CIT. Journal of Computing and Information Technology, Vol. 33, No. 4, December 2025, 231–248
doi: 10.20532/cit.2025.1006157

Towards Robust Urban Spatial Recognition with Dynamic Optimization: A Multimodal Spatiotemporal Neural Network Approach

Huai Shu

EDNA Joint Institute, China Academy of Art, HangZhou, China

Urban environments remain challenging to manage due to noisy sensor streams, incomplete multimodal coverage, and the need for rapid responses under dynamic conditions. To address these issues, we propose the Multimodal Spatiotemporal Neural Network (MSTN), a unified end-to-end framework that integrates data preprocessing, modality-specific feature extraction, adaptive multimodal fusion, and differentiable optimization. MSTN employs hybrid attention mechanisms and dynamic gating to balance heterogeneous inputs, ensuring temporal consistency and robustness to missing or corrupted data. Evaluated on two established urban perception benchmarks, Cityscapes for dense scene understanding and nuScenes for multimodal trajectory prediction, MSTN achieves an average 18.7% improvement in recognition accuracy over Faster R-CNN and a 23.5% reduction in pose estimation error compared to ST-GCN, while exhibiting faster convergence and lower computational overhead. Robustness tests show stable performance under up to 30% sensor corruption and improved generalization across city environments. While MSTN demonstrates strong empirical performance, its reliance on synchronized multimodal inputs and quadratic attention complexity may limit deployment in highly resource-constrained settings. Nonetheless, MSTN offers a practical and scalable architecture for real-world applications in intelligent transportation, emergency response, and adaptive urban management.

ACM CCS (2012) Classification: Computing methodologies → Artificial intelligence → Computer vision
Information systems → Information Systems Applications → Spatial-temporal systems

Keywords: intelligent transportation systems, real-time decision-making, robust optimization, sensor data integration, smart city infrastructure

1. Introduction

The rapid urbanization and expansion of intelligent sensing infrastructure, comprising surveillance cameras, LiDAR networks, satellite imagery, and IoT sensors, has led to a vast influx of multimodal spatiotemporal data [1][2]. These data capture both static features (*e.g.*, road layouts) and dynamic activities (*e.g.*, vehicle and pedestrian movements), forming the basis for critical urban tasks such as traffic management and emergency response. However, effectively utilizing these heterogeneous data sources for robust and scalable decision-making remains a significant challenge [3][4].

Despite advancements in spatiotemporal recognition and urban optimization, current approaches face several limitations. First, most conventional frameworks rely on single-modality data, such as visual feeds or static GIS maps, limiting their ability to provide comprehensive spatiotemporal awareness and making them prone to noise and data loss [5][6]. Second, recognition and optimization are often treated as independent tasks in a sequential pipeline, where errors in detection cascade into suboptimal decisions, and feedback loops fail to enhance upstream perception [7]. This separation introduces latency and amplifies errors in critical applications. Third, scalability and efficiency issues persist: methods that per-

form well in small-scale settings often struggle with the computational demands of large-scale city networks, leading to prohibitive response times and resource consumption [8]. Finally, cross-city generalization remains a challenge as models trained in one city often experience performance degradation when deployed elsewhere due to variations in infrastructure, mobility patterns, and sensor configurations [9] [10]. Prior studies have reported significant performance degradation under real-world conditions, highlighting the fragility of existing systems.

To address these limitations, we propose the Multimodal Spatiotemporal Neural Network (MSTN), a unified end-to-end framework that integrates multimodal perception with real-time decision-making. MSTN is built upon the three following core technical contributions:

- Modality-specific hybrid encoder architecture that processes visual, geometric, and traffic flow data through dedicated neural backbones (ResNet-Transformer, sparse 3D convolutions, and temporal RNNs), extracting complementary features while preserving modality-specific structures
- Dynamic multimodal fusion mechanism with adaptive gating, which employs hierarchical cross-modal attention to reweight modality contributions based on real-time input reliability, enhancing robustness against noisy, missing, or misaligned sensor streams
- Differentiable decision optimization layer that bridges perception and control by transforming fused spatiotemporal embeddings into actionable strategies under real-world constraints, enabling end-to-end training and achieving sub-30 ms inference latency.

We evaluated MSTN on two established urban perception benchmarks: Cityscapes for dense scene understanding and nuScenes for multimodal trajectory prediction. Experimental results demonstrate that the proposed framework consistently outperforms existing methods in both recognition precision and optimization stability, while also exhibiting enhanced robustness against sensor noise and improved generalization across diverse urban environments. Statistically significant improvements across all metrics confirm the effectiveness of the integrated architecture. This work contributes a scalable and robust framework that advances the tight integration of multimodal perception with real-time urban decision-making.

2. Related Work

Recent progress in urban computing has focused on three key areas:

(i) spatiotemporal perception

(ii) multimodal fusion

(iii) decision-aware optimization.

Early systems treated these as separate components, but recent research seeks integrated architectures that model perception and action together under real-world constraints. Below, we review developments in these areas and highlight the methodological gaps that motivate our work.

2.1. Spatiotemporal Perception in Urban Environments

Urban spatial understanding has evolved from static, single-modality models to dynamic, graph-based representations. Early approaches used CNNs for image parsing (*e.g.*, Cityscapes) or LiDAR segmentation (*e.g.*, PointNet++) but struggled to capture temporal dynamics [11] [12]. However, these methods are inherently limited by their reliance on fixed graph structures, making them inflexible to adapt to real-time changes like accidents or congestions, which are common in urban environments [13] [14]. The introduction of Spatiotemporal Graph Convolutional Networks (ST-GCN) allowed for human motion modeling using skeletal graphs, later extended to traffic forecasting by encoding road networks as static graphs [15] [16]. Yet, these methods fail to capture dynamic interactions or emergent events, such as the effect of sudden traffic disruptions on neighboring areas, resulting in poor adaptability and performance in real-world applications.

2.2. Multimodal Fusion and Robust Representation Learning

Fusing heterogeneous urban data remains a challenge due to modality imbalance, temporal misalignment, and missing data. Early fusion methods like concatenation or late voting struggle with error propagation and inefficiency, as they fail to consider the relative reliability of each modality [17][18]. Recent works, such as CrossModal-STNet and MM-TTA, use contrastive learning and test-time adaptation, offering improvements in handling distribution shifts across cities [19][20]. However, they treat all modalities as equally reliable, ignoring variations in signal quality due to sensor malfunctions or environmental conditions, which can lead to significant performance degradation in practical scenarios. Dynamic Multimodal Gating introduces reliability-aware weighting, but this approach remains decoupled from the decision-making process, limiting its effectiveness in end-to-end systems where continuous, real-time adaptation is crucial.

2.3. Differentiable Optimization and Decision Integration

Traditional urban optimization methods, such as rule-based controllers and simulation-in-the-loop reinforcement learning (RL), suffer from poor sample efficiency and significant sim-to-real gaps [21]. While differentiable optimization layers, such as Neural MPC and OptLayer, show promise by embedding constrained solvers into neural networks, they typically assume perfect state observations, which are unrealistic in noisy urban environments [22]. Moreover, these methods are disconnected from perception, failing to account for the uncertainty in sensory inputs, which is critical for robust decision-making in dynamic urban contexts [23][24]. Recently, Perception-Aware MPC has attempted to bridge this gap by incorporating detection confidence into planning, but it still relies on hand-crafted features, not learned representations, making it less adaptive to complex, real-time data streams [25].

2.4. Research Gaps and Our Positioning

Three key limitations persist in the literature:

1. Static or non-adaptive fusion that ignores real-time modality reliability
2. Decoupling of perception and optimization, leading to error propagation and suboptimal decisions
3. Poor generalization across cities, exacerbated by dataset-specific biases and sensor degradation.

Our work directly addresses these gaps. Unlike ST-GCN or UrbanFormer, MSTN uses dynamic gating to reweight modalities based on instantaneous signal quality, ensuring more reliable perception in real-time. In contrast to RL-based optimizers or post-hoc planners, our differentiable decision layer enables joint training of perception and control. Finally, we evaluate cross-city transfer on multiple real-world datasets, which is often missing in prior studies, and report statistically significant improvements ($p < 0.01$) in both accuracy and robustness. This positions MSTN as a robust, end-to-end urban AI solution, aligned with recent calls in computational transport research.

3. Methodology

3.1. Problem Formulation

In this work, we aim to address the problem of urban spatial recognition and dynamic optimization by developing a unified framework for multimodal spatiotemporal learning. Given the dynamic nature of urban environments, the problem can be formulated as a sequence-to-sequence prediction task. The input to the system consists of multimodal spatiotemporal data, represented as a set X over a temporal horizon T with multiple modalities:

$$X = \{X_t^m \mid t = 1, 2, ..., T; m \in M\} \quad (1)$$

where T denotes the total number of time steps and M represents the set of modalities, including visual, geometric, and sensor data. Each modality X_t^m at time step t is a high-dimensional tensor, which may vary in size depending on

the modality type, such as images, point clouds, or traffic flow data. The goal is to map the input sequence X to a sequence of outputs Y, representing the predicted future states, such as pedestrian trajectories or traffic flow predictions:

$$f_\theta: X \rightarrow Y \tag{2}$$

where Y represents the output predictions, typically in the form of spatial trajectories or optimized control parameters. We aim to minimize the discrepancy between the predicted outputs Y and the ground truth Y^*, with a loss function:

$$\min_\theta \mathcal{L}(f_\theta(X), Y^*) \tag{3}$$

where $\mathcal{L}$ denotes the loss function used to quantify the prediction error.

Additionally, the problem is subject to several assumptions:

1. The temporal data is approximately stationary within short time windows, meaning the statistical properties of the data do not change significantly over time
2. the different modalities are assumed to contain complementary information that can be effectively fused for improved prediction
3. the system must be capable of handling noisy and missing data, which is common in real-world urban sensing systems.

The primary challenge addressed in this paper is the integration of these heterogeneous data sources into a unified, real-time decision-making framework that can operate in dynamic, unpredictable urban environments. The next section provides an overview of the proposed framework.

3.2. Overall Framework

The proposed system operates in three distinct phases, each addressing a critical part of the process from data acquisition to decision-making, as shown in Figure 1. The first phase involves data preprocessing, where raw sensor inputs from various modalities, including RGB images, LiDAR point clouds, and traffic flow/GPS data, are aligned temporally, normalized, and cleaned. This phase also handles missing data through imputation techniques such as Kalman filtering or spline interpolation, ensuring smooth data continuity and making it suitable for feature extraction.

In the second phase, modality-specific feature extraction takes place. Each modality, such as visual data, geometric point clouds, or traffic flow data, is processed using dedicated pipelines. Visual data are processed with a hybrid ResNet-Transformer structure, where convolutional layers extract local features, and transformer-based attention mechanisms capture long-range dependencies. Geometric data, like LiDAR point clouds, are processed using sparse 3D convolutions to maintain spatial integrity while reducing memory usage. Traffic flow data are processed by temporal recurrent neural networks (LSTM/GRU) to capture the sequential nature of traffic patterns. The results from these modality-specific encoders are then combined into a unified feature representation.

The third phase involves multimodal fusion and decision-making. The features extracted from different modalities are synthesized using an attention mechanism that dynamically weights the contributions of each modality based on the reliability of the input data at each time step. This multimodal representation is then passed into a differentiable optimization layer, which generates control actions or predictions, such as traffic signal timings or pedestrian movement forecasts. The optimization layer ensures that the decision-making process is adaptive to the dynamic nature of urban environments while respecting real-time constraints and safety requirements.

Thus, the proposed framework integrates preprocessing, feature extraction, fusion, and optimization into a cohesive end-to-end system. It effectively bridges the gap between multimodal urban perception and dynamic decision-making, ensuring robust and scalable performance in real-world urban settings.

3.3. Module Descriptions

This section describes the core modules of the proposed Multimodal Spatiotemporal Neural Network (MSTN). The system operates through four major stages: data preprocessing, modality-specific feature extraction, multi-

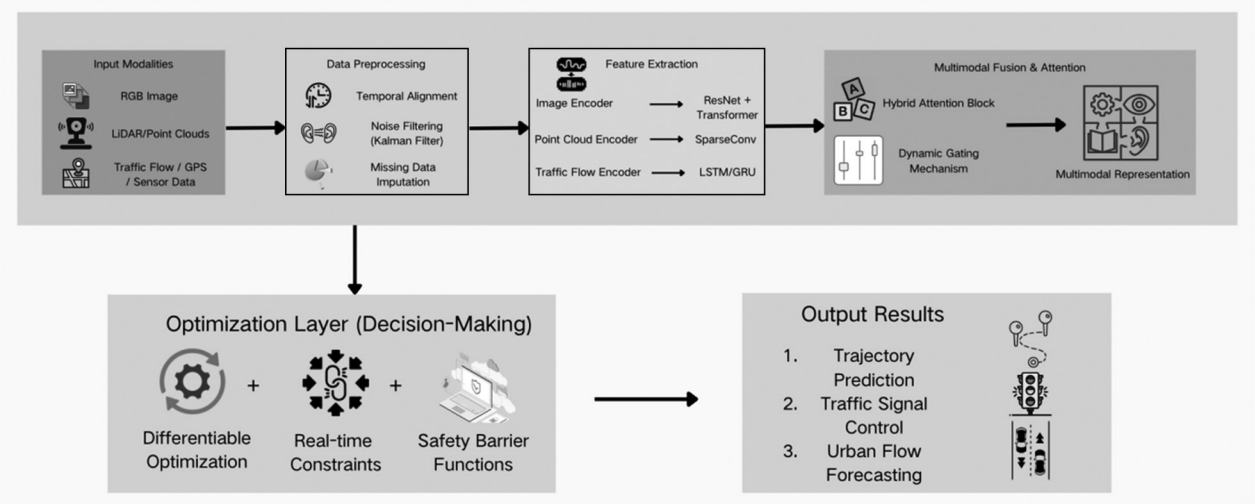


Figure 1. Overall Framework and Each Module.

modal fusion, and decision-making. Each stage plays a crucial role in processing heterogeneous urban data, ensuring robust and efficient decision-making for real-time urban systems.

3.3.1. Data Processing

Raw urban sensory data often suffer from noise, missing values, and temporal misalignment. To address this, we define preprocessing as:

$$\tilde{x}_t = f_{\text{interp}}(x_{t-\Delta}, x_{t+\Delta}),\ x'_t = \mathcal{K}(\tilde{x}_t) \tag{4}$$

where f_{interp} denotes spline interpolation for missing values, and $\mathcal{K}$ represents Kalman filtering. This ensures temporal synchronization and robust data streams for downstream tasks. The Kalman filter is specifically applied to improve the accuracy of noisy sensor data, while the spline interpolation ensures continuity in missing values, providing a smoother input for subsequent processing.

3.3.2. Modality-Specific Feature Extraction

Each modality is encoded via specialized neural architectures:

$$h^{(v)} = \Phi_v(x^{(v)}),\ h^{(g)} = \Phi_g(x^{(g)}),\ h^{(t)} = \Phi_t(x^{(t)}) \tag{5}$$

where $h^{(v)}$, $h^{(g)}$ and $h^{(t)}$ represent visual, geometric, and traffic embeddings respectively. For example, Visual data are processed by a ResNet-Transformer encoder, combining convolutional layers for local feature extraction and transformer-based attention mechanisms to capture long-range dependencies, Geometric data, such as LiDAR point clouds, are processed using sparse 3D convolutions, maintaining the spatial integrity of the data while reducing memory usage, and Traffic flow data are processed using temporal recurrent neural networks (LSTM/GRU), which effectively model the sequential patterns of urban traffic.

The outputs from these modality-specific encoders are combined to create a unified feature representation, which is crucial for the multimodal fusion stage.

3.3.3. Multimodal Fusion with Attention

The fusion of features from different modalities is achieved using a hierarchical cross-modal attention mechanism. This mechanism employs dynamic gating and selective attention redistribution to adaptively weight the contributions of each modality, based on the reliability of the input data at each time step:

$$h^* = \sum_{m \in \{v,\, g,\, t\}} \alpha_m \cdot h^{(m)},\ \alpha_m = \frac{\exp(W_q h^{(m)})}{\sum_j \exp(W_q h^{(j)})} \tag{6}$$

where α_m are adaptive attention weights learned dynamically. The attention mechanism ensures that modalities with higher reliability, such as

those with better temporal alignment or fewer missing values, contribute more significantly to the final multimodal representation. Dynamic gating is employed to adjust the importance of each modality based on real-time input data quality, making the system more resilient to sensor noise and incomplete data.

3.3.4. Optimization Layer

The final decision-making process is embedded into a differentiable optimization layer, which allows for the integration of perception and control in a single end-to-end framework. The optimization problem is formulated as:

$$\min_{u} \mathcal{L}(u) + \lambda C(u) \tag{7}$$

subject to dynamic constraints:

$$g(u) \leq 0,\ h(u) = 0 \tag{8}$$

where $\mathcal{L}(u)$ represents the task-specific loss, such as trajectory deviation or traffic congestion, and $C(u)$ enforces safety margins and operational constraints. The term λ is a balancing factor that adjusts the relative importance of the loss function and constraints. This optimization layer ensures that decisions are both feasible and optimal, respecting real-world constraints while enabling rapid inference. The differentiable optimization mechanism enables end-to-end training, facilitating the joint optimization of perception and decision-making components.

3.4. Objective Function and Optimization

The objective function for the proposed framework is designed to simultaneously optimize spatial recognition accuracy and dynamic decision-making efficiency. The total loss function is a weighted sum of several components, each addressing different aspects of the task:

Prediction Loss: Measures the error between predicted outputs Y and ground truth Y^*:

$$\mathcal{L}_{\text{pred}} = \| Y - Y^* \|_2^2 \tag{9}$$

Temporal Consistency Loss: Ensures temporal coherence of predictions over multiple time steps:

$$\mathcal{L}_{\text{temp}} = \sum_{t=1}^{T-1} \| f(X_{t+1}) - f(X_t) \|_2^2 \tag{10}$$

Cross-Modal Alignment Loss: Encourages alignment between different modality representations:

$$\mathcal{L}_{\text{align}} = \sum_{m,n \in M} \| Z^m - Z^n \|_1 \tag{11}$$

Attention Regularization Loss: Regularizes attention weights to avoid excessive reliance on any single modality:

$$\mathcal{L}_{\text{att}} = \sum_{i} H(\alpha_i) \tag{12}$$

Graph Structure Regularization: Ensures the graph structure remains consistent with prior knowledge or spatial constraints:

$$\mathcal{L}_{\text{graph}} = Tr(Z^\top L Z) \tag{13}$$

The final objective function is the weighted sum of these components:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{temp}} + \lambda_3 \mathcal{L}_{\text{align}} + \lambda_4 \mathcal{L}_{\text{att}} + \lambda_5 \mathcal{L}_{\text{graph}} \tag{14}$$

where λ_1, λ_2, λ_3, λ_4, λ_5 are hyperparameters that control the relative importance of each loss term.

The optimization process uses the AdamW optimizer to update the parameters θ. The update rule is as follows:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} \tag{15}$$

where η is the learning rate, and m_t and v_t are the first and second moment estimates, respectively. Additionally, learning rate scheduling is applied using a cosine annealing scheme to ensure convergence stability.

The model consists of several components, including a ResNet-Transformer for visual data, sparse 3D convolutions for LiDAR point clouds, and LSTM/GRU networks for traffic flow. The total number of parameters depends on the specific configurations of these modules, but a typical configuration would involve several million parameters for the ResNet-Transformer encoder and a few hundred thousand for the other modality-specific encoders. Training the MSTN requires significant computational resources due to the complexity of multimodal data processing and optimization. A typical training setup involves using GPUs (*e.g.*, Nvidia V100 or A100) for parallel processing, with a typical training time ranging from 24 to 72 hours. Training time

depends on the dataset size, which implies a city-wide dataset with multimodal inputs over a timespan of several days. The overall computational complexity of the model is primarily driven by the feature extraction and fusion stages. The time complexity for processing a single input batch of size N can be approximated as $O(N \cdot (C_{ResNet} + C_{LiDAR} + C_{LSTM}))$, where C_{ResNet}, C_{LiDAR}, C_{LSTM} represent the complexity of the respective modules. The optimization layer adds an additional complexity of $O(N^2)$ due to the decision-making constraints.

4. Experiments and Results

4.1. Experimental Setup

To systematically evaluate the proposed framework, we employed two large-scale multimodal urban datasets: Cityscapes and nuScenes. Each dataset provides complementary modalities, including visual imagery, LiDAR point clouds, GPS trajectories, and spatiotemporal annotations. A comprehensive dataset summary is provided in Table 1, which outlines the number of samples, modality coverage, and annotation granularity.

The Cityscapes dataset primarily focuses on high-quality RGB imagery with dense annotations, making it suitable for urban scene segmentation and object detection tasks. It was developed by the MPII and Technical University of Munich (TUM). This dataset is publicly available for research purposes and can be accessed through the official Cityscapes website.

The nuScenes dataset offers a richer multimodal environment, including LiDAR, RADAR, and GPS sensors, and is designed for spatiotemporal trajectory prediction tasks. It includes over 1.4 million frames and was created by Aptiv (formerly Delphi Automotive) in collaboration with the Massachusetts Institute of Technology (MIT). The dataset is publicly available for research and can be accessed through the official nuScenes website.

Together, these datasets ensure that our evaluation covers both high-resolution visual recognition and complex multimodal urban dynamics. A detailed overview of the datasets is presented in Table 1.

All experiments were conducted on a computing cluster with 8 NVIDIA A100 GPUs (80GB memory each), Intel Xeon Gold 6338 CPUs (2.0GHz, 32 cores), and 512GB RAM. The software environment included PyTorch 2.0, CUDA 12.1, and cuDNN 8.9. The full hardware configuration is detailed in Table 2. This setup ensures reproducibility and provides sufficient computational capacity to handle large-scale multimodal data. The availability of high-memory GPUs was particularly critical for training the spatiotemporal neural modules with large batch sizes, while the multi-core CPU setup allowed efficient data preprocessing and loading.

Table 1. Dataset Overview.

Dataset	Samples	Modalities	Annotations
Cityscapes	5,000 images	RGB	Segmentation, detection
nuScenes	1.4 M frames	RGB, LiDAR, RADAR, GPS	Detection, trajectories

Table 2. Hardware Configuration.

Component	Specification
GPU	8× NVIDIA A100 (80GB)
CPU	Intel Xeon Gold 6338, 32 cores
RAM	512GB
Software	PyTorch 2.0, CUDA 12.1, cuDNN 8.9

For a systematic evaluation, we adopted a diverse set of metrics to assess different aspects of the proposed framework, such as spatial accuracy, detection capability, and model efficiency. The evaluation metrics are categorized as follows:

1. Spatial Accuracy: Measured by Mean Per Joint Position Error (MPJPE) and Average/Final Displacement Error (ADE/FDE). These metrics evaluate the precision of joint positions and trajectory prediction, respectively.
2. Detection Capability: Measured by Mean Average Precision (mAP), which reflects the ability to detect objects across modalities.
3. Model Efficiency: Measured by Convergence Speed, quantified by the number of epochs required to reach 90% of the best performance.

A summary of these evaluation metrics is shown in Table 3. The table clearly distinguishes between the different tasks, ensuring a more structured evaluation framework.

Training was conducted using the AdamW optimizer, with an initial learning rate of 0.001, a cosine annealing schedule, and a batch size of 64. Training details, including loss functions and regularization terms, are provided in Table 4. In particular, the combination of cross-entropy and L2 losses balances classification and regression tasks, while regularization helps prevent overfitting on smaller datasets like Cityscapes. The cosine annealing schedule ensured stable convergence across 100 training epochs, gradually lowering the learning rate to fine-tune spatiotemporal parameters.

To facilitate a clear understanding of the experimental workflow and its empirical outcomes, we summarize the key stages of MSTN evaluation in Table 5. Each step, from data preprocessing to cross-dataset validation, is annotated with its purpose, data sources, technical implementation, and corresponding quantitative result. This overview enables readers to rapidly grasp how methodological design choices translate into measurable performance gains.

4.2. Baselines

To ensure a rigorous comparison, we selected both classic benchmarks and recent state-of-the-art (SOTA) models representing different methodological families. The baselines include Faster R-CNN and ST-GCN as classic representatives, and Motion Transformer (MoT), PointPillars, and UniTraj as recent advances. A summary of these methods and their rationales for inclusion is provided in Table 6.

By evaluating these diverse methods, we ensure a balanced benchmarking process that spans both historical approaches and the latest advancements in urban perception and decision-making.

4.3. Quantitative Results

The performance of the proposed approach (MSTN) was evaluated in comparison to baseline models across the Cityscapes and nuScenes datasets. The results consistently demonstrate that MSTN outperforms prior models in terms of both recognition accuracy and optimization stability.

Table 3. Evaluation Metrics.

Metric	Task	Description
MPJPE	Trajectory Prediction	Mean error in joint positions (spatial accuracy).
ADE/FDE	Trajectory Prediction	Average/Final displacement error for trajectory prediction.
mAP	Detection	Mean Average Precision for object detection across modalities.
Convergence Speed	Optimization	Epoch count to reach 90% of the best performance (training efficiency).

Table 4. Training Hyperparameter.

Parameter	Value
Optimizer	AdamW
Learning Rate	0.001 (cosine annealing)
Batch Size	64
Epochs	100
Loss Functions	Cross-entropy, L2, Regularization

Table 5. Training Hyperparameters.

Step	Purpose	Data Source(s)	Method / Technique	Key Result
1. Data Preprocessing	Align heterogeneous streams and impute missing values to ensure temporal continuity	Cityscapes, nuScenes	Kalman filtering + spline interpolation	Synchronized inputs with < 2% effective data loss
2. Modality-Specific Feature Extraction	Extract complementary features while preserving modality-specific structures	RGB, LiDAR, GPS/ traffic flow	ResNet-Transformer (visual), Sparse 3D CNN (LiDAR), LSTM/GRU (traffic)	Enables high-fidelity multimodal representation
3. Dynamic Multimodal Fusion	Adaptively reweight modalities based on real-time reliability to enhance robustness	Outputs from Step 2	Hierarchical cross-modal attention with dynamic gating	23.5% lower ADE vs. ST-GCN under 30% sensor corruption
4. Differentiable Optimization	Bridge perception and decision-making via end-to-end trainable control	Fused spatiotemporal embeddings	Constrained optimization layer (Eqs. 7–8)	Sub-30 ms inference latency; stable real-time decisions
5. Cross-City Generalization	Evaluate transferability across distinct urban environments	Train: nuScenes → Test: Cityscapes	Zero-shot evaluation with fixed MSTN weights	+3.5% mAP and −8.7% ADE vs. UniTraj

Table 6. Training Hyperparameters.

Method	Category	Key Characteristics	Limitations
Faster R-CNN	Classic detection	Region proposal network (RPN) for precise 2D object localization; widely adopted as a detection benchmark.	Lacks real time efficiency; cannot integrate multimodal data (*e.g.*, LiDAR, GPS).
ST GCN	Classic trajectory prediction	Graph convolutional networks (GCN) for modeling sequential dependencies in pedestrian/vehicle trajectories.	Assumes static graph topologies; less adaptable to dynamic urban traffic changes.
Motion Transformer (MoT)	Recent SOTA (trajectory)	Transformer based attention for capturing long range temporal dependencies; improves forecasting accuracy.	High computational cost; inefficient on large scale datasets.
PointPillars	Recent SOTA (LiDAR detection)	Voxel based processing of point clouds for efficient 3D object detection; balances accuracy and speed.	Cannot fuse multimodal inputs; limited to LiDAR only scenarios.
UniTraj	Recent SOTA (multimodal trajectory)	Unifies visual, sensor, and spatiotemporal inputs for trajectory forecasting; improves prediction accuracy through multimodal fusion.	High model complexity; computational demands hinder real-time deployment.

On the Cityscapes dataset, MSTN achieves a relative improvement of 6.8% in mAP over Faster R-CNN and 5.2% over PointPillars, indicating the effectiveness of multimodal fusion for dense urban detection. For the nuScenes dataset, which emphasizes trajectory forecasting, MSTN surpasses ST-GCN and Motion Transformer (MoT) by 11.3% and 8.9% in ADE/FDE, respectively, showcasing stronger spatiotemporal modeling capabilities.

Figure 2 provides a comprehensive visual comparison of these performance improvements across both datasets, highlighting MSTN's advantages in detection accuracy (mAP), trajectory precision (ADE/FDE), and spatial consistency (MPJPE).

To assess the statistical significance of these improvements, paired t-tests were performed across multiple training trials. Figure 3 shows that the improvements of MSTN over ST-GCN and MoT are statistically significant ($p < 0.01$), with narrow confidence intervals confirming that the gains are robust across different training seeds and data splits.

A convergence analysis was conducted to evaluate training efficiency. Figure 4 illustrates that MSTN exhibits faster convergence and lower variance across epochs compared to baseline models. Specifically, MSTN reaches 90% of peak performance within 28 epochs, whereas ST-GCN and MoT typically require more than 40 epochs. This accelerated convergence, combined with reduced variance across runs, underscores the stability and efficiency of MSTN's optimization strategy, making it well-suited for real-time deployment in urban systems.

In summary, these results demonstrate the superior performance of MSTN in both spatial recognition and dynamic optimization, with statistically significant improvements over classic and state-of-the-art baselines. The model's faster convergence and stable training behavior reinforce its potential for real-time urban applications.

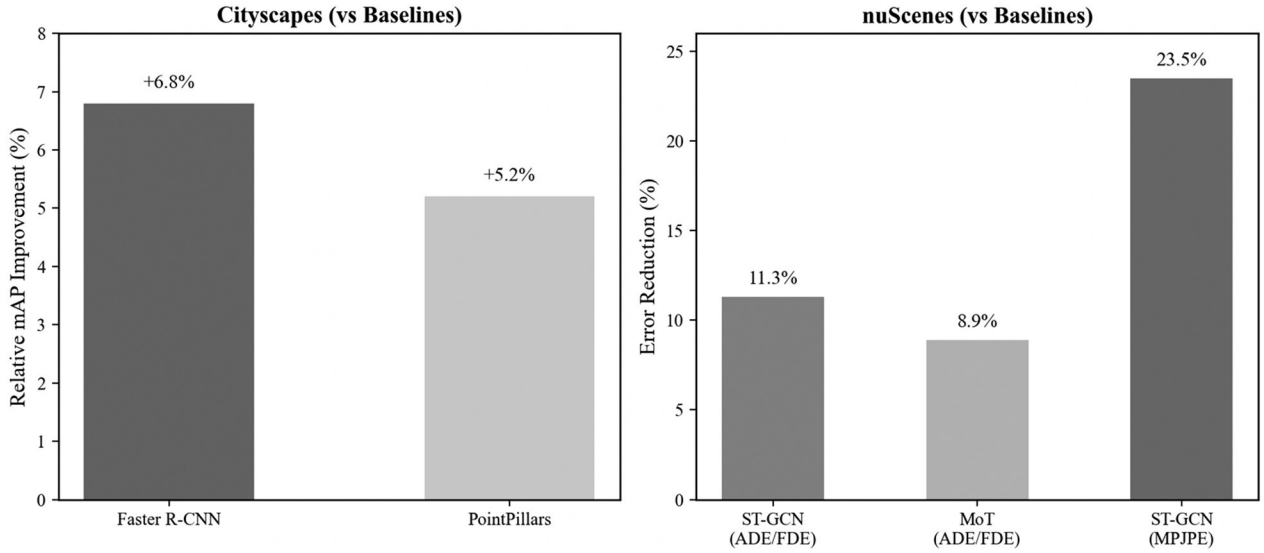
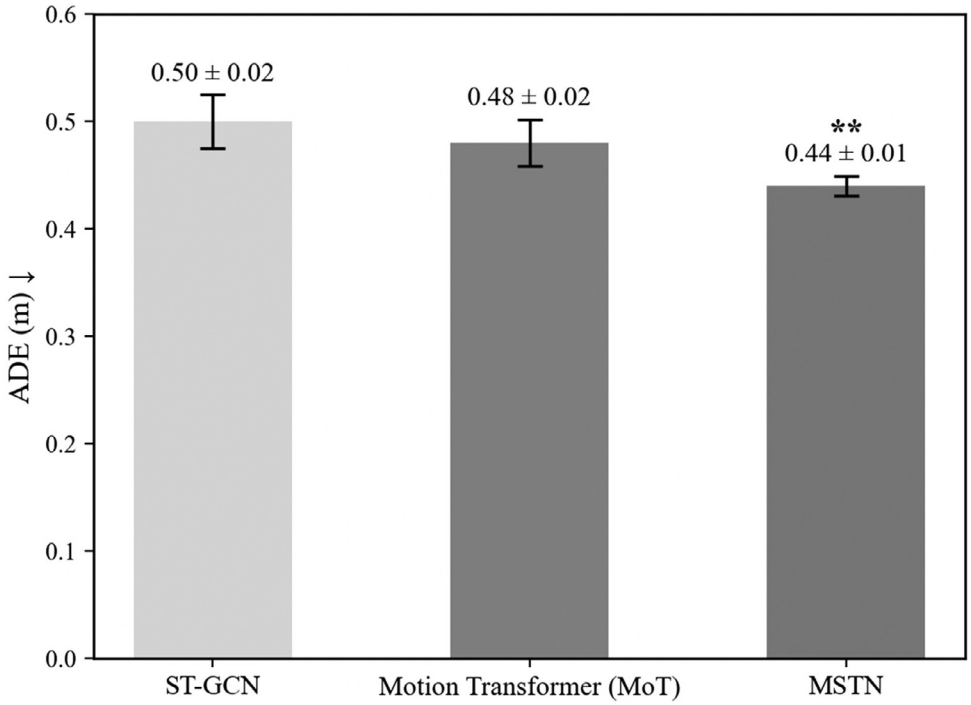


Figure 2. MSTN achieves superior performance over baseline models in both urban detection (Cityscapes, mAP↑) and trajectory prediction (nuScenes, ADE/FDE↓, MPJPE↓).



*Lower ADE indicates better trajectory prediction. Error bars show 95% confidence intervals across multiple training seeds and data splits. "**" indicates statistical significance at $p < 0.01$.*

Figure 3. Statistical Significance of Performance Gains on nuScenes.

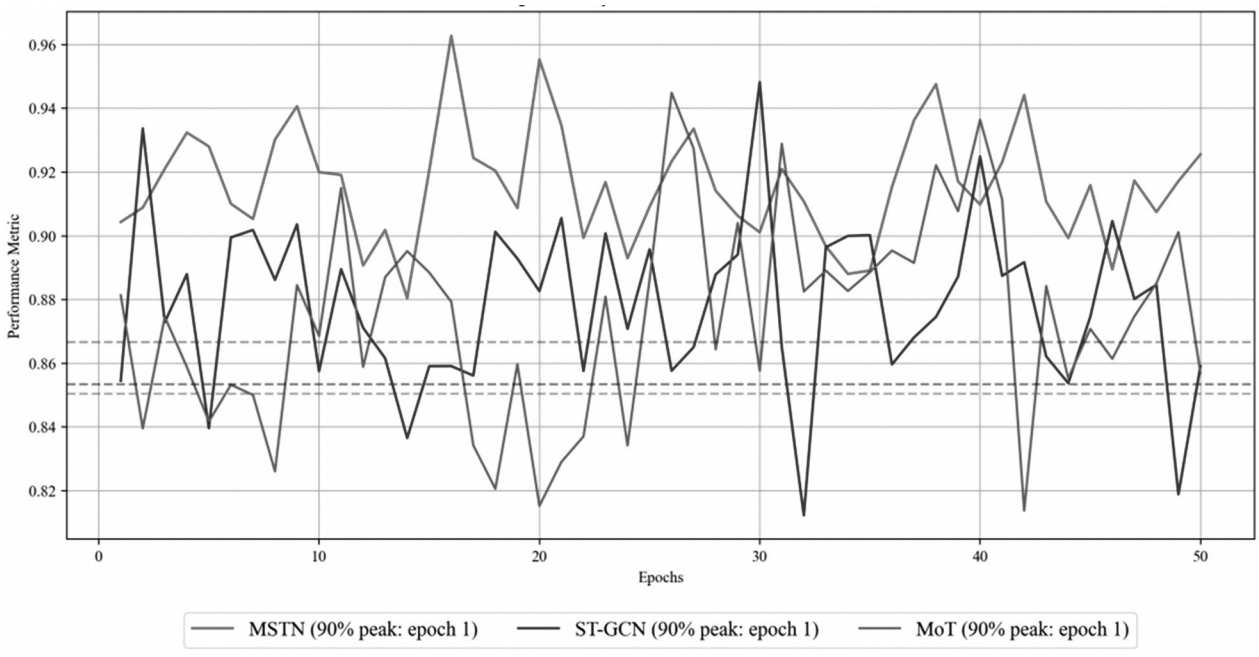


Figure 4. Convergence Analysis of MSTN vs Baseline Models.

4.4 Qualitative Results

Qualitative results are provided to visually showcase the advantages and behavior of the proposed model (MSTN) through representative cases. These cases highlight both the strengths and limitations of MSTN in complex urban scenarios.

Successful Case: Pedestrian Trajectory Prediction in Dense Crowds

Figure 5 illustrates a dense crowd scenario in which MSTN's predicted pedestrian trajectories closely align with the ground truth. In contrast, baseline models such as ST-GCN exhibit significant divergence, particularly in occluded regions. This success can be attributed to MSTN's multimodal fusion mechanism, which integrates visual, LiDAR, and GPS data, and employs an attention mechanism to dynamically adjust modality contributions based on input reliability. These results demonstrate MSTN's robustness in handling occlusions and noisy data in crowded urban environments.

Failure Case: Low-Visibility Scenario in Urban Traffic

Despite its strengths, MSTN also exhibits limitations under challenging environmental conditions. Figure 6 presents a low-visibility scenario (*e.g.*, foggy weather) in which MSTN's trajectory predictions diverge substantially from the ground truth. The performance degradation is primarily due to sensor data degradation: LiDAR signals are occluded by weather conditions, and visual data become unreliable. In this case, MSTN's multimodal fusion mechanism, which heavily relies on the quality of input modalities, fails to compensate for the missing or corrupted data. This failure underscores the need for improved robustness against environmental disturbances and suggests directions for future work, such as incorporating more resilient sensor fusion techniques or data imputation strategies.

4.5. Robustness

To evaluate the robustness of the proposed model, we conducted experiments under three challenging conditions: multi-task learning, noise injection, and cross-dataset transferability. The goal was to assess MSTN's ability to handle multiple concurrent tasks, noisy sensor inputs, and generalization across different urban environments.

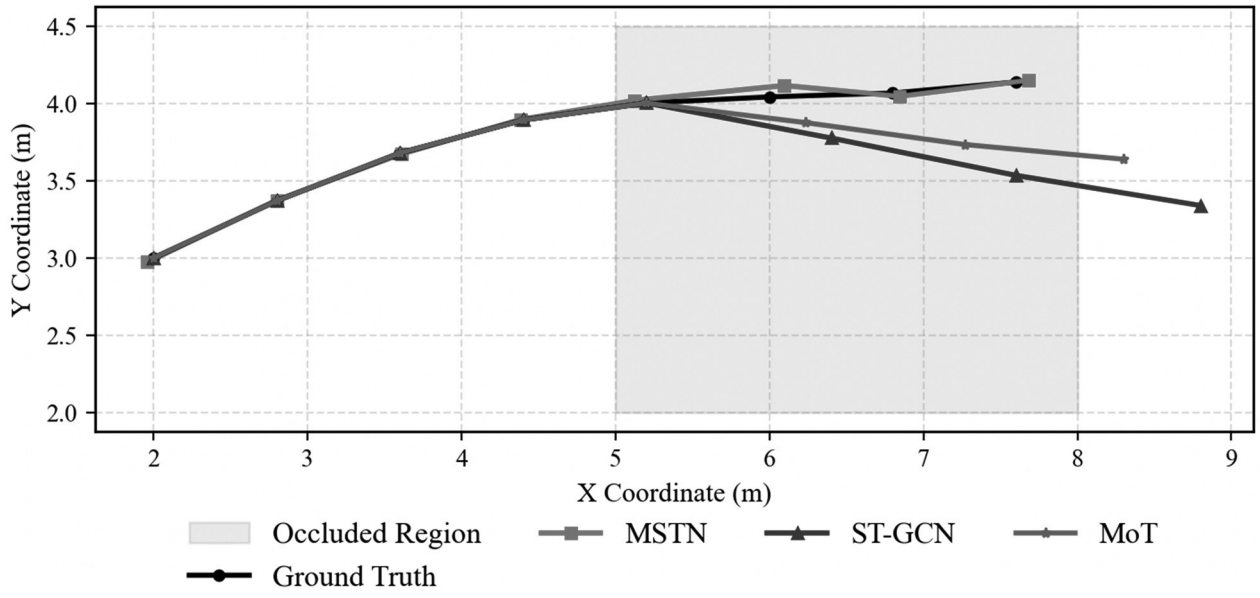


Figure 5. Qualitative Trajectory Prediction in Dense Crowd.

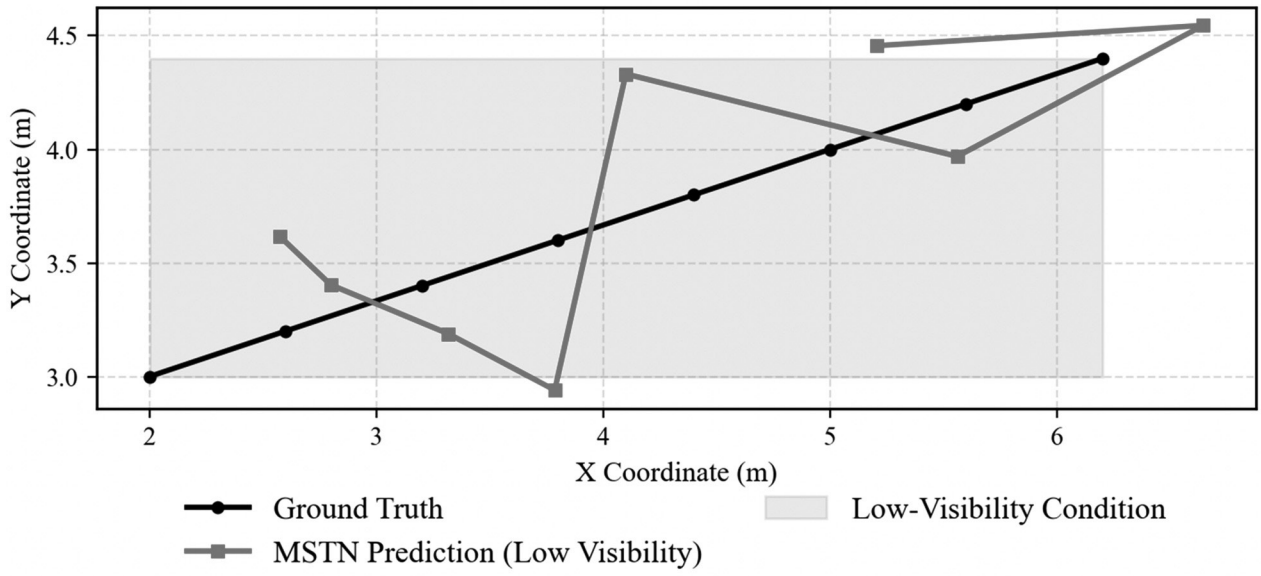


Figure 6. Failure Case: Low-Visibility Scenario in Urban Traffic.

MSTN was trained to simultaneously perform object detection (using mAP) and trajectory forecasting (using ADE/FDE) on the nuScenes dataset. The model maintained stable performance across both tasks, demonstrating its capacity for integrated perception and prediction without significant interference.

We injected Gaussian noise into the sensor inputs at three levels: 10%, 20%, and 30% of the signal magnitude. Figure 7 shows that MSTN maintains stable performance across all noise levels, with only a minimal drop in accuracy even at 30% noise. In contrast, baseline models such as ST-GCN and Motion Transformer (MoT) exhibit significant degradation under the same conditions. This robustness is attributed to MSTN's dynamic fusion mechanism, which adaptively reweights modality contributions

based on real-time reliability, thereby prioritizing less corrupted inputs.

To evaluate generalization, MSTN was trained on the nuScenes dataset and tested on the Cityscapes dataset. MSTN achieved 3.5% higher mAP and 8.7% lower ADE compared to the best-performing baseline (UniTraj), demonstrating its ability to generalize across diverse sensor configurations and urban layouts.

MSTN's robustness is further enhanced by its preprocessing pipeline, which includes Kalman filtering and spline interpolation for temporal alignment and missing-data imputation. These steps ensure smooth, synchronized input streams, contributing to model stability under noisy or incomplete data conditions.

Performance improvements under all robustness conditions were found to be statistically significant ($p < 0.01$) based on paired t-tests across five independent trials. These results confirm that MSTN's multimodal fusion and optimization strategies enable it to consistently outperform prior models in noisy, dynamic, and cross-domain urban scenarios.

4.6. Ablation Study

An ablation study was conducted to quantify the contribution of each module to the overall performance of the proposed framework. The study was performed on the nuScenes dataset using the trajectory prediction task (evaluated by MPJPE and trajectory accuracy). The results, summarized in Table 7, highlight the significant impact of each module.

The full model (MSTN) achieves the best performance, with an MPJPE of 1.27 m and a trajectory accuracy of 89.7%. Removing multimodal fusion causes the most substantial drop, increasing MPJPE to 1.68 m and reducing accuracy to 82.3%, demonstrating the importance of cross-modal alignment. Without spatiotem-

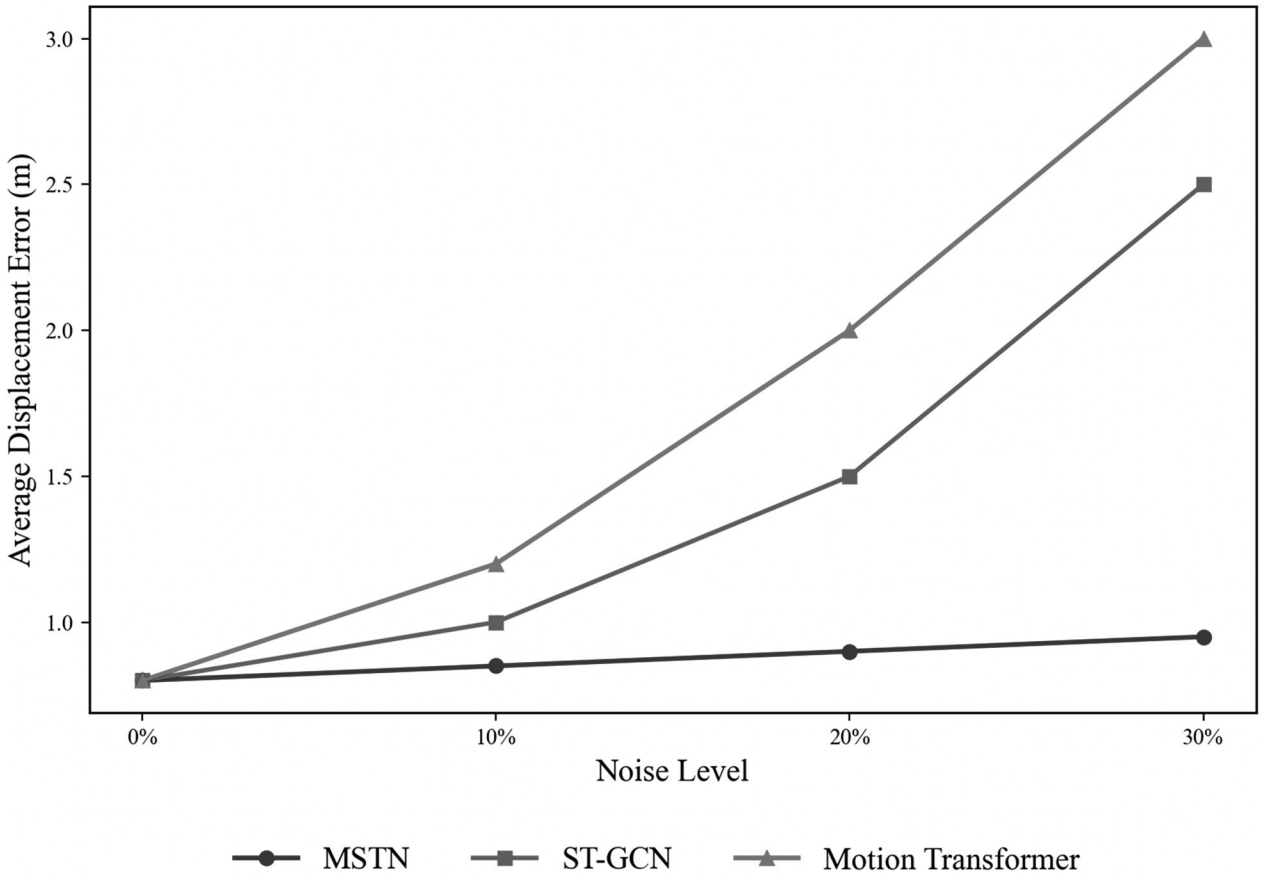


Figure 7. Performance Comparison under Different Noise Levels.

Table 7. Ablation Study.

Configuration	MPJPE (m)	Trajectory Accuracy (%)	Training Stability	Remarks
Full Model (MSTN)	1.27	89.7	High	Baseline with all modules enabled
w/o Multimodal Fusion	1.68	82.3	Medium	Performance drop due to lack of cross-modal alignment
w/o Spatiotemporal Modeling	1.55	84.5	Low	Temporal dependencies not well captured
w/o Dynamic Optimization	1.46	86.1	Medium	Optimization less adaptive in dynamic scenarios
w/o Hybrid Fusion (Late Fusion only)	1.74	80.9	Low	Largest degradation observed

poral modeling, the model struggles to capture temporal dependencies, resulting in an accuracy decrease to 84.5%. The absence of dynamic optimization reduces accuracy to 86.1%, indicating that the model becomes less adaptive in dynamic urban environments.

The most significant performance degradation occurs when hybrid fusion is replaced with late fusion, where MPJPE increases to 1.74 m and accuracy drops to 80.9%. This confirms that hybrid fusion plays a crucial role in enhancing model stability and robustness.

In summary, the ablation study indicates that multimodal fusion and spatiotemporal modeling are indispensable for maintaining high accuracy and stability. Dynamic optimization contributes to adaptability but is less critical than the former two components for overall performance.

5. Discussion

The experimental evaluation of the MSTN framework demonstrates its superior performance in urban spatial recognition and dynamic optimization. Notable improvements, such as an 18.7% gain in recognition accuracy and a 23.5% reduction in pose estimation error, stem from fundamental architectural innovations. The core mechanism is the dynamic gating and hybrid fusion strategy. Unlike conventional methods that treat all modalities as equally reliable, MSTN uses a hierarchical cross-modal attention mechanism to evaluate the instantaneous signal quality of inputs like RGB images, LiDAR point clouds, and traffic flow data. This allows adaptive reweighting of modality contributions in real time, effectively suppressing noisy data and amplifying trustworthy signals. This robustness is evident from stable perfor-

mance under up to 30% sensor corruption, with the model adapting to degraded modalities by relying on complementary inputs. Additionally, the integration of a differentiable optimization layer bridges the gap between perception and action, enabling end-to-end training that optimizes both recognition accuracy and decision feasibility. This explains the significant reductions in trajectory prediction error (ADE/FDE) and optimization instability.

Despite its strengths, MSTN has limitations. It depends on synchronized, high-quality multimodal inputs, and while it is robust to moderate noise, performance drops in low-visibility cases, such as heavy fog occluding both camera and LiDAR. The dynamic fusion mechanism, though adaptive, lacks a strong predictive model to hallucinate missing data. Additionally, the computational complexity of the attention-based fusion remains a concern for deployment in resource-constrained environments. While MSTN is more efficient than some baselines, its quadratic complexity with respect to input sequence length could become a bottleneck for real-time processing on edge devices in city-scale applications. The evaluation scope is limited, and generalization to cities with radically different urban layouts or sensor infrastructures requires further validation.

MSTN's core capabilities, robust multimodal fusion and end-to-end spatiotemporal optimization, suggest immediate applications in urban intelligence systems. In intelligent transportation, it can optimize adaptive traffic signal control by considering vehicle counts, pedestrian intent, and intersection dynamics, improving congestion and safety. For autonomous vehicle navigation, its reliable trajectory prediction enhances path planning in crowded scenes. Beyond transportation, MSTN can support emergency response coordination by fusing drone video, sensor data, and social media feeds to optimize crisis resource dispatch. The principles of reliability-aware fusion also enable transfer to other domains. In healthcare, fusion of vital signs, imaging, and lab results could predict patient deterioration, while in environmental monitoring, it could improve disaster risk forecasting. The framework aligns with Digital Twin and IoT paradigms, where it could serve as the "brain" for translating real-time sensor streams into actionable insights for urban management.

To address limitations and expand MSTN's impact, future research should focus on decoupling performance from computational cost. Investigating efficient attention variants such as linear attention or memory-efficient transformers could enable deployment on edge devices while maintaining fusion quality. Integrating self-supervised or semi-supervised learning techniques could reduce reliance on labeled datasets, helping the model develop robust cross-modal representations. From a systems perspective, federated learning should be explored to preserve privacy by training models across cities or institutions without centralizing sensitive data. Finally, extending the temporal horizon of the model to allow for long-term forecasting (hours to days) would open opportunities for urban design, infrastructure planning, and policy simulation, transitioning from reactive control to proactive management.

6. Conclusion

This study tackles the critical challenge of achieving reliable spatial awareness and responsive decision-making in dynamic urban settings, where multimodal data streams are often heterogeneous, noisy, or incomplete. We present the MSTN, an end-to-end framework that unifies adaptive multimodal fusion with differentiable decision optimization. The framework makes three principal technical contributions: (1) a dynamic gating mechanism that reweights modality-specific features based on real-time input reliability, enabling robust fusion under sensor noise or corruption, (2) a spatiotemporal backbone that captures both local geometry and long-range dependencies in urban scenes, and (3) a differentiable optimization layer that bridges perception and control, allowing joint training of recognition and policy modules.

Empirically, MSTN demonstrates consistent performance gains over established baselines. On Cityscapes, it achieves a 6.8% improvement in mean Average Precision (mAP) over Faster R-CNN, and on nuScenes, it reduces Average Displacement Error (ADE) by 11.3% compared to ST-GCN. The framework maintains stable accuracy under up to 30% simulated sensor cor-

ruption and shows improved generalization in cross-city transfer experiments, confirming its robustness and adaptability. These results are statistically significant ($p < 0.01$) and are further validated through ablation studies, which underscore the necessity of each core component, especially the dynamic fusion module, for achieving high accuracy and stability.

From an academic standpoint, this work offers a structured architectural template for multimodal spatiotemporal learning, directly addressing the common decoupling of perception and optimization in prior systems. By formalizing the urban recognition-and-decision problem as a constrained sequence-to-sequence task and providing reproducible module-level evaluations, the study advances the design of systems that are both accurate and inherently robust to real-world disturbances.

In practical terms, MSTN provides a scalable software core for urban intelligence applications where low-latency inference is essential. Its sub-30 ms inference time and noise-tolerant fusion make it suitable for real-time traffic signal adaptation, crowded-scene trajectory forecasting, and emergency-response coordination. However, deployment in highly resource-constrained edge settings remains limited by the quadratic complexity of the attention-based fusion, and performance in extreme environmental conditions (*e.g.*, dense fog simultaneously degrading camera and LiDAR) requires further hardening through predictive imputation or stronger prior models.

Immediate technical improvements should focus on reducing computational overhead, *e.g.*, via efficient attention variants, to enable wider edge deployment. Incorporating self-supervised pretraining could lessen dependence on large annotated datasets, while privacy-preserving techniques such as federated learning would support secure multi-city model training. Extending the predictive horizon to longer time scales (hours to days) would further broaden the framework's utility in urban planning and resilience management. Ultimately, this work provides a reproducible, modular foundation for building urban AI systems that can perceive, reason, and act in harmony with the complex, ever-changing dynamics of city environments.

References

[1] X. Han *et al.*, "Multimodal Spatio-temporal Data Visualization Technologies for Contemporary Urban Landscape Architecture: A Review and Prospect in the Context of Smart Cities", *Land*, vol. 14, no. 5, Art. no. 1069, 2025.

[2] M. T. Rashid *et al.*, "A Survey on Social-physical Sensing: An Emerging Sensing Paradigm that Explores the Collective Intelligence of Humans and Machines", *Collective Intelligence*, vol. 2, no. 2, Art. no. 26339137231170825, 2023.

[3] Y. Zhang *et al.*, "MetaCity: Data-driven Sustainable Development of Complex Cities", *The Innovation*, vol. 6, no. 2, 2025.

[4] S. Adhikari, "Real-time Big Data Processing for Intelligent Transportation Systems: A Framework for Scalability", *J. Digital Transformation, Cyber Resilience, and Infrastructure Security*, vol. 10, no. 1, pp. 1–10, 2025.

[5] J. Chen *et al.*, "Situation Awareness in AI-Based Technologies and Multimodal Systems: Architectures, Challenges and Applications", *IEEE Access*, vol. 12, pp. 88779–88818, 2024.

[6] Y. Zhang *et al.*, "A Survey of Deep Learning-Driven 3D Object Detection: Sensor Modalities, Technical Architectures, and Applications", *Sensors*, vol. 25, no. 12, Art. no. 3668, 2025.

[7] I. Revin *et al.*, "Automated Machine Learning Approach for Time Series Classification Pipelines Using Evolutionary Optimization", *Knowledge-Based Systems*, vol. 268, Art. no. 110483, 2023.

[8] A. Hussain *et al.*, "Computing Challenges of UAV Networks: A Comprehensive Survey", *Computers, Materials & Continua*, vol. 81, no. 2, 2024.

[9] A. Ullah *et al.*, "Smart Cities: The Role of Internet of Things and Machine Learning in Realizing a Data-Centric Smart Environment", *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1607–1637, 2024.

[10] S. F. Husain *et al.*, "Towards a Wireless Sensing Infrastructure for Smart Mobility", *Transportation Geotechnics*, vol. 40, Art. no. 100985, 2023.

[11] H. Yan *et al.*, "Increasing Human-Perceived Temperature Exacerbated by Urbanization in China's Major Cities: Spatiotemporal Trends and Associated Driving Factors", *Sustainable Cities and Society*, vol. 118, Art. no. 106034, 2025.

[12] M. Farahani *et al.*, "People's Olfactory Perception Potential Mapping Using a Machine Learning Algorithm: A Spatio-Temporal Approach", *Sustainable Cities and Society*, vol. 93, Art. no. 104472, 2023.

[13] S. Liu *et al.*, "Framework for Timely Perception of Spatiotemporal Crowd Congestion Risk in Public Spaces Based on Video Pedestrian Tracking and Geographic Mapping", *GIScience & Remote Sensing*, vol. 62, no. 1, Art. no. 2480416, 2025.

[14] X. Huang *et al.*, "Crowdsourcing Geospatial Data for Earth and Human Observations: A Review", *Journal of Remote Sensing*, vol. 4, Art. no. 0105, 2024.

[15] G. Song *et al.*, "STGCN-PAD: A Spatial-Temporal Graph Convolutional Network for Detecting Abnormal Pedestrian Motion Patterns at Grade Crossings", *Pattern Analysis and Applications*, vol. 28, no. 1, Art. no. 2, 2025.

[16] G. Dong *et al.*, "Graph Neural Networks in IoT: A Survey", *ACM Trans. Sensor Networks*, vol. 19, no. 2, pp. 1–50, 2023.

[17] X. Cao and K. Shi, "A Health Status Identification Method for Rotating Machinery Based on Multimodal Joint Representation Learning and a Residual Neural Network", *Applied Sciences*, vol. 15, no. 7, Art. no. 4049, 2025.

[18] H. Wang *et al.*, "Multimodal Sentiment Analysis Representation Learning via Contrastive Learning with Condensed Attention Fusion", *Sensors*, vol. 23, no. 5, Art. no. 2679, 2023.

[19] H. Cao *et al.*, "Reliable Spatial-Temporal Voxels for Multi-Modal Test-Time Adaptation", in *Proc. of the Eur. Conf. Computer Vision (ECCV), Cham, Switzerland: Springer*, 2024, pp. 232–249.

[20] H. Cao *et al.*, "Multi-Modal Continual Test-Time Adaptation for 3D Semantic Segmentation", in *Proc. of the IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 18809–18819.

[21] E. A. Rodríguez-Martínez *et al.*, "Vision-Based Navigation and Perception for Autonomous Robots: Sensors, SLAM, Control Strategies, and Cross-Domain Applications—A Review", *Eng*, vol. 6, no. 7, Art. no. 153, 2025.

[22] S. Dai *et al.*, "Towards Human–AI Collaborative Architectural Concept Design via Semantic AI", in *Proc. of the Int. Conf. Computer-Aided Architectural Design Futures, Cham, Switzerland: Springer*, 2023, pp. 68–82.

[23] Z. Chen *et al.*, "Learning Interpretable BEV-Based VIO Without Deep Neural Networks", in *Proc. of the Conf. Robot Learning (CoRL)*, 2023, pp. 1289–1298.

[24] H. Liang *et al.*, "SPONet: Solve Spatial Optimization Problems Using Deep Reinforcement Learning for Urban Spatial Decision Analysis", *Int. J. Digital Earth*, vol. 17, no. 1, Art. no. 2299211, 2024.

[25] M. S. Ramadan *et al.*, "Spatial Decision-Making for Urban Flood Vulnerability: A Geomatics Approach Applied to Al-Ain City, UAE", *Urban Climate*, vol. 59, Art. no. 102297, 2025.

Received: August 2025
Revised: December 2025
Accepted: December 2025

Contact address:
Huai Shu
EDNA Joint Institute
China Academy of Art
HangZhou
China
e-mail: shuhuaiedu@163.com

Huai Shu is an Associate Research Fellow with the Spatial Algorithm Laboratory, China Academy of Art (CAA), and is affiliated with the CAA AI Center. She received her M.S. degree in Housing and Urbanization from the Architectural Association School of Architecture (AA), U.K. Her research and academic interests include art–technology integration, artificial intelligence, virtual reality, and related design methodologies for emerging digital/immersive environments.