

Defect-Preserving Token Allocation for Industrial Surface Inspection

Authors

Liam Carter^{1*}, Emily Foster², Daniel Hughes³

Affiliations

School of Computer Science, Faculty of Engineering, The University of Sydney, J12, Sydney, NSW 2006, Australia.

*Corresponding author: liam.carter@sydney.edu.au

Abstract

Industrial surface inspection requires fast and reliable recognition of small defects such as scratches, pits, stains, cracks, and coating bubbles. Vision Transformer models can capture long-range visual context, but their full-token computation is often unnecessary for large uniform regions on steel plates, fabrics, printed circuit boards, and coated surfaces. This study investigates defect-preserving token allocation for efficient industrial inspection. We propose the Defect Token Allocation Network (DTAN), which assigns computational priority according to local anomaly contrast, patch-level uncertainty, and cross-layer defect persistence. Instead of using a fixed pruning ratio, DTAN keeps suspected defect patches at full attention depth, compresses repetitive background patches, and rechecks borderline patches through a lightweight verification branch. The study used inspection data collected from 14 production lines, including 286,000 high-resolution images and 412,600 annotated defect instances. Defect sizes ranged from 6×8 pixels to 420×360 pixels, which allowed evaluation on both tiny and large surface anomalies. On the NEU-Surface, DAGM, KolektorSDD2, and internal factory datasets, DTAN reached 91.8% defect-level F1-score while reducing transformer computation by 33.9%. For defects smaller than 32×32 pixels, recall remained at 88.6%, compared with 82.4% for a standard token-pruning model. In online testing, average inspection latency decreased from 47.3 ms to 29.5 ms per image, allowing stable processing at 1.6 m/s conveyor speed. False alarms caused by repetitive texture were reduced by 21.7%. These results show that defect-aware token allocation can improve the deployment efficiency of transformer inspection models without losing small industrial defects.

Keywords: Industrial inspection; surface defect detection; Vision Transformer; token allocation; efficient inference; anomaly localization; quality control

1. Introduction

Surface defect inspection is an essential component of industrial quality control because even small scratches, pits, cracks, stains, inclusions, and coating irregularities may influence product appearance, mechanical performance, subsequent processing, and final product reliability. With the increasing automation of manufacturing systems, machine-vision inspection has gradually

replaced part of manual inspection in steel production, textiles, printed circuit boards, semiconductor components, coated materials, and other surface-sensitive products. Compared with manual inspection, automated systems provide higher consistency and can support continuous monitoring at production-line speed. However, reliable defect inspection remains difficult under practical factory conditions. Many defects occupy only a very small proportion of the image, whereas most regions contain normal material surfaces with uniform, periodic, or repetitive textures. Variations in illumination, surface reflectance, imaging distance, production speed, defect scale, and background pattern can further increase intra-class variation and reduce the visual contrast between defective and normal regions. Convolutional neural networks have been widely applied to this problem because of their ability to capture local edges, texture changes, and morphological characteristics [1]. Nevertheless, convolution-based architectures mainly aggregate information within local receptive fields, which may limit their ability to represent wider structural relationships when defects must be distinguished from large-scale or repetitive background patterns. Vision Transformer architectures provide an alternative approach by modeling interactions between spatially separated image regions through self-attention. This capability has led to increasing interest in transformer-based methods for industrial defect recognition, localization, segmentation, and anomaly detection [2,3]. Related research on adaptive intelligent systems has also shown that complex computational tasks can benefit from dividing processing responsibilities among specialized components and iteratively optimizing intermediate results rather than applying an identical processing procedure to all information [4]. This principle is relevant to visual inspection because different image regions do not contribute equally to defect recognition. In industrial images, a few local patches may contain critical defect evidence, while a large number of background patches mainly describe repeated surface patterns. Recent industrial anomaly-detection methods have further improved recognition performance under limited or imbalanced defect samples. PatchCore constructs a memory bank of representative normal features for nearest-neighbor anomaly estimation [5,6], while UniAD uses a unified reconstruction-based framework for multiple industrial anomaly categories [7,8]. DeSTSeg combines teacher-student representation learning with segmentation to improve pixel-level anomaly localization [9], and SimpleNet reduces architectural complexity while maintaining competitive anomaly-detection performance [10]. WinCLIP additionally explores vision-language representations for zero-shot and few-shot industrial anomaly detection [11,12]. These developments have substantially improved the ability of learning-based systems to detect previously unseen defects, yet most of them primarily optimize detection accuracy or anomaly localization quality rather than explicitly controlling how computation is distributed across individual image regions.

Computational efficiency is particularly important for industrial inspection because production systems must process continuous image streams under strict cycle-time requirements [13,14]. A model that achieves high accuracy but requires excessive computational resources may

be unsuitable for high-speed production lines, especially when several cameras operate simultaneously or inspection must be performed on edge devices. EfficientAD demonstrated that industrial anomaly-detection models should be evaluated not only in terms of detection performance but also with respect to inference efficiency and practical deployment cost [15,16]. However, conventional CNNs and standard Vision Transformers generally apply approximately the same computational procedure to all spatial locations. This uniform processing strategy is inefficient when most patches correspond to visually redundant normal backgrounds [17]. More importantly, the computational imbalance becomes greater for high-resolution industrial images, where increasing image resolution improves the visibility of small defects but also produces a much larger number of tokens. Consequently, a direct application of transformer models can substantially increase memory consumption and self-attention cost even though only a small proportion of tokens may be relevant to defect recognition. Recent studies on efficient Vision Transformers have shown that substantial computation can be reduced by dynamically selecting, pruning, merging, or terminating image tokens. DynamicViT progressively removes low-importance tokens during inference according to learned importance scores [18,19]. A-ViT introduces adaptive token computation by allowing individual tokens to exit at different transformer layers [20]. EViT preserves informative tokens while reorganizing less important information to reduce redundant processing [21], whereas Token Merging reduces token numbers by combining semantically similar representations instead of directly discarding them [22]. Dynamic Transformer further adjusts model computation according to the estimated difficulty of individual inputs, avoiding a fixed computational path for every image [23,24]. These approaches demonstrate that uniform token processing is not always necessary and that substantial computational savings can be achieved without major accuracy degradation in natural-image recognition tasks. Nevertheless, directly transferring these token-efficiency strategies to industrial defect inspection is not straightforward. Natural-image classification usually relies on relatively large semantic objects, and important image regions often produce strong feature responses at relatively early network stages. Industrial surface defects have different characteristics. A fine crack, pinhole, tiny pit, or localized coating defect may occupy only one or two image patches and may initially produce only weak activation compared with dominant background textures. If such patches are classified as unimportant and removed, merged, or terminated in early transformer layers, subsequent blocks cannot reconstruct the lost defect evidence. This problem is particularly critical for defects near the spatial resolution limit of the imaging system. Moreover, a fixed token-retention ratio does not adequately reflect differences among industrial samples. An image containing a large isolated defect may require relatively little background computation, whereas an image containing several weak defects embedded in highly repetitive textures may require more detailed processing. Surface complexity, defect density, local contrast, and illumination variation may therefore require different computational budgets even within the same production line. Another limitation of many existing industrial inspection studies is the gap between

benchmark evaluation and actual manufacturing conditions. Public datasets are valuable for controlled comparison, but they frequently contain relatively small numbers of images, limited acquisition conditions, and restricted defect distributions. Real production systems may encounter changes in camera exposure, vibration, material batches, roller conditions, surface contamination, and machine speed. Such variations can alter the appearance of normal textures and cause false alarms even when no physical defect is present. At the same time, extremely small defects may become more difficult to preserve after resizing or patch embedding. Conventional evaluation metrics such as overall accuracy, AUROC, or mean average precision may therefore provide an incomplete description of practical inspection performance. For industrial deployment, small-defect recall, texture-induced false-positive rate, inference latency, computational cost, memory consumption, and stability under continuous operation are also important. A useful inspection model should reduce redundant processing while maintaining sensitivity to weak and spatially limited defect evidence.

This study proposes the Defect Token Allocation Network (DTAN), an adaptive transformer framework designed specifically for computationally efficient industrial surface inspection. Instead of applying a predetermined token-pruning ratio, DTAN dynamically allocates computational depth according to local anomaly contrast, token uncertainty, and defect evidence accumulated across transformer layers. Tokens with strong or persistent defect characteristics are preserved for full-depth attention processing so that fine defect information can continue to interact with wider contextual features. Tokens representing highly repetitive and low-risk background regions receive reduced computation, thereby avoiding unnecessary attention operations. For tokens whose early-layer representations remain uncertain, a lightweight verification branch performs additional assessment before their computational paths are reduced. This mechanism is intended to decrease the probability that weak but genuine defects are discarded solely because their early features resemble background textures. The proposed framework is evaluated using 286,000 high-resolution industrial images collected from 14 production lines, containing 412,600 annotated defects covering a wide range of defect sizes and surface conditions. Experiments are conducted on the NEU-Surface, DAGM, KolektorSDD2, and internal factory datasets to examine both cross-dataset performance and production-oriented applicability. The study is designed to determine whether defect-aware token allocation can reduce transformer computation and end-to-end inspection latency while preserving the recall of small and low-contrast defects. It also investigates whether adaptive computation can reduce false alarms caused by repetitive surface textures and remain stable when imaging and production conditions vary. Beyond improving model efficiency, the broader objective is to establish a more suitable computational principle for industrial vision: computational resources should be concentrated on image regions with high defect risk rather than distributed uniformly across all tokens. Such a strategy may provide a practical route toward deploying transformer-based

inspection systems on high-resolution, high-throughput production lines where detection sensitivity, computational cost, and real-time performance must be considered simultaneously.

2. Materials and Methods

2.1 Image Samples and Inspection Conditions

The study used 286,000 high-resolution surface images collected from 14 industrial production lines. The samples included steel plates, coated materials, fabrics, and printed circuit boards. A total of 412,600 defect instances were labeled, including scratches, pits, stains, cracks, coating bubbles, and other surface defects. Defect sizes ranged from 6×8 pixels to 420×360 pixels. This range covered both small and large defects. Images were collected under normal production conditions using fixed industrial cameras and controlled lighting. The NEU-Surface, DAGM, and KolektorSDD2 datasets were also used for testing. Images with severe blur, incomplete views, or invalid labels were removed. The remaining samples were divided into training, validation, and test sets by production batch. Images from the same batch were kept within the same subset.

2.2 Experimental Design and Control Groups

DTAN was compared with a standard Vision Transformer and a fixed token-pruning model under the same experimental conditions. The standard Vision Transformer retained all image tokens and served as the full-computation baseline. The fixed pruning model removed a preset proportion of tokens without using defect information. DTAN allocated computation according to local anomaly contrast, patch uncertainty, and defect information from earlier layers. Suspected defect patches remained active through the full attention process. Repeated background patches received less computation, while uncertain patches were checked by a lightweight verification branch. Ablation tests were conducted by removing the anomaly-contrast term, the cross-layer defect term, or the verification branch. All models used the same image size, training samples, augmentation settings, and evaluation criteria.

2.3 Measurement Methods and Quality Control

Model performance was measured using defect-level precision, recall, F1-score, false-alarm rate, computational cost, and inference latency. Recall was also calculated separately for defects smaller than 32×32 pixels because these defects were more likely to be lost during token reduction. A predicted region was counted as correct when it matched the corresponding ground-truth defect under the preset overlap criterion. Inference latency was measured after several warm-up runs. Each model was then tested repeatedly under the same hardware settings, and the median runtime was recorded. The number of active tokens was recorded at each transformer layer to measure the reduction in computation. Defect labels were checked twice.

Samples with unclear defect boundaries were reviewed again before analysis. Training and test samples from the same production batch were kept separate.

2.4 Data Processing and Model Formulation

Each image was divided into non-overlapping patches and converted into token features before entering the transformer. For patch i at layer l , the token priority score was calculated from local anomaly contrast, prediction uncertainty, and defect information retained from earlier layers:

$$P_i^{(l)} = \alpha C_i^{(l)} + \beta U_i^{(l)} + \gamma H_i^{(l)}$$

where $C_i^{(l)}$ is the local anomaly contrast, $U_i^{(l)}$ is the prediction uncertainty, and $H_i^{(l)}$ is the defect information retained from earlier layers. The parameters α , β , and γ are the weights of the three terms. Tokens with high $P_i^{(l)}$ values were kept for full attention. Tokens with low scores were reduced, while uncertain tokens were sent to the verification branch. Detection performance was measured using the F1-score:

$$F1 = \frac{2PR}{P+R}$$

where P is precision and R is recall. The reduction in computation was calculated as:

$$CR = 1 - \frac{F_{DTAN}}{F_{full}}$$

where F_{DTAN} is the computational cost of DTAN and F_{full} is the computational cost of the full-token transformer.

2.5 Training and Statistical Analysis

All models were trained with the same optimizer, learning-rate schedule, batch size, preprocessing method, and stopping rule. Model selection was based on validation results, and the test set was used only for final evaluation. The main experiments were repeated using independent training runs, and the mean values were reported. F1-score, recall, false-alarm rate, computational cost, and inference latency were compared across methods. Results were also grouped by defect size and dataset to examine performance on small and large defects. Paired statistical tests were used when two models were evaluated on the same images. Statistical significance was defined as $p < 0.05$. Online tests were performed under fixed conveyor speed and image-acquisition settings to measure performance during continuous inspection.

3. Results and Discussion

3.1 Defect Detection and Small-Defect Recall

DTAN achieved a defect-level F1-score of 91.8% across the four test datasets. The main improvement was found in small-defect detection. For defects smaller than 32×32 pixels, recall reached 88.6%, compared with 82.4% for the fixed token-pruning model. This represents an

increase of 6.2 percentage points. As shown in Fig. 1, DTAN kept more patches that contained weak or uncertain defect signals. This helped reduce the loss of small scratches, pits, and cracks during token reduction. Small defects remain difficult for transformer-based inspection because they may occupy only one or two image patches. Islam et al. (2026) also reported lower accuracy for small steel defects when using multi-scale deformable attention. Defect Transformer showed that local spatial information is important when transformer features are used for surface inspection [25,26]. These results support the use of defect-related information when deciding which tokens should remain active.

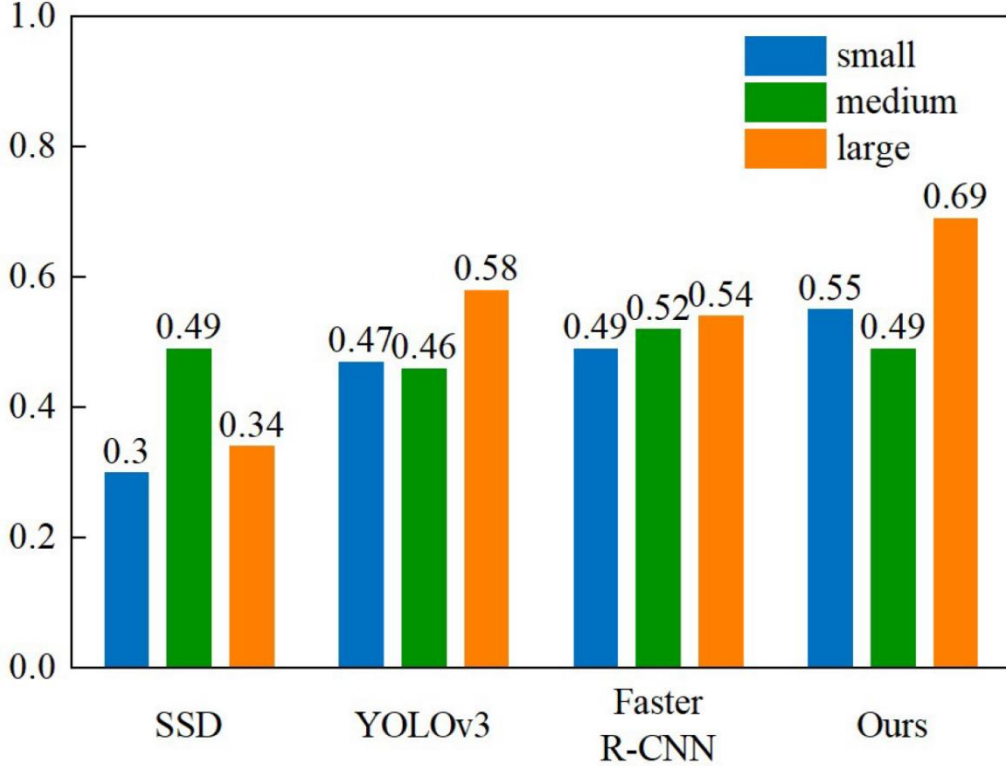


Fig. 1. Defect-level F1-score and small-defect recall of DTAN and the fixed token-pruning model.

3.2 Computational Cost and Inspection Speed

DTAN reduced transformer computation by 33.9%. Average inference latency decreased from 47.3 ms to 29.5 ms per image, corresponding to a reduction of about 37.6%. The processing rate increased from about 21.1 to 33.9 images per second and supported stable inspection at a conveyor speed of 1.6 m/s. The results are shown in Fig. 2. DynamicViT reported a 31–37% reduction in FLOPs through token pruning and also improved image-processing speed [27,28]. However, DynamicViT was mainly tested on image classification. In industrial inspection, a removed token may contain the main part of a small defect. Liang et al. (2026) also used token merging to reduce the cost of a transformer-based metal defect model. DTAN differs from fixed pruning and token merging because token removal is linked to defect probability. This allows computation to be reduced mainly in background regions while more processing is kept for possible defects.

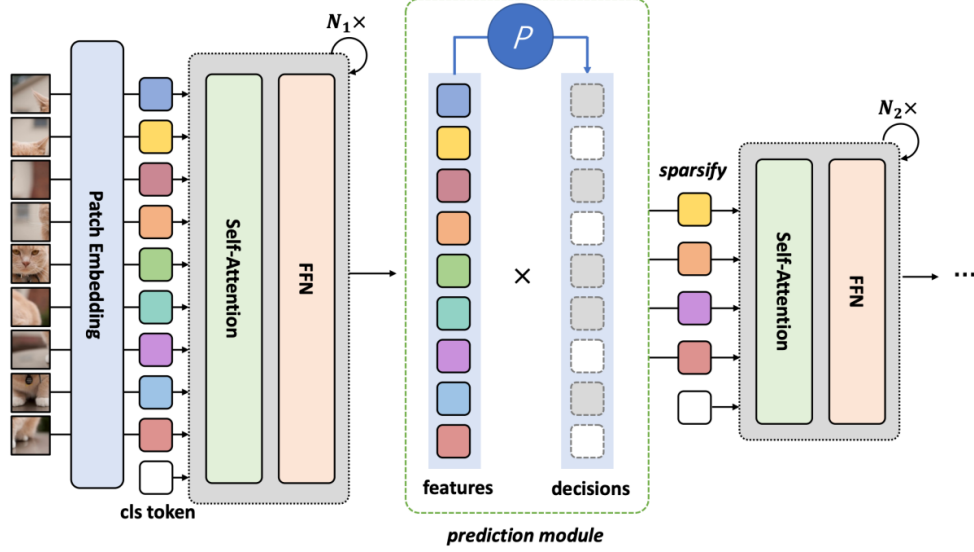


Fig. 2. Computational cost and inference latency of DTAN, the full-token model, and the token-pruning model.

3.3 False Alarms on Repeated Surface Textures

DTAN reduced false alarms caused by repeated textures by 21.7%. The reduction was more evident on fabrics, coated surfaces, and materials with regular local patterns. These patterns may produce strong attention values even when no true defect is present. A pruning method based only on attention scores may therefore keep many background tokens. DTAN also considers local anomaly contrast and defect information from earlier layers. Background patterns that remain similar to nearby regions can receive less computation in later layers. The verification branch is used for uncertain patches before a final token decision is made. Gnieciak et al. (2025) also showed that combining transformer features with multi-scale information can improve the separation of normal texture and abnormal regions. The present results show that repeated surface patterns should be considered during token selection, especially when false alarms are an important part of industrial inspection performance [29,30].

3.4 Performance Across Datasets and Production Conditions

DTAN was evaluated on NEU-Surface, DAGM, KolektorSDD2, and internal factory data from 14 production lines. The factory dataset contained 286,000 images and 412,600 labeled defects. Defect sizes ranged from 6×8 to 420×360 pixels. This dataset covered wider changes in material type, defect size, and production conditions than many public benchmark datasets. Yuan et al. (2026), for example, used 2778 high-resolution images from five metal defect datasets together with MVTec AD. Bao et al. (2026) evaluated their model on the 1800-image NEU-DET dataset. The larger production dataset used here provides a broader test of DTAN under practical inspection conditions. However, the factory images were still collected from a limited number of production lines. Changes in cameras, lighting, materials, or conveyor settings may affect model

performance. Further testing on new factories and previously unseen defect types is needed to examine performance under other production conditions.

4. Conclusion

This study developed DTAN for fast industrial surface defect inspection. DTAN assigns different amounts of computation to image patches based on anomaly contrast, prediction uncertainty, and defect information from earlier transformer layers. The method achieved a defect-level F1-score of 91.8% and reduced transformer computation by 33.9%. For defects smaller than 32×32 pixels, recall reached 88.6%, which was 6.2 percentage points higher than that of the fixed token-pruning model. Average inference latency decreased from 47.3 ms to 29.5 ms per image, while false alarms caused by repeated textures were reduced by 21.7%. These results show that reducing background tokens can improve inspection speed while keeping important small-defect information. The main contribution of DTAN is its defect-aware token allocation process, which keeps likely defect patches, reduces repeated background regions, and checks uncertain patches before token removal. The method can be used for real-time inspection of steel, coated materials, fabrics, printed circuit boards, and other industrial surfaces. However, the current tests were limited to selected public datasets and 14 production lines. Changes in cameras, lighting, materials, and production settings may affect the results. Future work should test DTAN on more factories, new defect types, and different edge devices.

References

1. Xiong, W., Guo, Y., & Zeng, Y. (2026). A Study on the Energy Efficiency of MEP Systems and Coordinated Dispatch Mechanisms for Multi-energy Systems in High-density Urban Buildings. Available at SSRN 7120518.
2. Olianin, D., Tsupryk, H., Hovorushchenko, T., Bahrii-Zaiats, O., & Andrushchak, I. (2025, June). Transformer neural networks in Industry 4.0. In CITI (pp. 331-336).
3. Hong, Z., Liang, S., Xu, T., & Chen, H. (2026). A Study on Failover Verification and Recovery Objective Prediction for Cross-Region Cloud Services.
4. Zhang, Z., Wang, J., Li, Z., Wang, Y., & Zheng, J. (2025). Annocoder: A mti-agent-based code generation and optimization model. *Symmetry*, 17(7), 1087.
5. Maiga, O., & Akhloufi, M. A. (2026, July). Unsupervised Rice Leaf Disease Detection Using PatchCore Feature Memory Banks. In 2026 IEEE International Conference on Signals and Systems (ICSigSys) (pp. 1-5). IEEE.
6. Huang, J., Yin, J., Xu, T., & Yang, J. (2026). Quality Assessment and User Adoption Prediction for Enterprise-Level Financial Reporting Data Products. Available at SSRN 7179118.

7. Yao, G., Zheng, J., Wang, Z., Zhang, W., Han, R., Zhao, C., ... & Liu, R. (2026, March). V-pruner: A fast and globally-informed token pruning framework for vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 40, pp. 34396-34404).
8. Chen, F., Li, S., Liang, H., Xu, P., & Yue, L. (2025). Optimization Study of Thermal Management of Domestic SiC Power Semiconductor Based on Improved Genetic Algorithm.
9. Barusco, M., Pezze, D. D., Borsatti, F., & Susto, G. A. (2026). AdapTS: Lightweight Teacher-Student Approach for Multi-Class and Continual Visual Anomaly Detection. *arXiv preprint arXiv:2603.17530*.
10. Zhang, Z. (2026). A Study on the Impact of Cost Presentation on Decision-making Bias and Cancellation Behavior in Online Subscriptions. Available at SSRN 7349839.
11. Qi, C., & Qiao, X. (2026). Building and Operating a Large Scale Multi-Agent System: A Case Study from Industry. Available at SSRN 6795198.
12. Du, Y., Liu, Q., Dong, Y., & Chen, Z. (2026). Cross-Domain Transfer and Few-Shot Recognition of Defect Images in Complex Industrial Settings.
13. Ercoşkun, B. (2025). Integration of Image Processing Techniques into Automotive Quality Control Processes: Methods, Applications, and Future Perspectives. *Gazi Journal of Engineering Sciences*, 11(3), 426-448.
14. Zhao, J., Fan, J., & Li, L. (2026). A Study on an Explainable Causal-Enhanced LLM Agent for Predicting the Forming Quality of Automotive Component Materials.
15. You, S. (2026). Verifiable Audit Mechanisms in AI Compliance Automation: Scalability. Available at SSRN 6547458.
16. Bao, Y., & Wang, H. (2026). Orion: A High-Throughput, Fault-Tolerant Event Routing Architecture for Hyperscale Data Streams. *Fault-Tolerant Event Routing Architecture for Hyperscale Data Streams* (January 01, 2026).
17. Cahyani, P. I. G., Suartana, K. D. D., & Yudistira, N. (2025). Input-adaptive visual preprocessing for efficient fast vision-language model inference. *arXiv preprint arXiv:2512.20839*.
18. Wang, S., Feng, Y., & Fang, X. (2026, May). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving. In *2026 6th International Symposium on Computer Technology and Information Science (ISCTIS)* (pp. 253-256). IEEE.
19. Yang, Q., & Du, Y. (2026). Predicting Rollback Risks and Controlling Gradual Rollout Traffic for Online Ranking Feature Deployments.
20. Pradeep, A., Nazari, S., Taheri, M., & Herglotz, C. (2026). Fusion: A Framework for Unified Sequential Token Adaptation in Vision Transformers. *arXiv preprint arXiv:2607.02612*.
21. Zhao, Z., & Welsch, R. E. (2026). Point-in-Time Financial RAG with Frozen LLMs and Market-Feedback Adaptive Retrieval. *arXiv preprint arXiv:2605.31201*.

22. Yuan, Y., Huang, J., Yin, J., & Xu, T. (2026). Confidence Calibration and Semantic Error Analysis for Ambiguous Coreference Resolution in User-generated Social Media Text. Available at SSRN 7393238.
23. Latibari, B. S., Homayoun, H., & Sasan, A. (2025). Optimizing vision transformers: Unveiling ‘focus and forget’ for enhanced computational efficiency. *IEEE Access*, 13, 27908-27927.
24. Jiao, Y., Shi, T., Zhao, B., & Wang, A. (2026). Retrieval-Guided Structured Reasoning and Interpretable Representation Learning for Large-Scale Video–Language Models.
25. Liang, S., Du, Y., Chen, W., & Liu, Z. (2026). Order Allocation and Emergency Transportation Decisions Amid Multi-Source Procurement Disruptions.
26. Islam, M. S., Bappy, M. M., Tushar, S. R., Zamiela, C., & Wasi, A. T. (2026). Beyond local receptive fields: vision transformers for real-time surface defect detection in FDM. *The International Journal of Advanced Manufacturing Technology*, 143(9), 5399-5420.
27. Liang, S., Hsu, K. H., & Liu, Z. (2026). Rolling Scheduling and Capacity Balancing for Server Assembly Under Engineering Change Disturbances.
28. Bao, Y., & Wang, H. (2026). Large-scale Metrics Migration: A Systematic Approach to Rebuilding Production Measurement Infrastructure. Available at SSRN 7185660.
29. Gnieciak, D., & Szandala, T. (2025). Large language models versus static code analysis tools: A systematic benchmark for vulnerability detection. *IEEE Access*, 13, 198410-198422.
30. Yuan, Y., Yin, J., Huang, J., & Xu, T. (2026). Weakly Supervised Anomaly Detection and Privacy Risk Scoring in Community Message Threads.