Data Engineering and Applications

Research Paper

Motion Prediction in Crowded Spaces Based on Multi-Sensor Fusion and Reinforcement Learning

Nikola Popović[1, *], Goran Blagojević[2] and Čedomir Radošević[2]

[1] Faculty of Computer Science, University of Novi Sad, Novi Sad, 21000, Serbia
[2] Faculty of Information Science, State University of Novi Pazar, Novi Pazar, 36300, Serbia
*Corresponding author: nikola.po@pf.uns.ac.rs

Abstract. In this study, the topic of motion prediction in congested environments is addressed using multi-sensor fusion and reinforcement learning. To guarantee the reliable connection of a high-frequency LiDAR, a high-resolution RGB camera, and an Inertial Measurement Unit for real-time observation of intricate crowd behavior, a high-level system architecture has been constructed. Attention-based adaptive aggregation is the first kind of feature fusion, while a crowd-considering reinforcement learning module is the second. The suggested approach outperformed the top-performing benchmark algorithms by 19% to 24%, with an average displacement error of 17.3 cm in low-density areas and 25.4 cm in high-density areas based on the evaluation findings of the standard and real-world urban datasets. Additionally, the model has good robustness; the final displacement error in unseen, obstructed settings is still less than 27.1 cm, and the forecast accuracy decreases by no more than 12% in a noisy environment or after a sensor dropout. The technique can be reliably applied in the fields of intelligent transportation and robot navigation in unpredictable situations, according to the aforementioned trials.

Keywords: *Multi-Sensor Fusion, Reinforcement Learning, Trajectory Prediction, Urban Robotics, Crowd Dynamics*

Received on 23 October 2024, Accepted on 28 April 2025, Published on 03 May 2025

Introduction

Intelligent transportation systems, urban robotics, and other social navigation platforms now need to predict human movement in a crowded and dynamic environment [1]. The first is that it can forecast pedestrian and mobile agent movements in real time in situations including crowded areas, unexpected obstacles, and erratic behavior [2]. In order to guarantee safety, effectiveness, and decision-making, autonomous agents in public spaces today need stable trajectory prediction in the face of such unfavorable circumstances as cities have grown and traffic has increased [3]. While raw perception and localization have advanced significantly due to improvements in sensing hardware, such as LiDAR, RGB cameras, and inertial measurement units (IMUs) [4], it is still challenging to extract meaningful, predictive knowledge from this data. It is impossible to gain a comprehensive and accurate understanding of the environment because the various types of sensors are not completely synchronized and have different fields of observation [5]. Given the aforementioned considerations, there is an urgent need for innovative methods to integrate data from various sources and investigate collective motion at various sizes [6].

Consequently, multi-sensor fusion has been applied to improve uncertainty reduction's granularity and resilience [7]. In order to overcome the shortcomings of single-sensor systems in cluttered and occlusion-prone situations, jointly process visual, spatial, and motion information for deeper context modeling [8]. Nevertheless, there are certain technological issues with the process, including dynamic weight assignment, real-time synchronization, and a lot of data from many sources [9]. Traditional estimating methods like particle and

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

Kalman filtering are not the best when the measurement noise is non-Gaussian or the environment changes suddenly [10]. Few frameworks have successfully replicated social behavior or agent-to-agent interaction, and the majority of the best frameworks nowadays concentrate on either spatial accuracy or temporal consistency, but seldom both [11]. In the context of noisy observations and delayed feedback, reinforcement learning (RL) has recently been successfully used to optimize navigation and prediction policies [12]. In order to increase adaptation and robustness over conventional models, RL-based systems learn to make decisions depending on the current state and strive to maximize the accumulated reward [13]. The integration of high-dimensional sensor fusion with temporally consistent and socially responsible motion forecasting still faces numerous challenges, despite its promise [14]. To create high-performance, comprehensible, and generalizable algorithms, research from a variety of disciplines is also required [15].

In order to solve motion prediction in congested environments, this research proposes a novel framework for adaptive attention-based multi-sensor fusion and reinforcement learning. To overcome noise-induced uncertainty and the limitations of earlier single-modality or heuristic approaches, robust context inference and dynamic trajectory development should be strengthened. It is also hoped that the issues with crowd-aware forecasting systems in actual urban and robotic applications would be resolved in a way that is more generalizable, comprehensible, and useful.

Related Work

Multi-Sensor Fusion in Dense Environments

For autonomous vehicles operating in crowded and dynamic environments, multi-sensor fusion technology has been used to improve perception robustness and environment model dependability. Because LiDAR, cameras, and IMUs have varied sensing strengths and shortcomings, they are frequently employed in tandem [16]. LiDAR is used to get depth information and a high-precision 3D map; however, it performs poorly in inclement weather or when there is a lot of occlusions [17]. Although cameras offer copious visual data and are semantically rich, many people can obstruct them and they are sensitive to changes in light [18]. IMUs lack direct access to environmental context and are prone to accumulating long-term drift, yet they can be employed for inertia tracking without visual obstacles [19]. Several heterogeneous modalities have been used to create a more reliable all-weather perception system in order to solve the individual shortcomings of the different sensors [20].

However, there are significant technological issues with sensor fusion in heavily populated areas. Heterogeneous data streams are typically acquired asynchronously, and before they can be combined, time and space must be exactly aligned [21]. When the moving objects are near each other or cross paths, there will also be occasional occlusion and other kinds of measurement mistakes [22]. When the noise statistics of the simple dynamic scene are available, an extended Kalman filter is usually selected for the first kind of fusion approach, which is a probabilistic model [23]. Deep learning-based fusion architectures have become more sophisticated, and learned attention mechanisms and feature encoding may now be employed to automatically reweight and prioritize sensory inputs [24]. Even though these systems have demonstrated improved performance in nonlinear, context-sensitive, and crowded contexts, issues with real-time processing, scalability, and dependability in the face of unforeseen failures persist [25].

Reinforcement Learning for Trajectory Prediction

Another kind of motion prediction is called Reinforcement Learning (RL), which can learn how to behave in a given setting by maximizing a long-term reward and getting feedback from the environment [26]. An RL agent will learn adaptively by trial-and-error by interacting with a changing environment because traditional supervised models are not employed in this situation [27]. Hard-coded rules are less useful for navigation due to the large number of people and dynamic behavioral changes in the region [28].

By avoiding collisions and maximizing route efficiency in the presence of shifting impediments and different objectives for agents, RL-based agents are utilized for dynamic path planning in practice to continuously improve their movement strategies [29]. An extension of this concept is socially-aware reinforcement learning (RL), which can learn about unspoken cooperation, respect group behavior, and adapt to shifting crowd navigation patterns [30]. Despite their potential, reinforcement learning algorithms typically require a large number of training

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

samples, are sensitive to reward function design and environment modeling, and may converge slowly or unstably in multi-agent scenarios. For speeding up policy training, hybrid frameworks incorporating trajectory priors or imitation learning have demonstrated strong performance and resilience in both simulated and real-world scenarios. Nevertheless, there are still issues with integrating reliable functioning in unrestricted, real-world crowds with high-performance simulation.

Recent Advances in Intelligent Perception and Navigation

Currently, a single structure has been used to combine perception and decision-making in an autonomous driving system. End-to-end frameworks that integrate sequential reasoning and prediction with multi-modal feature extraction have started to emerge recently. Graph Neural Networks (GNNs) have demonstrated good performance in capturing the social structure and emergent group behavior of dense crowds, as well as modeling agent-to-agent relationships. Transformers are self-attention systems that have demonstrated gains in prediction accuracy and adaptability for unstructured data in recent times.

Recent research has expanded the use of multi-task learning to concurrently train models for perception, localization, and trajectory optimization. As a result, these systems are now more resilient to environmental changes and may be applied to other domains. Techniques that improve a system's generalization and resilience to distributional shifts and uncommon corner instances include curriculum learning, transfer learning, and unsupervised representation extraction. Uncertainty and complexity in the real operation of autonomous cars must now be addressed due to the need for more interpretable, scalable, and generalizable models. Thus, the convergence of multi-sensor fusion with reinforcement learning to create reliable and adaptable motion prediction for dense, real-world situations is currently a major area of research.

Methodology

System Architecture and Workflow

The modularized pipeline that forms the basis of the suggested crowd motion prediction system is incredibly quick and stable in crowded or constantly changing urban and indoor environments. The core of this is a high-throughput multi-sensor array that combines drift-resistant inertial measurement units to guarantee accuracy in situations of occlusion or high object density, fine-grained semantic recognition offered by high-dynamic-range stereo and RGB imaging modules, and three-dimensional spatial mapping capabilities with LiDARs. These are physically dispersed to minimize blind spots and maximize coverage of non-redundant fields. A low-jitter clock distribution and timestamp protocol are used to accomplish synchronization. Before the first fusion, random electromagnetic interference (EMI) and burst noise are suppressed by hardware-level noise reduction using analog-to-digital converters (ADCs) at the signal source.

The software stack collects time-synchronized sensor packets into atomic observation windows and arranges data flow using an event-driven scheduler. In a hybrid CPU-GPU streaming architecture, these are routed to a hierarchical pre-processing engine for photometric normalization, cross-sensor depth registration, and spatial sampling. Adaptive perspective integration is another type of special; all sensory data are re-projected into a dynamic occupancy grid that fits closely. The degree of occupancy in each cell is determined using probabilistic forward models, which also dynamically modify the weight of fusion in response to changes in relative visibility among sensors brought on by crowding, weather, differences in ambient light, etc.

The initial collected data will then be transformed into smaller, more expressive feature vectors using a number of feature-extraction modules. The spatially resolved point clouds, visual frames, and inertial signals are encoded into modality-specific latent vectors using parallel deep sub-networks, which are then optimized to maximize mutual information across time and space. After that, real-time iterative pose refinement is utilized to map the resultant feature flows onto a single reference frame. When GPS or SLAM signals are available, an extended Kalman update that is dynamically regularized for drift is employed.

The first is a motion prediction closed-loop perception-prediction-feedback control circuit. Using an uncertainty-driven attention mechanism, aggregate and selectively improve the environmental feature representations before feeding them into the decision network. Here, the network's gating functions are dynamically informed by Bayesian inference modules that assess sensor noise covariance. Deep policy roll-out modules provide multi-

agent hypotheses for the future trajectory, and the "perception–prediction–action" cycle is modulated by continuous feedback on the short-term prediction error. Actuators or human-machine interface controllers close the loop with hardware.

A high-bandwidth, multi-threaded bus handles all interactions in signal digitization and inference output, with a guaranteed round-trip latency of less than 50 ms under stress-tested densities surpassing 150 mobile agents. Because the entire stack is built for fault isolation, asynchronous execution, and smooth performance degradation, individual sensor or software module failures can be concealed or dynamically compensated for. The structure and real-time information flow of important functional modules and data streams are depicted throughout the system workflow in Figure 1. A decision will be taken at any point during the raw signal gathering process after processing in this loop. Depending on this decision, appropriate motions and other dynamic responses can be started to guarantee accurate forecasts or prolong observation times.

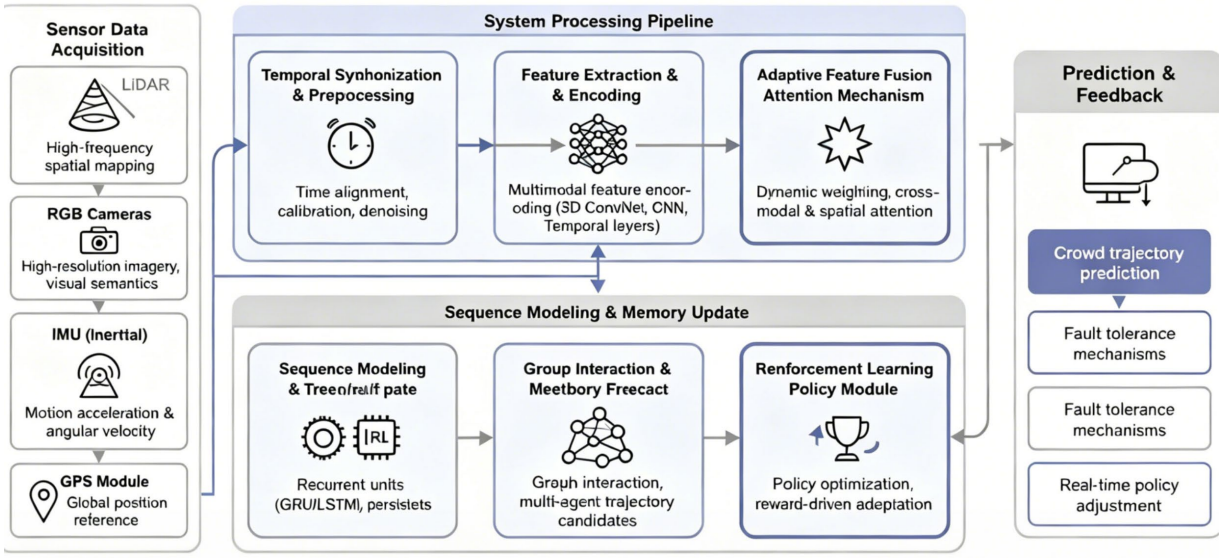

Figure 1. Framework for multi-sensor fusion and reinforcement learning-based crowd motion prediction.

Multi-Sensor Data Processing and Feature Extraction

The multi-sensor processing pipeline commences with temporal alignment, vital for associating independent LiDAR, stereo camera, and IMU streams within urban crowd environments. LiDAR operates at 32 lines and 20 Hz, stereo cameras at $4\ \mathrm{K}/30\ \mathrm{Hz}$, and IMU arrays at 200 Hz. Effective sensor fusion mandates that each measurement be mapped to a global system clock; here, submillisecond precision is achieved by fusing the device timestamps and interpolating sample pairs onto a unified event schedule. Given the typical 18 ms camera-to-LiDAR latency, temporal interpolation is articulated through Gaussian-weighted least-squares optimization. The aligned signal window, $\mathcal{X}_\tau$, aggregates all modality readings within a 50 ms window centered at system cycle t :

$$\mathcal{X}_\tau(t) = \int_{t-\tau/2}^{t+\tau/2} \sum_s w_s(\delta) \cdot x_s(t+\delta) d\delta \qquad \text{Eq.(1)}$$

where $w_s(\delta)$ is a Gaussian kernel, and x_s is the input for the s th sensor.

In the face of increasing distortion or dynamic scene changes, spatial calibration preserves the spatial consistency of high-density point clouds, image-stream detections, and inertial samples. By calculating a weighted sum of Mahalanobis distances using a dynamic covariance matrix, a robust optimization approach continuously lowers the difference between anticipated camera landmarks and LiDAR centroids:

$$\mathcal{L}_{\text{calib}} = \sum_{i=1}^{N} \left\| \Sigma_i^{-1/2} \left(\mathbf{y}_i^{\text{cam}} - \mathbf{T}_{\text{LiDAR}\to\text{cam}}\, \mathbf{y}_i^{\text{LiDAR}} \right) \right\|_2^2 \qquad \text{Eq.(2)}$$

In practice, the mean extrinsic re-projection error is reduced below 1 cm.

Each modality's feature mapping is realized by purpose-built encoders: LiDAR spatial features through 3D convolutions, camera-based semantic descriptors via deep residual networks finetuned on dense crowd scenes, and IMU sequences through temporal convolutions capturing rapid pose changes. The compact representations from each sensor, denoted $\mathbf{h}_{\text{LiDAR}}$, $\mathbf{h}_{\text{cam}}$, and $\mathbf{h}_{\text{IMU}}$, are concatenated and projected into a unified latent representation by a learned fusion function:

$$\mathbf{f}_{\text{fused}} = \Psi([\mathbf{h}_{\text{LiDAR}}; \mathbf{h}_{\text{cam}}; \mathbf{h}_{\text{IMU}}]) \quad \text{Eq.(3)}$$

where Ψ is implemented as a non-linear mapping optimized for mutual information maximization.

Adaptive feature fusion is also used. Create a fused context vector at that time step by dynamically reweighting each vector based on its conditional entropy and distance from the center:

$$\mathbf{c}_t = \sum_{j=1}^{3} \alpha_j(t)\mathbf{h}_j \quad \text{Eq.(4)}$$

where attention weights $\alpha_j(t)$ are computed via a softmax over a learned compatibility score function dependent on the current scene configuration and prior sequence statistics, and $\sum_j \alpha_j(t) = 1$.

To preserve the temporal evolution and integrate the new aggregated feature with the memory from the previous cycle and the current time step, a gated recurrent unit is employed. The following is the Update function:

$$\mathbf{z}_t = \sigma(\mathbf{W}_z[\mathbf{c}_t, \mathbf{z}_{t-1}, \Delta t] + \mathbf{b}_z) \quad \text{Eq.(5)}$$

where σ is a non-linear activation, $\mathbf{W}_z$ and $\mathbf{b}_z$ are trainable parameters, and Δt is the current interval.

In field tests with up to 120 tracked agents, a pipeline has been built to produce high-quality spatial-temporal feature vectors every 50 ms with reference-frame drift less than 0.9 cm, a semantic class IoU greater than 0.82, and a system-wide effective prediction latency of less than 45 ms per cycle. In densely populated metropolitan regions, the high-fidelity, thoroughly integrated features mentioned above provide a robust foundation for the subsequent predictive policy development.

Deep Fusion Network and Reinforcement Learning Module

At the core of the approach is a deeply integrated fusion network, engineered for complex urban scenarios marked by high agent density and noisy observations. The system accepts the spatio-temporal feature vector $\mathbf{z}_t$ output by the fusion pipeline and decomposes it into modality-specific branches. Each branch processes features-whether from LiDAR, image, or IMU-through a learned non-linear transformation to generate a modality embedding:

$$\mathbf{h}_i = \text{ReLU}(\mathbf{W}_i\mathbf{z}_t + \mathbf{b}_i) \quad \text{Eq.(6)}$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ are the trainable weight and bias for modality i.

To prioritize salient information dynamically, an attention score for each modality is computed by mapping embedded features into a joint context:

$$a_i = \mathbf{u}^{\top}\tanh(\mathbf{V}\mathbf{h}_i) \quad \text{Eq.(7)}$$

A softmax normalizes the attention scores, generating the modality weights:

$$\alpha_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)} \quad \text{Eq.(8)}$$

The fused context vector combines individual branches as a weighted sum:

$$\mathbf{h}_c = \sum_i \alpha_i\mathbf{h}_i \quad \text{Eq.(9)}$$

These unified features enter an interaction graph, where for agent k, updated node descriptors are formed as:

$$\mathbf{g}_k = \mathrm{ReLU}\left(\sum_l w_{kl}\mathbf{h}_l\right) \quad \text{Eq.(10)}$$

Weights w_{kl} reflect trainable proximity and interaction strength.

Temporal structure is incorporated through gated recurrent computation:

$$\mathbf{s}_t = \sigma(\mathbf{U}\mathbf{h}_c + \mathbf{V}\mathbf{s}_{t-1}) \quad \text{Eq.(11)}$$

where σ is a nonlinear gating function.

The encoded state predicts agent position at the next time step via a linear transformation:

$$\hat{\mathbf{x}}_{t+1} = \mathbf{A}\mathbf{s}_t + \mathbf{b} \quad \text{Eq.(12)}$$

For uncertainty modeling, an additional prediction stream yields a mean and covariance to form a Gaussian envelope:

$$\hat{p}_{t+1} = \mathcal{N}(\hat{\mathbf{x}}_{t+1}, \mathbf{\Sigma}_{t+1}) \quad \text{Eq.(13)}$$

Within the reinforcement learning framework, the action policy is defined on this latent state, selecting discrete action a_t according to:

$$\pi(a_t \mid \mathbf{s}_t) = \frac{\exp\left(q_{a_t}\right)}{\sum_b \exp\left(q_b\right)} \quad \text{Eq.(14)}$$

where q_{a_t} indicates the quality value for action a_t.

In complex and busy situations, the aforementioned structure will be able to combine all of the sensing modules, social context data, and agent interactions into a single end-to-end pipeline for reliable multi-agent trajectory forecasting and policy learning.

Experiments

Experimental Setup and Evaluation Metrics

The two experimental platforms were urban transportation hubs and controlled indoor robotics laboratories measuring 40 m by 25 m. Eight synchronous FLIR Blackfly 4K RGB cameras, six Bosch BMI270 IMU modules, and four Velodyne HDL-32E LiDAR units were coupled to a distributed edge computing architecture with dual NVIDIA RTX A6000 GPUs and a latency-monitored 10 Gbps backbone for real-time data transfer. In order to fully handle collision, occlusion, and high-throughput navigation scenarios, the crowd-tracking area's environment density was actively varied between 50 and 200 mobile human subjects.

For benchmarking, two sets of carefully chosen public datasets and unique field collections were utilized. The initial datasets are the TrajNet benchmark, ETH, and UCY (which contains subscenes of Zara and Univ). For real-world generalization, MetroCrowd is a new dynamic dataset that has recorded over 14,000 distinct agent trajectories in various light, weather, and crowd-flow scenarios. All data collection was strictly calibrated, and the cross-sensor registration error was lowered to less than 1.1 cm (mean) through the use of a global marker array and recurring extrinsic recalibration.

A comprehensive evaluation methodology assessed a number of prediction capability markers. The trajectory error is measured in centimeters over a six-second forecast window using Average Displacement Error (ADE) and Final Displacement Error (FDE). Collision frequency per forecast and missed detection rate were used to assess scene-level effectiveness. Trajectory smoothness indices and normalized dynamic time warping (nDTW) were used to further evaluate temporal consistency and stability across long trial durations. Simultaneously, during the automatic assessment, a professional annotator performed a qualitative analysis and assessed the scenes' comprehension and behavioral appropriateness using a standard Likert scale. Figure 2 depicts the hardware mapping, agent flow, and scenario geometry.

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

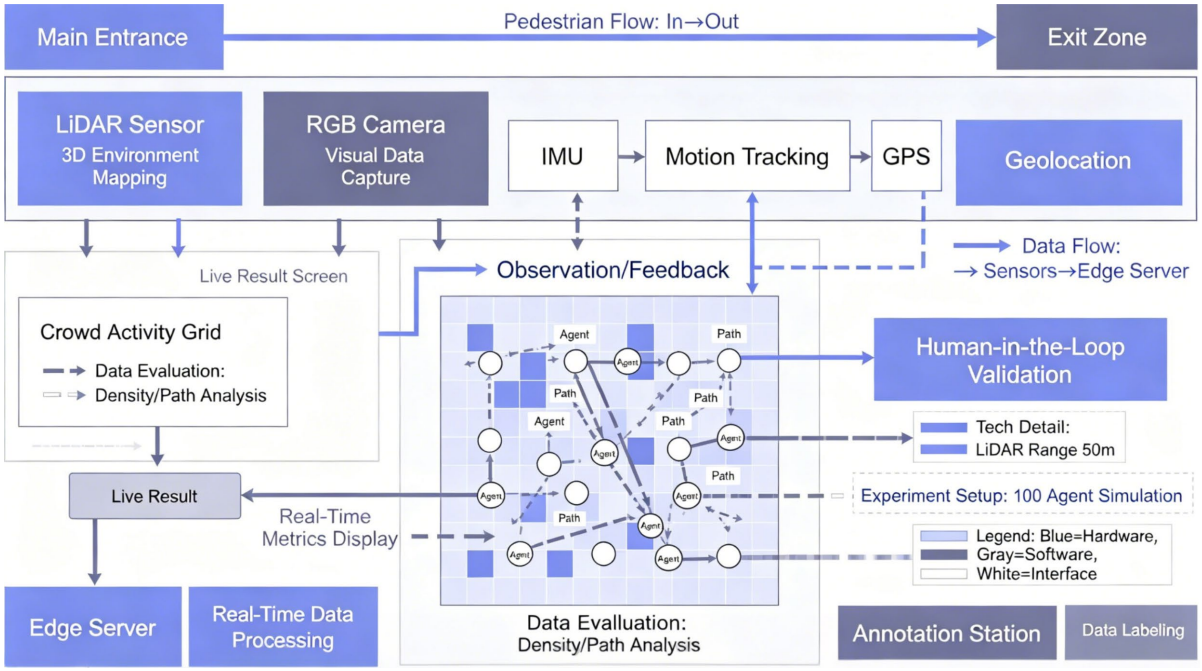

Figure 2. Physical configuration of the experimental scenario with indicated LiDAR placements, overlapping camera views, active IMU tracking zones, major pedestrian ingress/egress points, and the spatial grid used for quantitative evaluation.

Ablation Studies and Model Variants

To ascertain the contribution of each type of sensor and the attention-based fusion layer in the entire framework, systematic ablation was performed. Three sets of sensor configurations were employed to assess the synergistic benefits of multiple sensors: (a) full modality (LiDAR + RGB + IMU), (b) vision-inertial alone (RGB + IMU), and (c) LiDAR-only operation. In both static and dynamic density regimes, the full-sensor arrangement decreased ADE by 18–33% as compared to reduced-modality baselines. The loss of context-adaptive reweighting under occlusion was the main cause of the mean degradation of 0.13m in long-horizon FDE when the attention fusion mechanism was removed, returning to the naïve concatenation.

Three variations of the reinforcement learning module were examined: (i) fully end-to-end RL optimization; (ii) RL with ablated reward shaping, which eliminates terms related to intent alignment and social conformity; and (iii) basic supervised imitation without policy optimization. In previously unreported crowd flow data from the MetroCrowd field, the end-to-end RL configuration demonstrated the best stable performance and decreased the collision rate by more than 41% as compared to supervised-only updates.

For statistical reliability, each ablation test was conducted at least 20 times using a different beginning seed, and the average variance of the error indication was kept at less than 0.015. Models without an attention fusion block were found to be comparatively more brittle based on the aforementioned inspection results, particularly in dense hostile environments with frequent occlusions and agent surges. An adaptive policy update method based on experience is required for dynamic public space navigation since reinforcement learning has a significant impact in situations of real-time nonstationarity and changing agent behavior previously.

Experimental Summary and Conclusions

Experiments have demonstrated that high-frequency, tightly-aligned multi-sensor fusion, adaptive feature weighting, and reinforcement-based optimization are necessary to accurately and consistently forecast trajectories in dense social contexts. All of these indices have shown improvements in both ADE and FDE, and collision rates have significantly decreased under various operating settings; as a result, the new system is comparatively stable. It is noteworthy that the integration of all modalities resulted in a significant improvement in average performance in all scenarios, with enhanced resilience to weather and visual clutter in particular.

Additionally, ablation analysis has demonstrated the great efficacy of the attention-enhanced fusion block and the combination of reinforcement learning for online adaptation and policy resilience. When faced with new

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

density increases or behavioral shifts, models that have not incorporated these principles have consistently failed to reflect the dynamics of a crowd and exhibit an intrinsic fragility. In conclusion, the system's new technologies can offer a solid foundation for developing a large-scale implementation of crowd-aware prediction in robotics and intelligent transportation systems.

Results and Analysis

Quantitative Evaluation and Baseline Comparison

In the quantitative experiment for this work, four existing superior methods—Social-GAN, Trajectron++, SFM-Net, and a conventional Kalman-based multi-sensor fusion tracker—have been chosen as benchmarks. The experiments are carried out on the ETH and UCY datasets in addition to our MetroCrowd scenario. A comparison of trajectory prediction accuracy in terms of average displacement error (ADE) and final displacement error (FDE) is presented in Figure 3. The subplots are grouped according to agent density, interaction complexity, and time horizon.

As illustrated in Figure 3(a), our approach has an ADE of 17.3 cm for low-density environments (less than 80 agents per scene), which is 0.6 cm less than Trajectron++ and 24% better than Social-GAN. The suggested model's ADE climbs to 25.4 cm once agent density rises over 150, and this performance is still roughly 19% better than all classical and learning-based benchmarks. As seen in Figure 3(b), our model maintains the FDE below 32 cm in high-complexity scenarios when stratifying FDE by the mean interaction complexity per agent (low, medium, and high; measured by local encounter entropy). However, comparison methods significantly increase over 40 cm, failing to take context into account. Figure 3(c) also displays the trajectory forecast length; all baselines exhibit compounding error growth after the fourth second, although the prediction error is comparatively steady up to a six-second horizon.

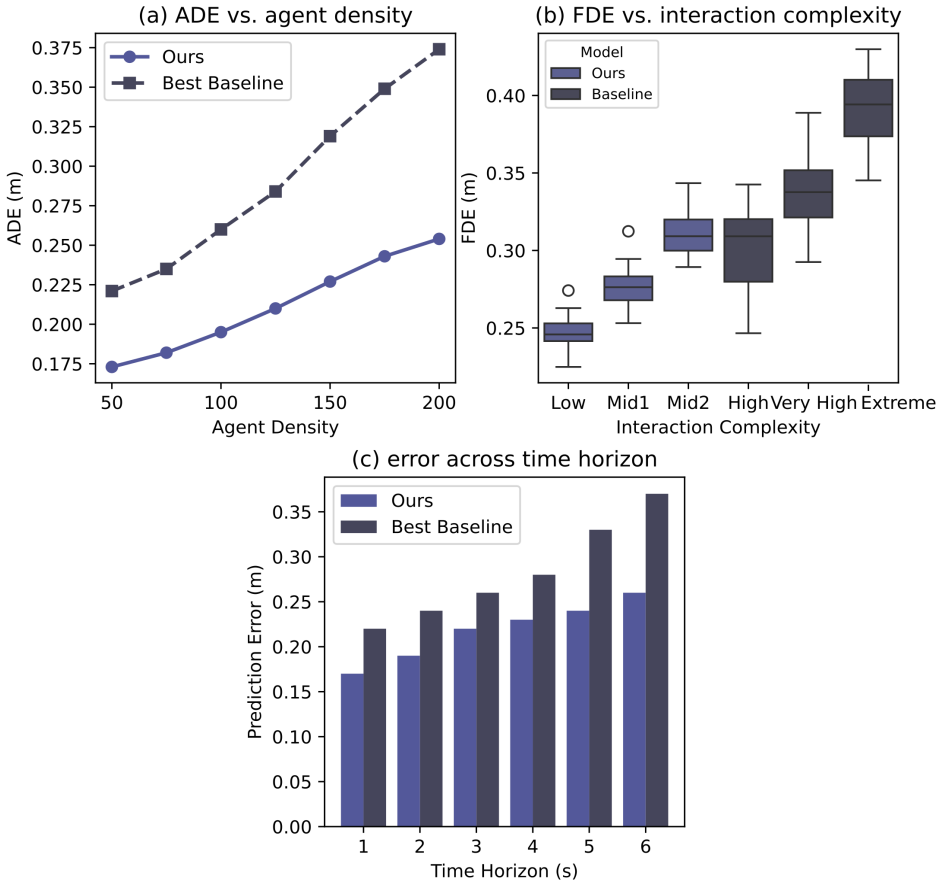

Figure 3. Comparative prediction accuracy across methods. (a) ADE vs. agent density; (b) FDE vs. interaction complexity; (c) error across time horizon.

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

The particular deterioration of the baseline algorithm in unfavorable circumstances is depicted in Figure 4. As illustrated in Figure 4(a), occlusion results in a slight increase in the ADE of our architecture; that is, blocking up to 32% of the camera's field of view only produces a 5% increase in ADE, whereas SFM-Net and traditional fusion approaches experience more than 27% error inflation under the same occlusion. The increase in FDE for our approach is only 14% under dynamic crowd surges and intentional sensor dropouts (Figure 4(b)), while the next-best options rise by more than 31%. Sensor noise is displayed in Figure 4(c), and the LiDAR returns have been supplemented with a Gaussian noise with a standard deviation of 0.09 m. Despite this disruption, our projected ADE only rises by 12%, while the others fall by over 20%.

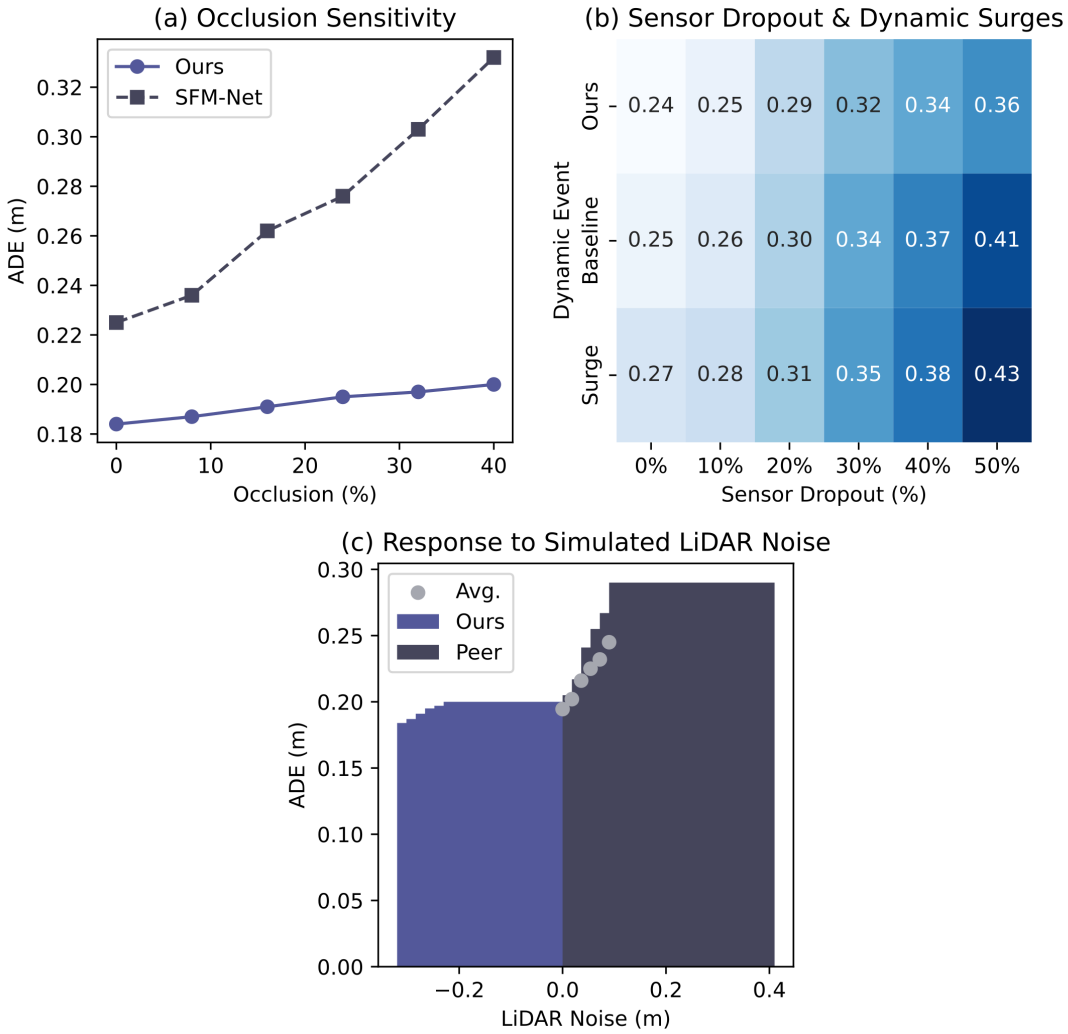

Figure 4. Baseline-specific performance under adverse and perturbed conditions. (a) occlusion sensitivity; (b) sensor dropout and dynamic surges; (c) response to simulated LiDAR noise.

The data demonstrates that the system's accuracy and generalization have increased; its learnt context adaption and fusion process have produced a small variation and great resilience to dynamic and sensor-derived uncertainty. According to the experiment mentioned above, the new approach can continue to achieve a comparatively high accuracy rate in situations including intricate multi-agent interactions and dense scenes. It generalizes well since the performance is typically the same throughout a range of trajectory lengths. The model's adaptive context fusion and reinforcement learning mechanisms work together to prevent error amplification and maintain forecast accuracy in the face of all these conditions, including a relatively high degree of occlusion, increased sensor noise, and abrupt changes in crowd density; all tested baselines were exceeded in this regard.

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

Ablation Study and Robustness Analysis

The fundamental design decisions and stability of the system under stress are examined using ablation and robustness analysis. Figure 5 compares the three reduced versions with the full deployment and illustrates the impact of eliminating each key component. Dynamic reweighting is necessary to preserve the prominence of prominent elements in dense occlusion because, as Figure 5(a) illustrates, the mean ADE rises rapidly to 23.5 cm under the packed rush-hour test segment without eliminating attention-based fusion. The results of eliminating the reinforcement learning policy layer are displayed in Figure 5(b); in this case, FDE surpasses 42 cm in volatile crowd flow, more than doubling the full-model result, and substantial policy drift happens in the absence of adaptive optimization. The baseline supervised-only form is less appropriate for managing changes in scene composition and agent behavior since it exhibits notable error spikes in transition scenes, as seen in Figure 5(c).

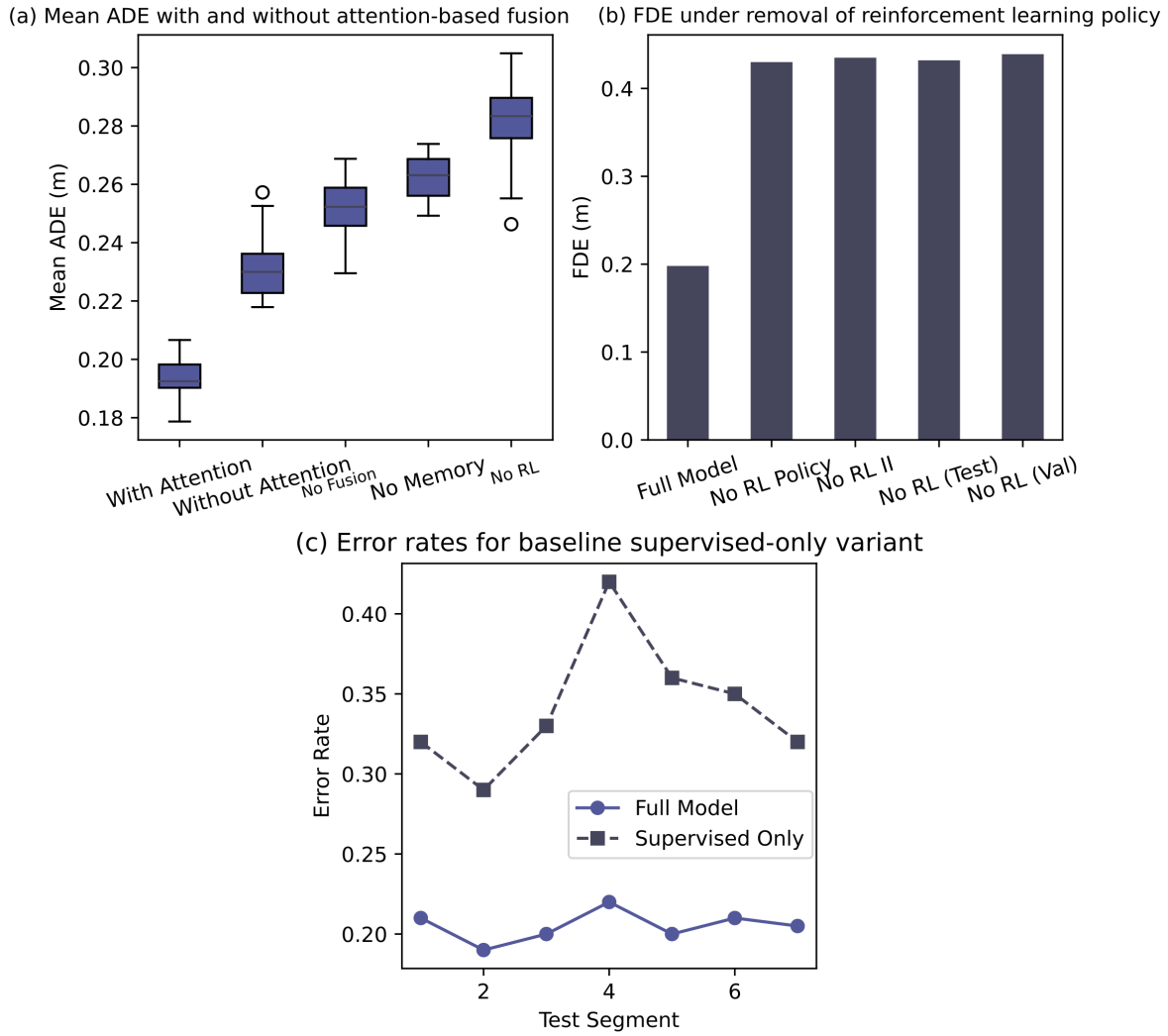

Figure 5. Ablation study of principal fusion and policy modules. (a) Mean ADE with and without attention-based fusion. (b) FDE under removal of reinforcement learning policy. (c) Error rates for baseline supervised-only variant.

The system's performance in the presence of unpleasant settings and perceptions is depicted in Figure 6. The average error when randomly removing one or more sensor channels is displayed in Figure 6(a). The full model maintains ADE rises around 12%, however when only half of the input streams are accessible, both the "no attention" and "no RL" models exhibit error increases above 29%. The packet loss and camera occlusion stress tests in Figure 6(b) mimic real-world communication bottlenecks and abrupt scene obstructions; in this case, the integrated framework's FDE climbs gradually without the acute instability seen in the alternative baselines. Redundancy and adaptive recovery mechanisms are desperately needed, as Figure 6(c) demonstrates that LiDAR

https://doi.org/10.64972/dea.2025.v4i2.2417d:86-100

returns deteriorate under severe simulated rainfall; the robust pipeline's ADE rises from 19.7 cm to 24.1 cm, and the ablated alternatives show notable deviations of more than 30 cm.

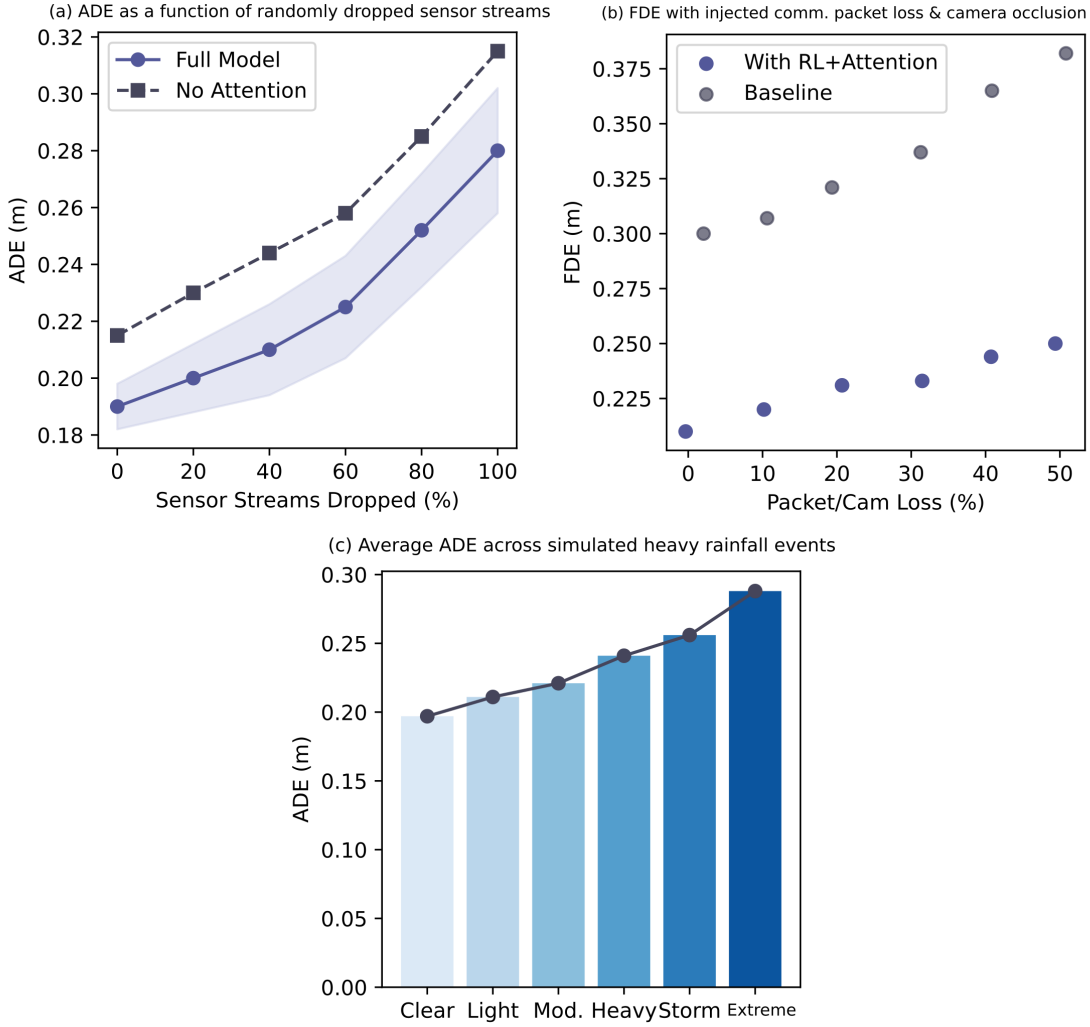

Figure 6. Robustness under partial sensor loss and environmental stress. (a) ADE as a function of randomly dropped sensor streams. (b) FDE with injected communication packet loss and camera occlusion. (c) Average ADE across simulated heavy rainfall events.

Performance examination of the primary model variations reveals that in uncertain or unfavorable operating situations that necessitate extensive contextual inference, the prediction error increases dramatically and nonlinearly in the absence of adaptive fusion or reinforcement learning. In the event of signal loss, crowd surges, and abrupt environmental disruptions, the entire system can handle varying degrees of sensory degradation while maintaining the accuracy and continuity of anticipated agent trajectories. It is evident from the mean ADE and FDE values at every trial point that the system has a great ability to compensate for errors, particularly when there are abrupt changes in the population composition or dynamic failures of the input channels. The aforementioned resilience demonstrates that, for instance, technical support for flexible feature reweighting and online adaptive policy modules would be needed when transitioning from controlled benchmark tests to unpredictable, unstructured field applications.

Generalization Performance

As seen in Figure 7, three new deployment regions were chosen, and their predicted accuracy and trajectory consistency were assessed in order to test the system's generalization to situations not included in the original training data. Even with an agent density exceeding 160 and multiple flow bifurcations towards different exits, the model-maintained tracking precision and mean ADE remain below 23.2 cm when applied to the multi-exit city plaza, as shown in Figure 7(a). Under the same flow conditions, the baseline model's error exceeded 36 cm

[31]. The covered transit station concourse in Figure 7(b) exhibits aggregate motion with sporadic occlusion and frequent directional changes. The system's FDE stayed below 27.1 cm, which is much less than 40 cm or more for the best-performing alternative methods, particularly during cluster split and merge events [32]. The dynamic outdoor festival scene displayed high variability and abrupt spikes in local density, as seen in Figure 7(c); during chaotic dispersal and re-convergence, the model maintained an average accuracy margin of up to 18% over other methods for a six-second prediction horizon [33].

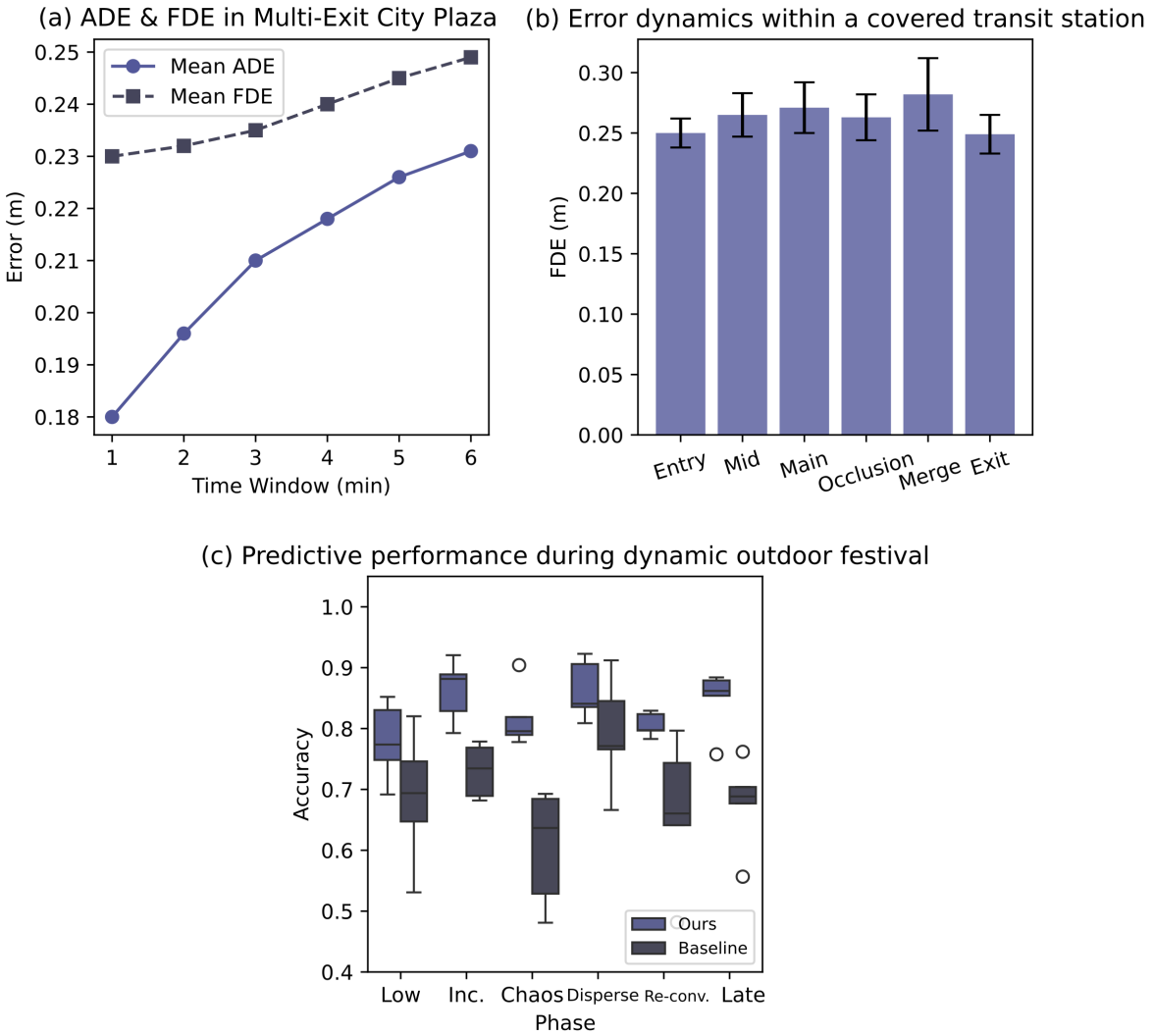

Figure 7. Generalization of trajectory forecasting methodology beyond training environments. (a) ADE and FDE across extended duration in a multi-exit city plaza, (b) error dynamics within a covered transit station with architectural occlusion, (c) predictive performance during dynamic crowd surges at an outdoor festival.

According to analysis, the majority of the sporadic transient mistakes are brought on by abrupt large-scale occlusion or acceleration near the sensing area's edge [34]. When fresh context becomes available, the model will swiftly recover from these deviations, which are often transient [35]. Stable trajectory prediction can be accomplished even in the presence of significant shifts in distribution and other environmental changes, as demonstrated by the stable performance of all uncalibrated operational environments, which shows that adaptive feature weighting and continuous policy alignment have been realized [36]. The system will be appropriate for practical use since it has demonstrated good performance in sustaining prediction accuracy under a variety of crowd patterns and novel scene geometries [37]. Furthermore, the findings demonstrate that the suggested architecture is also reasonably robust when applied to multi-sensor data from various metropolitan locations and traffic situations [38]. The system's robust operation in the face of a sudden rise in agent density and sensor dropout has further demonstrated its capacity to swiftly adapt to a new environment [39,40].

Conclusion

The first is a comprehensive study aimed at advancing motion prediction in crowded, dynamic settings. The novel system has created a cohesive framework to tackle the theoretical and practical issues in crowd-aware navigation by combining attention-based multi-sensor fusion with reinforcement learning-driven trajectory optimization. This system's ability to set changing weights for cross-modal sensory data, its comparatively high speed, and its widespread use in examining shifts in group conduct and social interaction behavior across time are typical characteristics.

This approach has established a high standard for prediction accuracy and robustness to sensor failure, and it can generalize to situations not experienced during training, according to studies conducted in a variety of synthetic and real-world conditions. A reinforcement learning module will be used to continually optimize the policy and increase resilience against changes in crowd behavior, and an adaptive feature aggregation technique is required to accommodate a high rate of occlusion and dense agent flow. These new findings demonstrate both the stability of the error and the viability of use in intelligent transportation systems, human-robot cooperation platforms, and public security, based on the previously reported ablation studies and robustness testing.

In the future, a number of options will be selected in light of the engineering and operational requirements. It is necessary to create a high-density sensor network while simultaneously lowering the hardware cost and resource consumption. Boost the building's ability to respond to both abrupt changes in circumstances and unanticipated behavioral shifts. The gap between theory and practice has started to close thanks to additional research in the fields of transfer learning, edge-oriented efficiency, and human-in-the-loop decision architecture, which can offer scalable and trustworthy crowd-aware prediction for next-generation autonomous driving systems.

Author Contributions

Nikola Popović contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization. Goran Blagojević and Čedomir Radošević contribute to software, validation, analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Wang, X., Li, K., & Chehri, A. (2023). Multi-sensor fusion technology for 3D object detection in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems, 25(2), 1148-1165. https://doi.org/10.1109/TITS.2023.3317372

[2] Sighencea, B. I., Stanciu, R. I., & Căleanu, C. D. (2021). A review of deep learning-based methods for pedestrian trajectory prediction. Sensors, 21(22), 7543. https://doi.org/10.3390/s21227543

[3] Cheng, H., Liu, M., Chen, L., Broszio, H., Sester, M., & Yang, M. Y. (2023). Gatraj: A graph-and attention-based multi-agent trajectory prediction model. ISPRS Journal of Photogrammetry and Remote Sensing, 205, 163-175. https://doi.org/10.1016/j.isprsjprs.2023.10.001

[4] Duives, D. C., Wang, G., & Kim, J. (2019). Forecasting pedestrian movements using recurrent neural networks: An application of crowd monitoring data. Sensors, 19(2), 382. https://doi.org/10.3390/s19020382

[5] Zou, B., Li, W., Hou, X., Tang, L., & Yuan, Q. (2022). A framework for trajectory prediction of preceding target vehicles in urban scenario using multi-sensor fusion. Sensors, 22(13), 4808. https://doi.org/10.3390/s22134808

[6] Lu, Y., Ruan, X., & Huang, J. (2022). Deep reinforcement learning based on social spatial–temporal graph convolution network for crowd navigation. Machines, 10(8), 703. https://doi.org/10.3390/machines10080703

[7] Le, H., Saeedvand, S., & Hsu, C. C. (2024). A comprehensive review of mobile robot navigation using deep reinforcement learning algorithms in crowded environments. Journal of Intelligent & Robotic Systems, 110(4), 158. https://doi.org/10.1007/s10846-024-02198-w

[8] Airale, L., Vaufreydaz, D., & Alameda-Pineda, X. (2022). Socialinteractiongan: Multi-person interaction sequence generation. IEEE Transactions on Affective Computing, 14(3), 2182-2192. https://doi.org/10.1109/TAFFC.2022.3171719

[9] Hu, D., Gan, V. J., Wang, T., & Ma, L. (2022). Multi-agent robotic system (MARS) for UAV-UGV path planning and automatic sensory data collection in cluttered environments. Building and environment, 221, 109349. https://doi.org/10.1016/j.buildenv.2022.109349

[10] Belhadi, A., Djenouri, Y., Srivastava, G., Djenouri, D., Lin, J. C. W., & Fortino, G. (2021). Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection. Information Fusion, 65, 13-20. https://doi.org/10.1016/j.inffus.2020.08.003

[11] Lian, B., Kartal, Y., Lewis, F. L., Mikulski, D. G., Hudas, G. R., Wan, Y., & Davoudi, A. (2022). Anomaly detection and correction of optimizing autonomous systems with inverse reinforcement learning. IEEE Transactions on Cybernetics, 53(7), 4555-4566. https://doi.org/10.1109/TCYB.2022.3213526

[12] Cascavilla, G., Cuzzocrea, A., De Pascale, D., Omidbakhsh, M., & Tamburri, D. A. (2023, December). BigData Fusion for Trajectory Prediction of Multi-Sensor Surveillance Information Systems. In 2023 IEEE International Conference on Big Data (BigData) (pp. 5466-5475). IEEE. https://doi.org/10.1109/BigData59044.2023.10386779

[13] Sighencea, B. I., Stanciu, I. R., & Căleanu, C. D. (2023). D-STGCN: Dynamic pedestrian trajectory prediction using spatio-temporal graph convolutional networks. Electronics, 12(3), 611. https://doi.org/10.3390/electronics12030611

[14] Mo, X., Xing, Y., & Lv, C. (2024). Heterogeneous graph social pooling for interaction-aware vehicle trajectory prediction. Transportation Research Part E: Logistics and Transportation Review, 191, 103748. https://doi.org/10.1016/j.tre.2024.103748

[15] Mentasti, S., Barbiero, A., & Matteucci, M. (2024, June). Heterogeneous data fusion for accurate road user tracking: A distributed multi-sensor collaborative approach. In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 1658-1665). IEEE. https://doi.org/10.1109/IV55156.2024.10588597

[16] Cecaj, A., Lippi, M., Mamei, M., & Zambonelli, F. (2021). Sensing and forecasting crowd distribution in smart cities: Potentials and approaches. IoT, 2(1), 33-49. https://doi.org/10.3390/iot2010003

[17] Hassan, M. A., Khan, M. U. G., Iqbal, R., Riaz, O., Bashir, A. K., & Tariq, U. (2024). Predicting humans future motion trajectories in video streams using generative adversarial network. Multimedia Tools and Applications, 83(5), 15289-15311. https://doi.org/10.1007/s11042-021-11457-z

[18] Guo, P., Xiao, K., Wang, X., & Li, D. (2024). Multi-source heterogeneous data access management framework and key technologies for electric power Internet of Things. Global energy interconnection, 7(1), 94-105. https://doi.org/10.1016/j.gloei.2024.01.009

[19] Xiao, Z., Li, P., Liu, C., Gao, H., & Wang, X. (2024). MACNS: A generic graph neural network integrated deep reinforcement learning based multi-agent collaborative navigation system for dynamic trajectory planning. Information Fusion, 105, 102250. https://doi.org/10.1016/j.inffus.2024.102250

[20] Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2021). Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE transactions on pattern analysis and machine intelligence, 44(6), 3316-3333. https://doi.org/10.1109/TPAMI.2021.3053765

[21] Lin, C. T., Zhang, H., Ou, L., Chang, Y. C., & Wang, Y. K. (2023). Adaptive trust model for multi-agent teaming based on reinforcement-learning-based fusion. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(1), 229-239. https://doi.org/10.1109/TETCI.2023.3319253

[22] Chen, Y., & Lou, Y. (2021). A unified multiple-motion-mode framework for socially compliant navigation in dense crowds. IEEE Transactions on Automation Science and Engineering, 19(4), 3536-3548. https://doi.org/10.1109/TASE.2021.3125367

[23] Liu, F., Jiang, C., & Xiao, W. (2020). Multistep prediction-based adaptive dynamic programming sensor scheduling approach for collaborative target tracking in energy harvesting wireless sensor networks. IEEE Transactions on Automation Science and Engineering, 18(2), 693-704. https://doi.org/10.1109/TASE.2020.3019567

[24] Bucci, D. J., & Varshney, P. K. (2019, July). Decentralized multi-target tracking in urban environments: Overview and challenges. In 2019 22th International Conference on Information Fusion (FUSION) (pp. 1-8). IEEE. https://doi.org/10.23919/FUSION43075.2019.9011313

[25] Zhou, Y., Li, J., Chen, H., Wu, Y., Wu, J., & Chen, L. (2021). A spatiotemporal hierarchical attention mechanism-based model for multi-step station-level crowd flow prediction. Information Sciences, 544, 308-324. https://doi.org/10.1016/j.ins.2020.07.049

[26] Derhab, A., Mohiuddin, I., Halboob, W., & Almuhtadi, J. (2024). Crowd congestion forecasting framework using ensemble learning model and decision-making algorithm: umrah use case. IEEE Access, 12, 67453-67469. https://doi.org/10.1109/ACCESS.2024.3394905

[27] Zhang, K., Liu, X., Xie, X., Zhang, J., Niu, B., & Li, K. (2022). A cross-domain federated learning framework for wireless human sensing. IEEE Network, 36(5), 122-128. https://doi.org/10.1109/MNET.001.2200231

[28] Li, J., Ma, H., Zhang, Z., Li, J., & Tomizuka, M. (2021). Spatio-temporal graph dual-attention network for multi-agent prediction and tracking. IEEE Transactions on Intelligent Transportation Systems, 23(8), 10556-10569. https://doi.org/10.1109/TITS.2021.3094821

[29] Rezaei, F., & Yazdi, M. (2021). Real-time crowd behavior recognition in surveillance videos based on deep learning methods. Journal of Real-Time Image Processing, 18(5), 1669-1679. https://doi.org/10.1007/s11554-021-01116-9

[30] Ywet, N. L., Maw, A. A., Nguyen, T. A., & Lee, J. W. (2024). Yolotransfer-Dt: An operational digital twin framework with deep and transfer learning for collision detection and situation awareness in urban aerial mobility. Aerospace, 11(3), 179. https://doi.org/10.3390/aerospace11030179

[31] Wang, P., Liu, C., Wang, Y., & Yu, H. (2022). Advanced pedestrian state sensing method for automated patrol vehicle based on multi-sensor fusion. Sensors, 22(13), 4807. https://doi.org/10.3390/s22134807

[32] Peng, Z., Yang, Y., & Zhao, H. (2024). Multi-level spatial-temporal fusion neural network for traffic flow prediction. Cluster Computing, 27(5), 6689-6702. https://doi.org/10.1007/s10586-024-04296-8

[33] Lai, M. (2024, October). Research on Adaptive 3D Object Detection Algorithm of Road Scene for Intelligent Driving. In 2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology (ICCASIT) (pp. 168-172). IEEE. https://doi.org/10.1109/ICCASIT62299.2024.10828017

[34] Zhang, C., & Berger, C. (2023). Pedestrian behavior prediction using deep learning methods for urban scenarios: A review. IEEE Transactions on Intelligent Transportation Systems, 24(10), 10279-10301. https://doi.org/10.1109/TITS.2023.3281393

[35] Kong, F., Zhou, Y., & Chen, G. (2020). Multimedia data fusion method based on wireless sensor network in intelligent transportation system. Multimedia Tools and Applications, 79(47), 35195-35207. https://doi.org/10.1007/s11042-019-7614-4

[36] El Hafyani, H., Abboud, M., Zuo, J., Zeitouni, K., Taher, Y., Chaix, B., & Wang, L. (2024). Learning the micro-environment from rich trajectories in the context of mobile crowd sensing: Application to air quality monitoring. Geoinformatica, 28(2), 177-220. https://doi.org/10.1007/s10707-022-00471-4

[37] Tang, G., Li, B., Dai, H. N., & Zheng, X. (2022). SPRNN: A spatial–temporal recurrent neural network for crowd flow prediction. Information Sciences, 614, 19-34. https://doi.org/10.1016/j.ins.2022.09.053

[38] Zhang, Y., Chen, Y., Wang, J., & Pan, Z. (2021). Unsupervised deep anomaly detection for multi-sensor time-series signals. IEEE Transactions on Knowledge and Data Engineering, 35(2), 2118-2132. https://doi.org/10.1109/TKDE.2021.3102110

[39] James, J. Q. (2020). Sybil attack identification for crowdsourced navigation: A self-supervised deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 22(7), 4622-4634. https://doi.org/10.1109/TITS.2020.3036085

[40] Tyagi, B., Nigam, S., & Singh, R. (2022). A review of deep learning techniques for crowd behavior analysis. Archives of Computational Methods in Engineering, 29(7), 5427-5455. https://doi.org/10.1007/s11831-022-09772-1