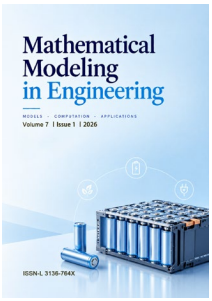
UNIVERSITY OF WEST ATTICA

eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 3, July, 2026



Edge-Adaptive VMamba-UNet for Congestion Prediction in Automated Warehouse Digital Twins

Szymon Kurowski[a,*]

[a] *Faculty of Mechatronics and Automation, Koszalin University of Technology, Koszalin, 75-453, Poland*

Received 23 April 2026; Revised 07 July 2026; Accepted 12 July 2026; Published 19 July 2026

Abstract

Automated warehouse digital twins require short-horizon forecasts that identify not only whether congestion will occur, but also where a queue front will cross an aisle, buffer, or sorter interface before the next dispatch decision. Existing methods commonly represent synchronized warehouse states as tabular sequences or smooth occupancy maps, which weakens operational boundaries and makes delayed or conflicting sensor observations difficult to handle. This study proposes an edge-adaptive VMamba-UNet for multi-horizon congestion prediction. Camera occupancy, automated guided vehicle telemetry, queue counters, order-release events, route conflicts, and sensor reliability are aligned on a fixed warehouse grid. A pre-encoder fusion module then weights motion discontinuities, topology-defined conflict edges, and confidence changes, while VMamba selective scanning propagates workload pressure across distant zones with linear complexity. The model produces continuous congestion intensity and warning-probability maps at 5, 10, and 15 min horizons. Experiments use a 42-zone simulated and replay-validated warehouse twin containing 2.8 million synchronized records. The proposed model achieves an MAE of 0.061, a warning F1-score of 89.7%, and a boundary F1-score of 83.5%. Compared with Transformer-UNet, MAE decreases from 0.083 to 0.061, warning F1 increases by 7.3 percentage points, and P95 inference latency decreases from 41.8 to 24.6 ms. Removing edge-adaptive fusion increases MAE to 0.070 and reduces boundary F1 to 77.2%. These results indicate that reliability-conditioned boundary preservation and selective state propagation can provide accurate, spatially actionable congestion forecasts within an edge-deployment time budget, although broader multi-site validation remains necessary.

Keywords: Warehouse Digital Twin, Edge-Adaptive Fusion, VMamba-UNet, Congestion Prediction, Automated Storage

1. Introduction

Automated warehouses have evolved from isolated material-handling cells into cyber-physical systems in which robots, conveyors, buffers, high-bay racks, sorters, and human-supervised exception stations interact continuously [1]. Throughput depends on coordinated equipment status, task release, route availability, and local service capacity [2]. A disturbance at an inbound dock can alter replenishment plans, occupy an AGV aisle, and reduce downstream picking speed within minutes [3]. Digital twins therefore synchronize observations with virtual representations of zones, resources, and queues to expose these interactions [4]. More recent systems extend twins beyond visualization and offline simulation toward predictive support for dispatching, slotting, and maintenance [5]. In this setting, congestion is not adequately represented by a single warehouse-level value. It is a spatial-temporal condition expressed through occupancy, queue growth, route conflict, vehicle motion, task urgency, and observation reliability. A continuous intensity map indicates expected local pressure, a warning-probability map supports intervention thresholds, and the position of the predicted boundary determines which aisle or buffer is operationally affected. Preserving all three outputs

* Corresponding author: szymon.k@tu.koszalin.pl

https://doi.org/10.68406/mme.2026.vol7iss3nm3:28-41

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

is essential because a numerically accurate forecast can still trigger an invalid action when its boundary is displaced into an adjacent zone.

Warehouse geometry further complicates prediction. Narrow vehicle lanes, irregular service demand, one-way conveyor clusters, and synchronized picking waves create dependencies that extend beyond a local image patch. Recurrent models can represent temporal accumulation but do not naturally preserve a two-dimensional route layout, whereas conventional convolutions are efficient but may lose dependencies between receiving, storage, and sorting areas [6]. Transformers model global interaction, but quadratic attention is costly for frequent twin updates at the edge [7]. Visual state-space models provide a linear-complexity alternative with a large effective receptive field [8]. VMamba propagates information through directional selective scans without constructing a full attention matrix [9]. Selective scanning alone, however, does not identify which discontinuities are operationally important. In a warehouse twin, relevant edges include lane-change interfaces, queue-spillover fronts, blocked-route borders, and abrupt changes in sensor confidence [10]. The technical gap is therefore the joint preservation of long-range workload propagation and thin, topology-dependent congestion boundaries under a fixed inference budget.

This paper addresses that gap with an edge-adaptive VMamba-UNet. The synchronized twin state combines camera occupancy, vehicle kinematics, queue counters, zone service rates, order-release pressure, and graph-derived lane conflicts. Before spatial compression, the fusion module estimates boundary salience from load gradients, conflict topology, and sensor reliability, and uses the resulting gate to retain operational edges. The encoder-decoder then applies selective state-space blocks to propagate pressure across distant zones while skip connections preserve cell-level detail. The model jointly predicts continuous congestion intensity and calibrated warning probability at three horizons. The study tests three design claims: early edge fusion should improve boundary localization, selective scanning should improve longer-horizon forecasting, and their combination should retain accuracy within an edge-computing latency budget.

The remainder of the paper develops this argument from representation to deployment evidence. Section 2 reviews warehouse digital-twin forecasting, state-space vision models, and edge-aware fusion. Section 3 defines twin-state alignment, operational edge cues, and congestion targets. Section 4 presents the VMamba-UNet architecture, selective scanning objective, and edge deployment workflow. Section 5 describes the dataset, evaluation protocol, comparative results, ablations, robustness tests, calibration, and operational interpretation. Section 6 summarizes the findings, applicability conditions, limitations, and future research directions.

2. Related Work

2.1. Digital Twin Based Warehouse Congestion Forecasting

Digital twins of warehouse systems have developed from static 3D display platforms into operational mirrors that collect order information, resource status, queue counters and equipment events. It is unknown whether the virtual environment has excluded the local constraints that cause traffic jams, such as travel lanes, docking buffers, storage access rules and temporary exception zones, and this may affect the prediction accuracy of the twin [11]. Simulation-based methods can find the congestion under an ideal plan, but often require manual adjustment of parameters and may not reflect changes in actual operating conditions. Data-driven forecasting methods reduce the calibration problem by learning from synchronous telemetry and historical workload data. Warehouses are still prone to the problems of state sparsity and event imbalance because heavy congestion accounts for a small proportion of operating hours but causes a large number of delays. A model trained to minimize the average error may appear accurate in low-flow conditions; however, it will fail to detect spillover fronts during high-flow events.

Recently, some research has begun to integrate twin representation with machine learning by converting warehouse states into zone graphs, occupancy grids or resource-time tensors [12]. Graphs can naturally show the connections of paths and conflict relationships, but they cannot fully utilize dense camera or Lidar data. Grid formulations are suitable for convolutional models and visual encoders, but they may lose the functional boundaries of adjacent cells in different operating regions. Although the hybrid tensor design can encode multiple sources in separate channels, an explicit fusion strategy is still needed to address sensor confidence and spatial discontinuity. Queueing-inspired models have the advantage of interpretability for waiting time and utilization, but the assumption of fixed dispatch decisions and battery state becomes restrictive in dynamic task priority scenarios [13]. These limitations motivate a prediction architecture that treats the warehouse twin as a structured spatial-temporal field while preserving the edge information that indicates where congestion begins, crosses, or stops [14].

2.2. State Space Vision Models and Edge-Aware Feature Fusion

Recently, the development of deep congestion prediction has included convolutional encoder-decoder networks, temporal convolutional networks, graph neural networks and attention models. UNet-style structures can be used to acquire local features and multi-scale semantic information conveniently, and thus are suitable for predicting congestion

maps at the same spatial resolution as the input state [15]. A drawback of the conventional UNet is that it cannot effectively spread the problem of global workload pressure without adding more layers or attention mechanisms. Papa et al. [16] noted that transformers address global dependency but impose significant memory and latency costs during frequent twin-state updates. Structured state-space models are other methods to address long-term dependencies by employing recurrent scanning with input-dependent parameters. Based on the modifications to the above models, selective scanning can handle image-like data in linear time with comparable accuracy [17].

Edge-aware fusion is relevant because warehouse congestion boundaries are thin and irregular but not purely visual. Segmentation networks generally infer edges from image gradients, boundary labels, or auxiliary losses, whereas a warehouse twin must also account for vehicle trajectories, lane topology, queue gradients, and sensing reliability. A stalled vehicle may create a critical route boundary without changing floor appearance, and a buffer can approach its release threshold while camera occupancy remains moderate. Fusion based only on feature magnitude can therefore suppress the earliest bottleneck signal during downsampling. Edge-aware segmentation research demonstrates the value of preserving fine boundaries [18], but its warehouse transfer requires redefining an edge as an operational discontinuity. The proposed module implements this transfer by weighting load gradients with route-conflict and reliability cues before the VMamba-UNet encoder, rather than reconstructing a visually sharp boundary only at the decoder output.

3. Edge-Aware Feature Fusion for Warehouse Twin Representation

3.1. Twin State Alignment and Prediction Target

The automated warehouse is represented in a synchronized digital-twin coordinate system divided into 124 floor cells across 42 operational zones, including inbound docks, storage aisles, replenishment lanes, picking stations, conveyor interfaces, and sorting buffers. Every 10 s, each cell receives a feature vector containing camera occupancy, AGV speed, heading dispersion, queue length, order-release pressure, route-conflict count, conveyor utilization, and sensor confidence. The aligned sequence is used to predict continuous congestion-intensity maps and binary warning maps at 5, 10, and 15 min horizons. The edge gate is bounded to the [0, 1] interval, and its contribution is evaluated through the component ablations. Warning labels are obtained by applying a training-derived delay threshold to the continuous target, ensuring that the classification task remains tied to an interpretable operational condition.

The fixed warehouse coordinate system is used for alignment, rather than the camera-centered coordinate system. Camera detections are mapped to floor cells, vehicle telemetry is aligned with the nearest feasible lane segment, and queue counters are broadcast to neighboring buffer cells according to the service direction. Order-release pressure is applied to the zones where the next task wave will be delivered, and the model can thus anticipate demand in advance of vehicle arrival. Sensor reliability is determined by the delay, missing rate and inconsistency of camera occupancy and vehicle telemetry. These reliability values are not employed as independent diagnostic metadata; instead, they are added to the input tensor, and often the cause of a false congestion alarm is stale or conflicting observations. Construction provides a specific foundation and reference for the model in the same Space.

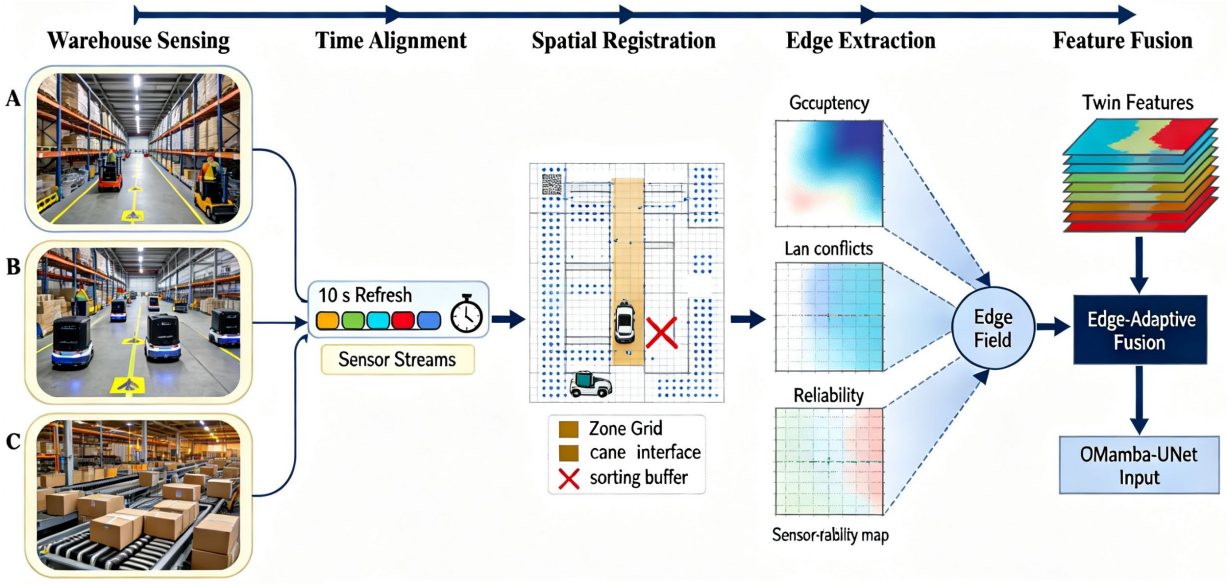


Figure 1. Workflow of warehouse twin state alignment and edge-adaptive feature fusion

Figure 1 is the warehouse twin data alignment and edge-adaptive fusion pipeline. First, there are physical sensing and event streams; next are time-window synchronization and zone-grid registration; then, edge indicators are built from

gradients and lane conflicts, and finally, fused twin tensors are passed to the VMamba-UNet predictor. The first Design Option is to perform edge detection before the main encoder. Therefore, the narrow-aisle boundary and queue-spillover-front will be affected at all subsequent scales and do not need to be reconstructed only at the decoder output.

3.2. Edge-Adaptive Feature Fusion

Let $X_t \in \mathbb{R}^{H\times W\times C}$ denote the synchronized warehouse twin tensor at time t, where H and W are the grid height and width and C is the number of feature channels. The edge cue E_t combines three sources: spatial gradients of occupancy and queue load, lane-graph conflict intensity, and sensor-confidence discontinuity. The following equation produces the normalized edge field from these three cues.

$$E_t(i,j) = \sigma(\alpha\|\nabla O_t(i,j)\|_1 + \beta C_t(i,j) + \gamma\|\nabla R_t(i,j)\|_1) \quad (1)$$

In the edge-field equation, O_t is the occupancy channel, C_t is the conflict intensity projected from the lane graph, R_t is the sensor reliability map, and σ denotes sigmoid normalization. The coefficients α, β, and γ are trainable scalars constrained to be non-negative. The resulting field increases boundary salience where physical load, graph conflict, or sensing reliability changes abruptly, while suppressing isolated fluctuations that are not supported by the other cues.

The fusion module is deliberately placed before heavy spatial compression. In early experiments, applying boundary enhancement after the encoder improved visual sharpness but did not improve warning recall, because the relevant queue fronts had already been blended into coarse latent cells. The adopted design keeps three streams in parallel: a source projection for raw multi-channel state, an edge projection for discontinuity-sensitive features, and a context projection for smooth workload pressure. The edge gate decides how much of each stream contributes at every cell. Unlike a simple attention mask, the gate is conditioned by warehouse topology and sensing reliability rather than feature magnitude alone. A lightly occupied but strategically important crossing node can therefore retain high edge salience when lane conflicts indicate possible spillover. Their interaction is formalized by the fusion equation below.

$$F_t = \phi_s(X_t) + E_t \odot \phi_e(X_t) + (1 - E_t) \odot \phi_c(X_t) \quad (2)$$

3.3. Congestion State Encoding

The fused tensor F_t is encoded into two complementary targets. The first target is a continuous congestion intensity map Y_t^h for horizon h, which estimates expected local pressure. The second target is a binary warning map B_t^h, which supports operational alarms. The continuous target combines queue length, local occupancy, expected service delay, and route-conflict pressure after robust normalization. The warning map becomes positive when the continuous value exceeds the horizon-specific threshold learned from delay distributions. The target construction is captured by the next equation.

$$Y_t^h(i,j) = \lambda_1 Q_t^h(i,j) + \lambda_2 O_t^h(i,j) + \lambda_3 D_t^h(i,j) + \lambda_4 C_t^h(i,j) \quad (3)$$

The weighting equation uses four coefficients identified on the training split and then fixed during model comparison, which prevents the proposed network from receiving a target-definition advantage. A 5 min horizon emphasizes queue growth and local vehicle interactions, while a 15 min horizon places more weight on order-release pressure and service-delay accumulation. This two-target design avoids forcing the model to choose between regression and warning classification. It allows the same backbone to produce a smooth congestion field for planning and a sharp warning layer for dispatch intervention.

4. VMamba-UNet Congestion Prediction Model

4.1. Network Architecture

Figure 2 presents the architecture of the proposed VMamba-UNet. The input tensor first passes through the edge-adaptive fusion module described in Section 3, then enters a four-stage encoder with patch embedding, VMamba selective scanning blocks, and stride-based downsampling. Skip connections carry edge-preserved features to a symmetric decoder. Each decoder stage contains a VMamba block followed by a lightweight convolutional refinement layer, which restores local boundary detail after selective scanning. The final head produces one continuous map and one warning-probability map for each prediction horizon.

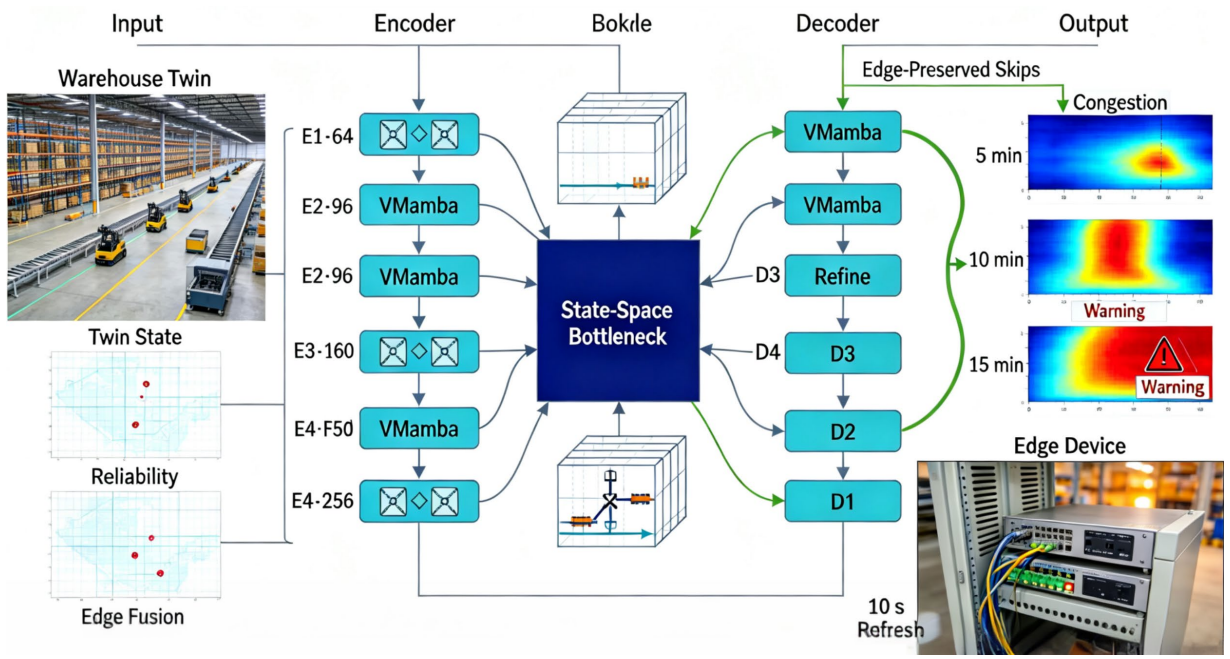


Figure 2. VMamba-UNet architecture for multi-horizon warehouse congestion prediction

The widths of the four stages in the encoder are 64, 96, 160 and 256. All four modules have a Normalization layer, a Selective Scan block, a Depthwise Local Mixing layer and a Residual Projection. The decoder has the same number of stages but fewer refinement blocks; the skip features already contain high-resolution boundary information. Multi-horizon prediction is conducted via shared decoder features and horizon-specific heads. Thus, no three separate models need to be trained, memory at the edge is reduced, and predicted maps from various horizons can be compared. The warning head receives the same decoded representation as the regression head, but its final layer is independently calibrated to change the probability threshold without changing the continuous congestion scale.

VMamba blocks extend the encoder receptive field through directional state propagation rather than dense pairwise attention. Horizontal, vertical, and reverse scans carry workload information along long aisles and between crossing nodes, buffer exits, and sorter interfaces. Input-conditioned state updates determine which features persist along each direction, while depthwise local mixing restores neighborhood detail. UNet skip connections serve a complementary role by transferring high-resolution, edge-preserved features to the decoder. Selective scanning therefore models remote pressure propagation, whereas the skip pathway limits boundary displacement during reconstruction; neither mechanism is presented as a substitute for the other.

4.2. Selective Scanning and Training Objective

For a flattened scan sequence derived from a warehouse state feature map, each VMamba block updates a hidden state with input-conditioned transition and input matrices. The operation is applied in horizontal, vertical, and reverse directions, and the directional outputs are merged before decoder refinement. For clarity, the selective state update is written in the following equation.

$$s_k = A(x_k)s_{k-1} + B(x_k)x_k, z_k = C(x_k)s_k + Dx_k \tag{4}$$

The selective-scan equation shows how the current feature x_k controls state propagation through $A(x_k), B(x_k)$, and $C(x_k)$, while the direct term Dx_k preserves an unfiltered local contribution. The training objective then combines regression error, warning classification loss, and edge consistency. Huber loss reduces sensitivity to rare sensor spikes, focal loss addresses the lower frequency of positive congestion cells, and the gradient term penalizes disagreement between predicted and target congestion boundaries. The resulting terms enter the objective equation shown below.

$$\mathcal{L} = \mathcal{L}_{Huber}(\hat{Y}, Y) + \eta\mathcal{L}_{Focal}(\hat{B}, B) + \mu\|\nabla\hat{Y} - \nabla Y\|_1 \tag{5}$$

4.3. Edge Deployment Workflow

The Deployment Workflow is for edge execution near the warehouse control system. Sensor streams are buffered in a rolling time window, mapped to the twin grid, and missing values are checked before inference. When the camera or telemetry channel is delayed, the reliability map is updated to reduce the weight of the old local observations by the fusion module. At each time step, the model outputs a congestion map, and then the control layer selects areas where

the predicted probability exceeds the warning threshold. The same output can be shown in the future congestion front on the digital twin map by the dashboard.

The inference service has a small cache of the previous fused state and the previous warning map. The cache is not used to reduce the new congestion; instead, a change detection mechanism is employed to avoid issuing repeated alerts for a zone that has remained congested without further deterioration. When a predicted alarm occurs in the new area, information about the time window, main contributing routes and local reliability will also be added by the system. Metadata can be employed by the warehouse control system to distinguish between a high-confidence route conflict and a low-confidence warning caused by a lack of camera frames. Therefore, the model will be used to provide forward-looking support for the dispatch system rather than replacing it entirely.

A reasonable application constraint is to complete the prediction well in advance of the next twin refresh cycle. Therefore, the proposed model has a relatively narrow channel and depthwise local refinement, and selective scanning is used instead of dense self-attention. Batch inference is employed for the three horizons, and the warning head shares most features with the regression head. The Design does not have redundant calculations, and the output semantics of continuous load and warning probability are separate.

5. Experimental Data and Analysis

5.1. Dataset Construction and Experimental Settings

Experiments use one simulated automated warehouse twin whose event sequences and state transitions were checked through replay consistency. The twin contains 42 operating zones, 124 grid cells, 58 AGVs, 14 inbound or outbound interfaces, and 9 conveyor-buffer groups. It provides 2.8 million synchronized records at 10 s intervals under normal replenishment, wave picking, dock surge, and mixed exception handling. The evidence therefore represents a controlled single-facility benchmark rather than multi-site production validation. To prevent future workload information from entering model development, records are divided into contiguous chronological blocks: 70% for training, 10% for validation, and 20% for testing [19]. Forecast horizons are 5, 10, and 15 min. The heavy-congestion threshold of 0.68 is derived only from the 85th percentile of local delay in the training partition. Baselines include ConvLSTM-UNet, TCN-UNet, Graph Attention Network, Transformer-UNet, Swin-UNet, and VMamba-UNet without edge-adaptive fusion.

Table 1 shows the main quantitative results of the test set. The mean absolute error (MAE) of the proposed model is 0.061, and its F1-score is 89.7%; it has been improved by 0.022 over Transformer-UNet in MAE and by 7.3 percentage points in the warning F1-score. The boundary F1 score has increased to 83.5% from 74.1%, and the dispatch decisions are now more sensitive in the vicinity of aisle intersections and buffer exits. VMamba-UNet without edge fusion has a 0.070 MAE and an 86.1% warning F1-score; therefore, selective scanning is convenient but does not solve the problem of boundary dilution. Swin-UNet has a good structure but is relatively slow; its 95th-percentile latency is 36.9 ms, and that of the proposed model is 24.6 ms. Percentile latency is employed, in addition to a mean-based indicator, to illustrate the behavior of the tail distribution after the fixed refresh deadline [20]. The baseline of graph attention is based on route relationships and does not retain cell-level map details after interpolating back from the zone-level prediction to the grid.

Table 1. Overall test-set comparison of congestion prediction models

Model	MAE	RMSE	SSIM	Warning F1 (%)	Boundary F1 (%)	P95 Latency (*ms*)
ConvLSTM-UNet	0.096	0.142	0.842	78.3	68.9	29.7
TCN-UNet	0.089	0.134	0.858	80.1	70.5	26.4
Graph Attention	0.092	0.137	0.821	81.6	71.7	33.2
Transformer-UNet	0.083	0.126	0.881	82.4	74.1	41.8
Swin-UNet	0.079	0.119	0.889	84.8	76.3	36.9
VMamba-UNet	0.070	0.106	0.902	86.1	77.2	23.9
Proposed Model	0.061	0.094	0.919	89.7	83.5	24.6

All models receive the same input channels, chronological partitions, forecast horizons, and target definitions. The first 12 h of each replay day are excluded to remove empty-queue warm-up and incomplete rolling windows. Continuous features are normalized with training-partition statistics only, while bounded utilization and probability channels remain in the [0, 1] interval. Random time masking is applied only during training. Evaluation covers regression error, spatial similarity, warning classification, and boundary quality because the forecast must support both map reconstruction and discrete intervention [21]. Reported metrics are MAE, RMSE, SSIM, warning F1-score, boundary F1-score, and P95 latency. Hyperparameters are selected from validation MAE and warning F1-score; the test partition is used once for

final comparison. Every latency benchmark uses batch size one, identical input tensors, the same execution environment, 100 warm-up inferences, and 1,000 timed runs. The models are trained with AdamW, cosine learning-rate decay, and early stopping after 18 validation epochs without improvement.

5.2. Comparative Prediction Performance

Figure 3(a) compares horizon-wise MAE for the five principal model families. At the 5 min horizon, the proposed model records 0.047 MAE, while Transformer-UNet and Swin-UNet record 0.066 and 0.063; at 15 min, the proposed error rises to 0.078, whereas ConvLSTM-UNet reaches 0.121. The proposed-to-ConvLSTM gap consequently expands from 0.031 to 0.043 as the forecast interval lengthens. This divergence indicates that directional state propagation becomes more valuable when pressure must travel from docks and sorting buffers into remote aisles. Figure 3(b) reports warning precision, recall, and F1-score. The proposed model obtains 90.8% precision, 88.7% recall, and 89.7% F1-score, giving a comparatively balanced operating point rather than improving precision by suppressing alarms. Its 3.9 percentage-point recall advantage over VMamba-UNet without edge fusion is consistent with earlier detection of thin spillover fronts before they occupy broad grid regions.

Figure 3(c) plots boundary error against local vehicle density for 1,200 sampled congestion fronts. The proposed model keeps the median error below 1.6 cells when density is under 0.72 vehicles per active lane cell, while Transformer-UNet increases from 1.9 to 3.8 cells across the displayed density range. The widening separation at high density shows that the improvement is concentrated where interacting routes make boundary placement difficult, rather than being a uniform offset. A one-cell displacement can move a warning from an aisle throat into an adjacent storage region, so this reduction directly improves the spatial validity of rerouting decisions. Figure 3(d) shows the per-zone inference-latency distribution under edge-device emulation. The proposed model has an interquartile range of 18.4-22.1 ms and a 95th-percentile latency of 24.6 ms, whereas Transformer-UNet reaches 41.8 ms at the same percentile. The narrower upper tail leaves a larger and more predictable margin inside the 10 s twin refresh cycle for data validation, rule checking, and dashboard transmission.

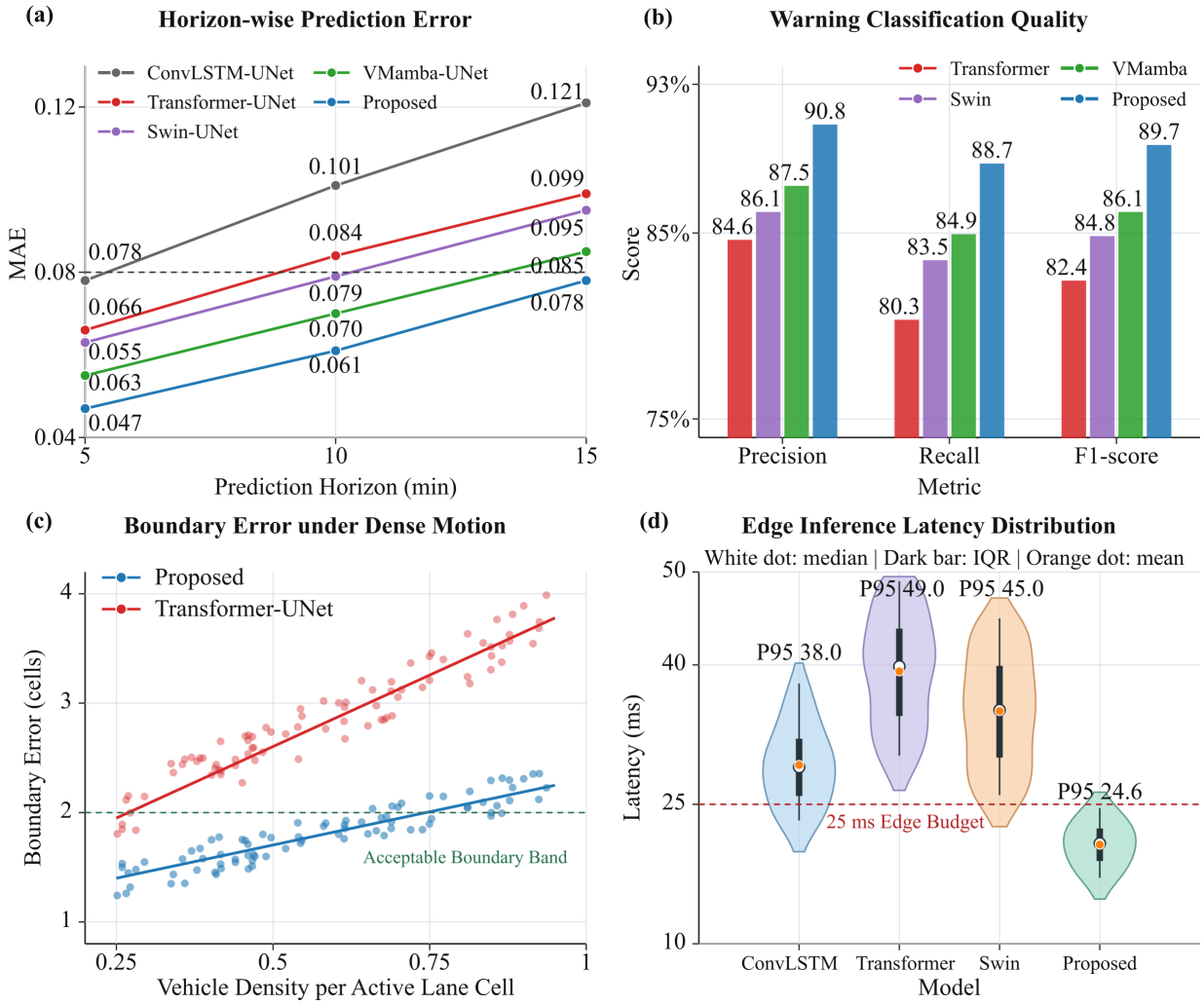


Figure 3. Comparative prediction performance: (a) horizon-wise MAE; (b) warning precision, recall, and F1-score; (c) boundary error versus vehicle density; (d) edge inference latency distribution

Across the four performance views, the gains in map regression and warning classification arise from related but distinguishable mechanisms. At short horizons, local queue and occupancy cues dominate, which limits the difference between globally aware models. At longer horizons, remote workload pressure must be carried across several operational zones, and selective scanning prevents the signal from decaying into a broad average. Edge-adaptive fusion affects a different failure mode: it preserves the location of the congestion front after downsampling. Their interaction explains why the proposed model improves both 15 min MAE and boundary F1-score without paying the latency penalty of dense attention. The largest practical separation occurs during dock surges, where receiving-buffer pressure moves through cross-aisles before the downstream grid becomes densely occupied.

Residual errors fall into two operational categories. Most false negatives arise from short manual exceptions that do not immediately produce a vehicle-density or queue-growth signal. A temporary quality inspection at a sorter entrance, for example, can halt pallet discharge before the route-based channels register sustained accumulation. False positives occur more often at cross-aisles during wave picking, where vehicles decelerate for a planned turn but clear the crossing before a queue develops. These errors usually persist for fewer than three refresh cycles. The proposed model reduces repeated alarms when reliable observations show deceleration without queue growth, but this result should be interpreted as improved spatial decision support rather than proof of closed-loop rerouting performance. Baseline behavior also varies by horizon: ConvLSTM-UNet remains competitive when short-term accumulation is locally visible, Transformer-UNet captures remote dependencies but tends to produce broad warning regions, and Swin-UNet shows mild discontinuity when a queue front crosses a window boundary. Directional scanning and edge-preserved skip features jointly reduce these errors, which explains why the gain in boundary F1-score is larger than the gain in SSIM.

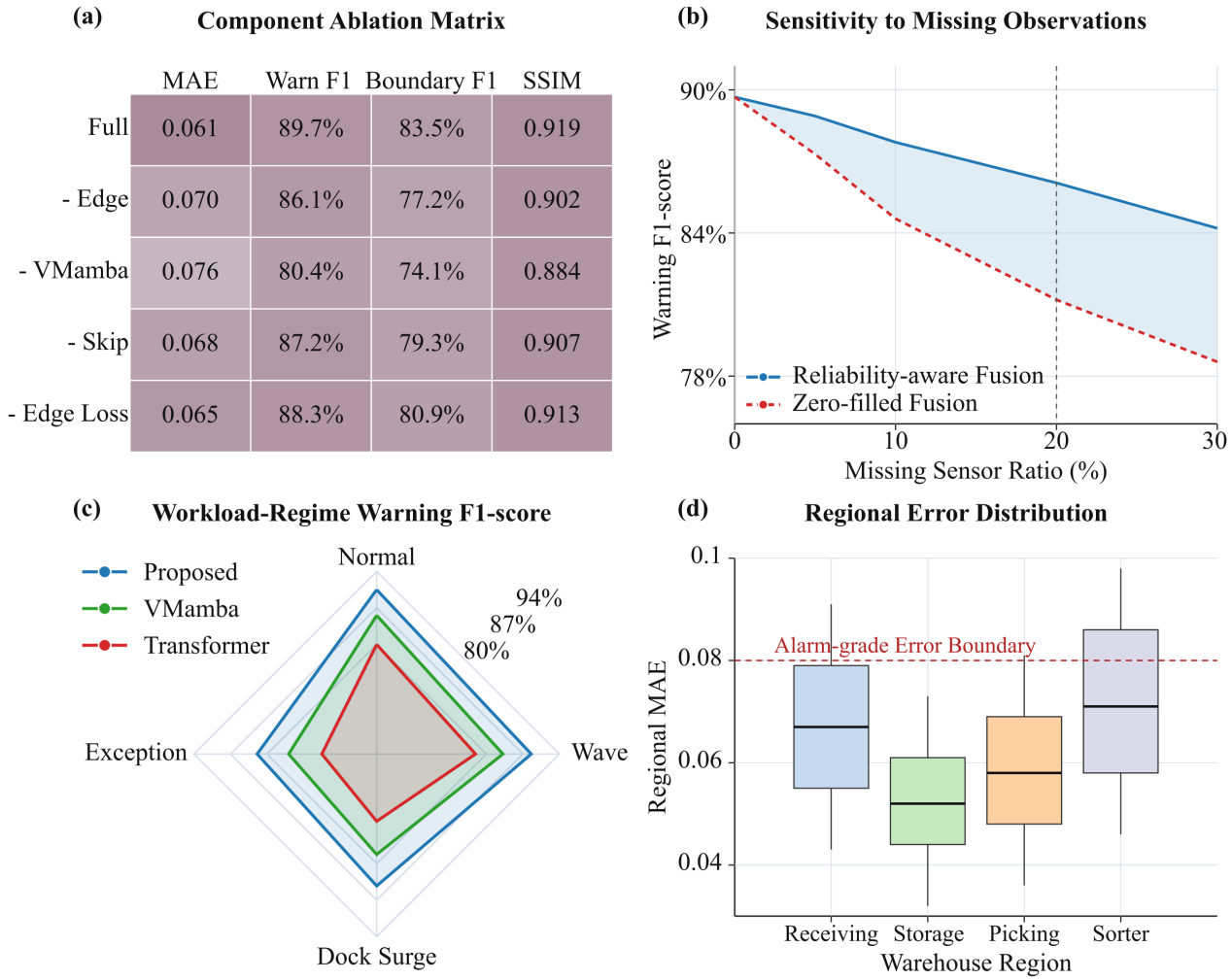


Figure 4. Ablation and robustness analysis: (a) component ablation matrix; (b) sensitivity to missing sensor ratio; (c) workload-regime warning performance; (d) regional error distribution

5.3. Ablation, Robustness, and Sensitivity Analysis

Figure 4(a) presents the ablation matrix across edge fusion, VMamba scanning, skip refinement, and edge consistency loss. Removing edge-adaptive fusion increases MAE from 0.061 to 0.070 and reduces boundary F1-score from 83.5% to 77.2%, whereas removing VMamba blocks produces 0.076 MAE and 80.4% warning F1-score. The different degradation patterns separate local boundary preservation from long-range propagation: edge fusion has the larger effect

on boundary quality, while VMamba removal causes the larger loss in field-level forecasting. Removing the edge consistency term decreases boundary F1-score by 2.6 percentage points, indicating that auxiliary gradient supervision remains useful after the architecture already preserves edges. Treating unavailable sensor channels as structured uncertainty rather than ordinary zero values is a recognized strategy for avoiding measurement-induced decision bias [22]. Figure 4(b) tests this property from 0% to 30% missing observations. Reliability-aware fusion declines from 89.7% to 84.2% F1-score, while zero-filled fusion falls to 78.6%; most of the separation appears beyond the 10% missing level, where imputation ambiguity begins to dominate local occupancy evidence.

Figure 4(c) displays warning F1-score under normal replenishment, wave picking, dock surge, and exception handling. The proposed model reaches 91.8%, 90.6%, 87.9%, and 86.4%, respectively. The 5.4 percentage-point span between normal and exception operation is smaller than the corresponding spans of the VMamba and Transformer variants, showing that the advantage is retained under abrupt disturbances rather than being confined to regular flow. Figure 4(d) gives the error distribution across warehouse regions. Storage aisles have the lowest median MAE at 0.052, while receiving buffers and sorter interfaces reach 0.067 and 0.071 and also exhibit wider interquartile ranges. Prior material-handling studies identify buffer interfaces as locations where release rules and service-time variability concentrate operational variance [23]. In the present results, the proposed model narrows the dispersion relative to Swin-UNet and Graph Attention Network in every region, but the residual spread at sorter interfaces indicates that abrupt equipment and manual events remain incompletely represented. Probability calibration is necessary when the same warning model must support different intervention costs and alarm tolerances [24]. Figure 5 examines warning calibration and horizon reliability in a single multilayer plot.

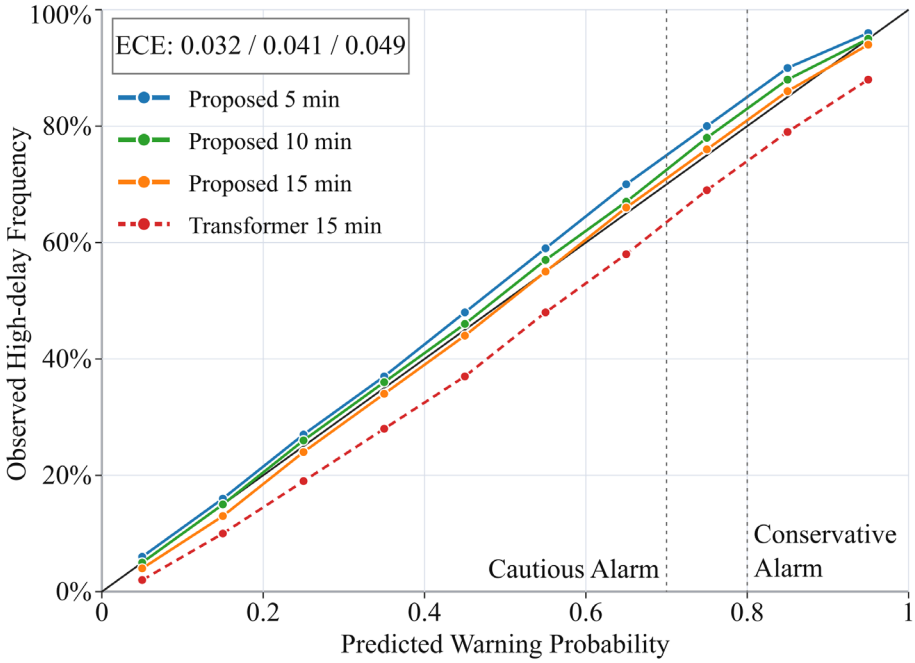


Figure 5. Warning probability calibration and horizon reliability

The proposed model has expected calibration errors of 0.032, 0.041, and 0.049 at the 5, 10, and 15 min horizons, compared with 0.064, 0.077, and 0.092 for Transformer-UNet. Error increases with horizon as state uncertainty accumulates, but the proposed curves remain closer to the diagonal. At a 0.70 operating threshold, the model identifies 88.1% of high-delay cells with an 8.6% false-alarm rate; increasing the threshold to 0.80 lowers recall to 81.4% and false alarms to 5.1%. These point estimates describe the fixed replay test set and should not be interpreted as cross-site confidence intervals. Their operational value lies in showing a controlled precision-recall adjustment around the training-derived reference threshold of 0.68.

Stress tests based on route-block events and sensor-delay bursts are commonly used to examine whether a predictive twin degrades smoothly under partial observation [25]. Under a 12 min route block near a storage cross-aisle, the proposed model raises downstream-buffer warning probability from 0.42 to 0.76 within two prediction cycles, while ConvLSTM-UNet reaches 0.61. The 0.15 probability separation is large enough to cross the 0.68 operating threshold only for the proposed model at that time. During a camera-delay burst affecting 18% of grid cells, reliability-aware fusion keeps MAE at 0.074, compared with 0.091 when the same model receives zero-filled channels without reliability gating. The smaller degradation shows that the fusion module moderates low-confidence evidence rather than merely sharpening boundaries under complete sensing. Sensitivity to the congestion threshold was evaluated from 0.60 to 0.76. At the lower threshold, all models improve recall but produce more false positives around planned vehicle slowdowns.

At the higher threshold, precision improves while early dock-surge warnings become harder to capture. The proposed model maintains the most stable F1-score across this interval, with a maximum variation of 3.4 percentage points, compared with 6.1 percentage points for Transformer-UNet and 5.7 percentage points for Swin-UNet. This stability is useful for practical deployment because each warehouse may choose a different alarm threshold depending on staffing level, tolerance for route changes, and service-level agreements. A model with fragile threshold behavior would require frequent recalibration after workload changes.

5.4. Operational Interpretation and Deployment Efficiency

Figure 6(a) ranks the zones with the largest predicted delay reduction after warning-triggered route reallocation. Receiving Buffer 2 decreases from 184 s to 137 s, and Cross-Aisle 3 decreases from 162 s to 124 s; the larger reductions occur upstream of several competing routes, where an early diversion prevents secondary queues. Figure 6(b) compares rerouting side effects. The proposed warning map records 11.8% unnecessary detours and 8.6% false alarms, below the corresponding values of the three baselines, which means the delay reduction is not purchased by indiscriminate route changes. Human-supervised exception stations and sorter interfaces are known to introduce timing variation that is only partly observable from automated telemetry [26]. Figure 6(c) localizes this limitation: the highest uncertainty clusters appear around the sorter interface and temporary exception area, while regular storage corridors remain comparatively certain. The spatial pattern connects uncertainty to missing operational context rather than to uniform model indecision.

Figure 6(d) reports the relationship between inference budget and map quality under channel-width scaling. Reducing the base width from 64 to 48 lowers 95th-percentile latency from 24.6 ms to 19.7 ms but increases MAE from 0.061 to 0.066. Increasing it to 80 improves MAE only to 0.059 while raising latency to 31.5 ms. The 64-channel point lies near the curve knee: moving to 80 channels buys 0.002 MAE at a cost of 6.9 ms, whereas moving to 48 saves 4.9 ms but loses 0.005 MAE. Source decomposition is useful for checking whether a predictor responds to physically plausible causes across operating regimes [27]. Figure 6(e) shows that wave-picking congestion is led by order-release pressure at 36% and route conflict at 27%, whereas dock-surge congestion is led by buffer occupancy at 41% and service delay at 31%. The shift in dominant components indicates that the shared model is not relying on one invariant proxy for congestion.

Ranked capture curves evaluate whether a limited intervention budget is directed toward the cells with the greatest future delay [28]. Figure 6(f) compares cumulative high-delay capture over the top-ranked warning cells. The proposed model captures 71.3% of future high-delay cells in the top 20% of warnings, versus 62.8% for Swin-UNet and 54.6% for ConvLSTM-UNet. At a 40% warning budget, the proposed capture reaches 91.2%, so the second 20% of inspected cells contributes a further 19.9 percentage points. The diminishing increment after 40% indicates that the ranking concentrates risk near the top instead of distributing high-delay cells uniformly across the warning list. This property is particularly relevant when controllers can reroute only a small set of vehicles or inspect only a few zones during each refresh cycle. The deployment efficiency analysis also shows that the model can support different operating policies without retraining. A conservative site can use only the top 10% of warning cells and still capture 45.2% of future high-delay cells, which is suitable when route changes are costly. A more aggressive site can act on the top 30% and capture 84.5% of high-delay cells, which is suitable during short dock surges or planned wave releases. The continuous congestion map provides additional context for these choices because it shows whether a warning is isolated, expanding, or connected to an upstream buffer. This combination of ranked warning and continuous intensity makes the output usable for both automated rule checks and human supervision.

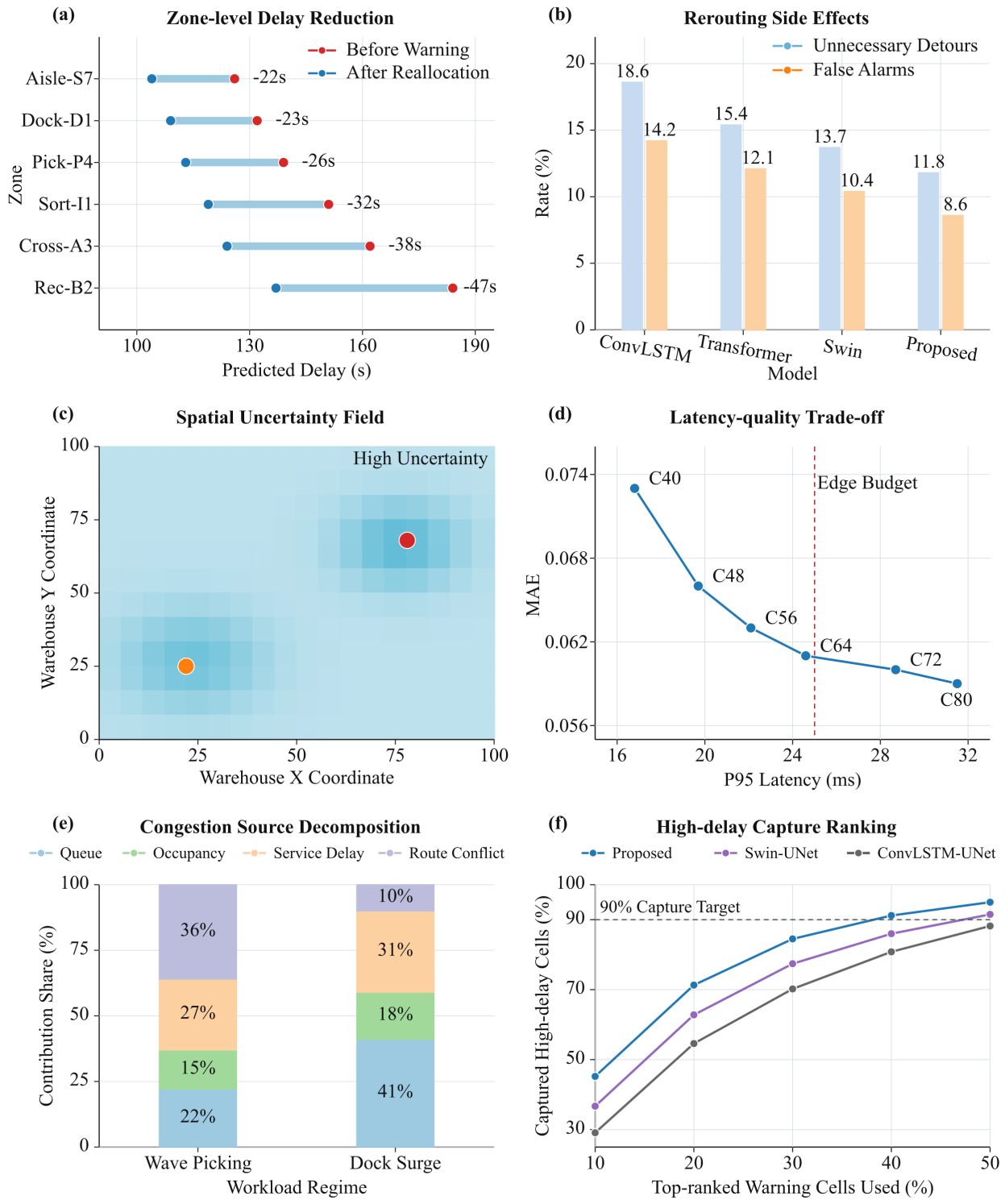


Figure 6. Operational deployment interpretation: (a) zone-level delay reduction; (b) rerouting effect; (c) uncertainty field; (d) latency-quality trade-off; (e) congestion source decomposition; (f) high-delay capture ranking

Edge deployment is useful only when prediction finishes before the next dispatch decision. With a 24.6 ms P95 inference time, the selected configuration leaves more than 9.5 s of the 10 s refresh period for state validation, rule evaluation, dashboard transmission, and command arbitration. This timing margin is measured for model inference within the stated benchmark protocol; it is not a guarantee of end-to-end deterministic control under every facility load. Residual errors concentrate in events that change faster than the synchronous feature window. Sorter interruptions can begin between updates, manual quality checks may omit structured completion events, and temporary right-of-way decisions can alter AGV trajectories without changing the planned route graph. These cases can delay a warning because upstream pressure develops before the relevant event is encoded. Planned deceleration creates the converse error: speed drops can temporarily raise the warning score even when the crossing clears without a queue. Event-completion signals and a short asynchronous update path for exception stations are therefore more direct remedies than increasing backbone

width. The latency-quality curve shows only marginal accuracy improvement beyond 64 channels, while online event integration targets the remaining information gap. The present system supports the downstream controller with forecasts and explanations; it does not autonomously optimize dispatch.

6. Conclusion

This study developed and evaluated an edge-adaptive VMamba-UNet for multi-horizon congestion prediction in an automated warehouse digital twin. The results show that topology- and reliability-conditioned boundary fusion complements selective state-space scanning by preserving local operational edges while propagating workload pressure across distant zones.

1. On the 42-zone, 2.8-million-record benchmark, the model achieved an MAE of 0.061, a warning F1-score of 89.7%, and a boundary F1-score of 83.5%. Compared with Transformer-UNet, it reduced MAE by 26.5%, increased warning F1 by 7.3 percentage points, and reduced P95 inference latency from 41.8 to 24.6 ms under the same benchmark protocol.

2. The advantage increased with forecasting difficulty. At 15 min, MAE was 0.078 compared with 0.121 for ConvLSTM-UNet. Removing edge-adaptive fusion increased MAE to 0.070 and reduced boundary F1 to 77.2%. Under 30% missing observations, reliability-aware fusion retained an F1-score of 84.2%, whereas zero-filled fusion reached 78.6%. These comparisons support distinct roles for long-range state propagation and reliability-conditioned boundary preservation.

3. The principal innovation is to treat operational topology and sensor reliability as inputs to a linear-complexity encoder-decoder rather than reconstructing boundaries only at the output. The evidence is limited to a simulated and replay-validated single-facility twin with a fixed grid, a 10 s refresh rate, and thresholds learned from historical delays. Manual exceptions and sorter interruptions remain difficult, and the model predicts congestion without directly optimizing downstream dispatch.

The work provides a practical basis for spatially precise, real-time warning support in warehouse twins. Future research should validate transfer across facilities, adapt online to layout and workload changes, integrate asynchronous exception events, and evaluate prediction-control co-design in closed-loop operations.

7. Declarations

7.1 Author Contributions

Szymon Kurowski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

7.2 Funding and Acknowledgments

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

7.3 Institutional Review Board Statement

Not applicable.

7.4 Informed Consent Statement

Not applicable.

7.5 Data Availability

Not applicable.

7.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

8. References

[1] Ellithy, K., Salah, M., Fahim, I. S., & Shalaby, R. (2024). AGV and Industry 4.0 in warehouses: A comprehensive analysis of existing literature and an innovative framework for flexible automation. The International Journal of Advanced Manufacturing Technology, 134(1-2), 15–38. https://doi.org/10.1007/s00170-024-14127-0

[2] Zhao, Z., Cheng, J., Liang, J., Liu, S., Zhou, M., & Al-Turki, Y. (2024). Order picking optimization in smart warehouses with human-robot collaboration. IEEE Internet of Things Journal, 11(9), 16314–16324. https://doi.org/10.1109/jiot.2024.3352658

[3] Lv, Z., Qiao, L., Mardani, A., & Lv, H. (2024). Digital twins on the resilience of supply chain under COVID-19 pandemic. IEEE Transactions on Engineering Management, 71, 10522–10533. https://doi.org/10.1109/tem.2022.3195903

[4] Ferrari, A., Carlin, A., Rafele, C., & Zenezini, G. (2024). A method for developing and validating simulation models for automated storage and retrieval system digital twins. The International Journal of Advanced Manufacturing Technology, 131(11), 5369–5382. https://doi.org/10.1007/s00170-023-12660-y

[5] Cai, Y. (2024). Digital twin model construction for intelligent Internet of Things logistics and warehousing systems. Intelligent Decision Technologies, 18(3), 2407–2420. https://doi.org/10.3233/idt-240324

[6] Li, W., & Law, K. L. E. (2024). Deep learning models for time series forecasting: A review. IEEE Access, 12, 92306–92327. https://doi.org/10.1109/access.2024.3422528

[7] Le, D. P. C., Wang, D., & Le, V. T. (2024). A comprehensive survey of recent transformers in image, video and diffusion models. Computers, Materials & Continua, 80(1), 37–60. https://doi.org/10.32604/cmc.2024.050790

[8] Huang, J., Zhong, A., & Wei, Y. (2024). A new visual state space model for low-dose CT denoising. Medical Physics, 51(12), 8851–8864. https://doi.org/10.1002/mp.17387

[9] Wang, H., Liu, H., Huo, L., Chen, L., & He, Q. (2025). DBENet: Dual-branch encoder network based on the visual state space model for semantic segmentation of remote sensing image. International Journal of Remote Sensing, 46(20), 7684–7706. https://doi.org/10.1080/01431161.2025.2557591

[10] Zhou, Q., Wang, L., Gao, G., Kang, B., Ou, W., & Lu, H. (2024). Boundary-guided lightweight semantic segmentation with multi-scale semantic context. IEEE Transactions on Multimedia, 26, 7887–7900. https://doi.org/10.1109/tmm.2024.3372835

[11] Neugebauer, J., Heilig, L., & Voß, S. (2024). Digital twins in the context of seaports and terminal facilities. Flexible Services and Manufacturing Journal, 36(3), 821–917. https://doi.org/10.1007/s10696-023-09515-9

[12] de Assis, R. F., Faria, A. F., Thomasset-Laperrière, V., Santa-Eulalia, L. A., Ouhimmou, M., & Ferreira, W. D. P. (2024). Machine learning in warehouse management: A survey. Procedia Computer Science, 232, 2790–2799. https://doi.org/10.1016/j.procs.2024.02.096

[13] Eder, M. (2024). Analytical examination of shuttle-based storage and retrieval systems with multiple-capacity lifts. The International Journal of Advanced Manufacturing Technology, 133(9-10), 5053–5064. https://doi.org/10.1007/s00170-024-14058-w

[14] Hu, N., Zhang, D., Xie, K., Liang, W., Li, K. C., & Zomaya, A. Y. (2024). Dynamic multi-scale spatial-temporal graph convolutional network for traffic flow prediction. Future Generation Computer Systems, 158, 323–332. https://doi.org/10.1016/j.future.2024.04.052

[15] Ren, S., & Wang, Y. (2024). MTC-Net: Multi-scale feature fusion network for medical image segmentation. Journal of Intelligent & Fuzzy Systems, 46(4), 8729–8740. https://doi.org/10.3233/jifs-237963

[16] Papa, L., Russo, P., Amerini, I., & Zhou, L. (2024). A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7682–7700. https://doi.org/10.1109/tpami.2024.3392941

[17] Zhu, Y., Zhang, D., Lin, Y., Feng, Y., & Tang, J. (2025). Merging context clustering with visual state space models for medical image segmentation. IEEE Transactions on Medical Imaging, 44(5), 2131–2142. https://doi.org/10.1109/tmi.2025.3525673

[18] Zhou, T., Zhang, Y., Chen, G., Zhou, Y., Wu, Y., & Fan, D. P. (2025). Edge-aware feature aggregation network for polyp segmentation. Machine Intelligence Research, 22(1), 101–116. https://doi.org/10.1007/s11633-023-1479-8

[19] Wang, P., Gurmani, S. H., Tao, Z., Liu, J., & Chen, H. (2024). Interval time series forecasting: A systematic literature review. Journal of Forecasting, 43(2), 249–285. https://doi.org/10.1002/for.3024

[20] Yuan, M., Zhang, L., Duan, D., Zeng, L., Song, M. H., Li, Z., Xing, G., & Li, X. Y. (2025). Mitigating tail latency for on-device inference with load-balanced heterogeneous models. IEEE Transactions on Mobile Computing, 24(12), 13090–13105. https://doi.org/10.1109/tmc.2025.3588430

[21] Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Muhammad Marier, S., Liu, Y., Chen, Y., & Xia, F. (2025). Deep learning for

time series forecasting: A survey. International Journal of Machine Learning and Cybernetics, 16(7-8), 5079–5112. https://doi.org/10.1007/s13042-025-02560-w

[22] Xiao, Y., Zhang, P., & Li, Q. (2025). Fusion Diffusion Mamba for missing data reconstruction. IEEE Access, 13, 199051–199063. https://doi.org/10.1109/access.2025.3631133

[23] Boysen, N., Stephan, K., & Schwerdfeger, S. (2024). Order consolidation in warehouses: The loop sorter scheduling problem. European Journal of Operational Research, 316(2), 459–472. https://doi.org/10.1016/j.ejor.2024.02.042

[24] Balanya, S. A., Maroñas, J., & Ramos, D. (2024). Adaptive temperature scaling for robust calibration of deep neural networks. Neural Computing and Applications, 36(14), 8073–8095. https://doi.org/10.1007/s00521-024-09505-4

[25] Braik, A. M., & Koliou, M. (2024). A digital twin framework for efficient electric power restoration and resilient recovery in the aftermath of hurricanes considering the interdependencies with road network and essential facilities. Resilient Cities and Structures, 3(3), 79–91. https://doi.org/10.1016/j.rcns.2024.07.004

[26] Koreis, J., Loske, D., & Klumpp, M. (2025). Together, we travel: Empirical insights on human-robot collaborative order picking for retail warehousing. The International Journal of Logistics Management, 36(1), 1–20. https://doi.org/10.1108/ijlm-03-2023-0127

[27] Ghahfarokhi, A. F., Schäfer, J., Wagner, M. F., & Dorronsoro, B. (2025). Explainable artificial intelligence for time series using attention mechanism: Application to wind turbine fault detection. IEEE Access, 13, 180613–180631. https://doi.org/10.1109/access.2025.3621003

[28] Gaudreault, J. G., & Branco, P. (2024). Empirical analysis of performance assessment for imbalanced classification. Machine Learning, 113(8), 5533–5575. https://doi.org/10.1007/s10994-023-06497-5