
| ViNT Topo | 63 | 0.25 | 3.25 | 0.67 | 61 | 0.26 | 3.35 | 0.69 |
| NoMaD Std | 80 | 0.45 | 4.84 | 2.71 | 76 | 0.46 | 5.07 | 2.86 |
| NoMaD Topo | 84 | 0.32 | 3.79 | 1.77 | 81 | 0.32 | 3.78 | 1.74 |

Three patterns emerge. First, filtering to successful episodes does not uniformly improve compliance: CityWalker Topo-FT's QCR *increases* from 0.27 to 0.30 after OSR filtering, indicating some successful trajectories pass through queues rather than around them—an instance of the *illusion of social compliance* discussed in the main paper. Second, Topo-FT consistently reduces CPI on successful episodes across all three architectures: CityWalker Topo-FT (1.19 vs. 2.01), ViNT Topo (0.67 vs. 1.43), and NoMaD Topo (1.77 vs. 2.71). NoMaD Topo shows the broadest improvement, reducing all three metrics compared to Std, the only model in this table to achieve monotone compliance gains. Third, NoMaD is largely unaffected by the OSR filter (3–4 episodes filtered), confirming strong sidewalk adherence when it navigates successfully.

Notably, Topo-FT models succeed on fewer episodes than their Std counterparts (CityWalker: 67 vs. 74; ViNT: 63 vs. 69). Formation-avoidance detours can push episode duration past the 120 s timeout, converting a socially compliant trajectory into a navigation failure. NoMaD Topo is an exception ($n$=84 > 80), likely because its goal representation is more robust to re-acquisition after a detour.

### D.3 Topology Violations in Collision-Free Episodes

Collision rate is the most common safety metric in robot navigation evaluation. If compliance violations always co-occur with physical collisions, then QCR, GDS, and CPI would add no information beyond what CR already captures. This analysis tests the opposite: do robots that avoid all collisions still routinely violate social formations? Table 10 reports, for each model, the fraction of episodes with no physical collision and, among those collision-free trajectories, the fraction that nevertheless contain at least one topology violation (QCR $\geq 1$ or GDS $> 0$ or CPI $> 0$).

Table 10: Topology violations among collision-free episodes. *No-col.* is the fraction of episodes with no physical collision; remaining columns give the fraction of those episodes containing each violation type.

| **Method** | **Regime** | **No-col.** | **Any viol.** | **QCR $> 0$** | **GDS $> 0$** | **CPI $> 0$** |
|---|---|---|---|---|---|---|
| CityWalker | Zero-Shot | 59% | 53% | 10% | 24% | 37% |
| | Std-FT | 69% | 73% | 15% | 39% | 43% |
| | Topo-FT | 70% | 59% | 16% | 33% | 34% |
| ViNT | Std | 58% | 83% | 16% | 48% | 45% |
| | Topo | 61% | 66% | 18% | 31% | 34% |
| NoMaD | Std | 45% | 78% | 22% | 44% | 44% |
| | Topo | 56% | 79% | 18% | 41% | 46% |

Across all seven model–regime combinations, **53–83% of collision-free episodes contain at least one topology violation**. ViNT Std shows the highest rate (83%): despite strong navigation performance (SR = 0.69), it traverses scene geometry aggressively and passes through social formations on most collision-free trajectories. Topo-FT reduces the any-violation rate for both ViNT (83% $\rightarrow$ 66%) and CityWalker (73% $\rightarrow$ 59%), confirming that topological training genuinely reduces social intrusion, not merely navigation failure. GDS and CPI are the most frequent individual violation types (24–48% and 34–46%, respectively), while QCR is less common (10–22%). These results motivate the topological metrics as a complementary, non-redundant evaluation axis: avoiding physical collisions does not preclude structural norm violations.