



Scene-Adaptive Token Budgeting for Urban Traffic Perception

Lukas Meier^{1*}, Anna Keller², Marco Frei³

¹Department of Computer Science, ETH Zurich, Universitätstrasse 6, 8092 Zurich, Switzerland

Corresponding Author*: Lukas Meier E-mail: lukas.meier@ethz.ch

ARTICLE INFO	ABSTRACT
<i>Keywords</i> Urban traffic perception efficient transformer token budgeting autonomous driving object detection semantic segmentation real-time inference <i>Published:</i> 03 December 2026	Urban traffic perception systems must process complex visual scenes under changing weather, illumination, density, and object scale. Transformer-based perception models provide strong contextual modeling, but their computational cost limits real-time deployment on vehicle-grade hardware. This study explores scene-adaptive token budgeting for efficient urban traffic perception. We develop the Scene Token Budgeting Transformer (STBT), which estimates the required token budget from scene complexity indicators, including object density, lane-boundary uncertainty, foreground motion, and low-light degradation. Simple road scenes are processed with a reduced token budget, while crowded intersections, nighttime frames, and scenes with small vulnerable road users retain more visual tokens. The evaluation was organized around driving conditions rather than only dataset averages. Experiments used Cityscapes, BDD100K, and nuScenes camera data, with 196,000 annotated frames, 1.84 million object boxes, and 25 semantic road categories. In clear daytime road scenes, STBT reduced token computation by 46.2% with only a 0.3-point decrease in detection mAP. In rainy-night scenes, the model automatically retained 71.5% of foreground-related tokens and limited mAP loss to 1.1 points. On an NVIDIA Orin automotive platform, end-to-end perception latency decreased from 62.8 ms to 41.6 ms per frame. Pedestrian and cyclist recall at distances above 35 meters improved by 3.4 points compared with a fixed-ratio pruning strategy, mainly because the token budget was preserved for small foreground objects. The results indicate that scene-level token budgeting can support real-time traffic perception while maintaining safety-sensitive detection performance.

1. Introduction

Urban traffic perception is a fundamental component of autonomous driving and intelligent transportation systems. It supports a wide range of tasks, including object detection, lane recognition, semantic segmentation, drivable-area estimation, and scene understanding. Reliable perception is particularly important because road environments change continuously with traffic density, illumination, weather, vehicle speed, object scale, and occlusion. Pedestrians, cyclists, traffic signs, and other small road users may occupy only a limited number of pixels when they appear at long distances, while shadows, headlights, rain, and partial occlusion can further reduce their visual saliency. These challenges require perception models to preserve both fine local details and broader contextual relationships. Vision Transformers have become increasingly important for dense visual prediction because self-attention can model long-range dependencies across image regions and capture contextual information beyond the receptive field of conventional convolutional operations (Palanisamy et al., 2025; Yang, 2026). At the same time, recent globally informed token-pruning methods have shown that redundant visual tokens can be identified using

Citations: Meier, L., Keller, A., & Frei, M. (2026). Scene-Adaptive Token Budgeting for Urban Traffic Perception. *The Journal of Applied Engineering and Technologies*, 1(3), 147-155.

information distributed across the Transformer representation, providing a promising way to reduce computation without relying only on local attention responses (Yao et al., 2026). Such developments suggest that Transformer computation can be selectively allocated rather than applied uniformly to every image region. Transformer-based architectures have consequently been adopted in many autonomous-driving perception systems. Multi-camera and bird's-eye-view models combine information from different viewpoints to improve spatial reasoning, while temporal modeling further supports object localization and scene understanding across consecutive frames (Koh et al., 2025; Zhang, 2026). Query-based 3D perception methods also reduce the dependence on dense voxel representations by directly reasoning about objects and their spatial relationships (Du et al., 2026; Qi & Qiao, 2026). More recent sparse and geometry-aware architectures have improved the efficiency of multi-view feature aggregation and three-dimensional perception (Bozic et al., 2026; Xiong et al., 2026). Despite these advances, computational cost remains a major limitation. High-resolution traffic images produce large numbers of visual tokens, and these tokens are repeatedly processed through attention and feed-forward layers. The resulting memory consumption and latency can become substantial when several cameras are processed simultaneously (Hong et al., 2026; Huang et al., 2026). This problem is particularly relevant for autonomous vehicles, where perception systems must share limited computation, memory bandwidth, and power resources with localization, planning, prediction, and control modules (Khatab et al., 2026; Yang & Du, 2026). A model that achieves high benchmark accuracy but requires excessive computation may therefore be difficult to deploy under real-time vehicle constraints.

A growing body of research has attempted to reduce Transformer complexity through lightweight architectures, sparse computation, and adaptive token processing. Efficient network designs reduce latency by simplifying attention operations and improving compatibility with practical hardware (Yuan et al., 2026a; Zhao & Welsch, 2024). Dynamic spatial sparsification selectively removes or bypasses image regions that contribute little to the final prediction (Pradeep et al., 2026), while token-pruning strategies estimate token importance and discard low-value representations during intermediate stages (Zhang et al., 2025). Other approaches preserve compressed information from discarded tokens through token squeezing or feature aggregation, reducing information loss after pruning (Bao et al., 2026; Liang et al., 2026b). Dynamic token halting further allows different tokens to stop computation at different network depths according to their estimated importance (Khan et al., 2026; Zhao et al., 2026). Adaptive pruning has also been introduced into object-detection pipelines to reduce redundant processing in background areas and other low-information regions (You, 2026). Together, these studies indicate that a considerable proportion of visual computation can be avoided without processing every token through the complete Transformer backbone. However, existing adaptive-computation methods are not fully aligned with the characteristics of urban traffic perception. Token importance is commonly estimated from attention weights, classification confidence, feature magnitude, or predefined pruning ratios. These indicators describe the current network representation, but they do not necessarily represent the perceptual difficulty of the entire driving scene. A clear daytime road containing several large vehicles may require substantially less visual computation than a crowded intersection containing pedestrians, bicycles, lane markings, reflections, and partially occluded vehicles. Similarly, nighttime illumination, rain, glare, and distant small objects may increase the amount of ambiguous visual information even when the total number of objects is not high. Applying the same pruning ratio to these different conditions may therefore create an inappropriate balance between computational efficiency and perception reliability. This limitation is especially

important for safety-relevant small objects. A token covering a distant pedestrian or cyclist may contribute only weakly to global attention because the object occupies a very small image region. Removing such a token can have little influence on average feature statistics while substantially reducing recall for vulnerable road users. Similar problems can occur around uncertain lane boundaries, partially visible vehicles, traffic signs, and low-contrast foreground objects. Aggressive pruning can therefore improve average inference speed while introducing condition-specific performance degradation that is not immediately visible from overall dataset metrics. Evaluations based only on mean detection accuracy or mean segmentation scores may further conceal this problem because common daytime scenes can dominate the test distribution. Several studies have shown that autonomous-driving perception performance varies considerably across environmental conditions, object scales, visibility levels, and scene complexity (Jiao et al., 2026; Pichappan, 2026). These observations indicate that computational allocation should be evaluated together with scenario difficulty rather than treated only as a global efficiency problem. Another limitation is that most token-pruning mechanisms determine whether individual tokens should be retained but provide relatively little control over the total computational budget assigned to an entire traffic scene. Token-level importance and scene-level difficulty represent related but different questions. Token selection determines which visual regions deserve further computation, whereas scene-level budgeting determines how much computation should be available before those decisions are made. Introducing a scene-dependent budget can therefore provide an additional control mechanism. Frames containing sparse traffic, clear illumination, and stable road boundaries can tolerate stronger compression, while frames containing dense road users, uncertain lane structures, foreground motion, or poor illumination may require a larger token budget. Such an approach is particularly relevant to edge deployment, where computational resources are limited and unnecessary processing in simple scenes directly reduces achievable frame rate and energy efficiency.

This study proposes the Scene Token Budgeting Transformer (STBT), a scene-adaptive framework that dynamically adjusts the number of visual tokens retained during traffic perception. Instead of applying a fixed pruning ratio across all inputs, STBT estimates scene complexity from complementary indicators related to object density, lane-boundary uncertainty, foreground motion, and low-light degradation. These signals are used to determine an appropriate token budget for each traffic frame. Relatively simple scenes are processed with fewer tokens to reduce redundant computation, whereas crowded intersections, nighttime scenes, motion-intensive environments, and frames containing difficult road users retain a larger portion of the visual representation. The objective is not merely to maximize pruning, but to allocate computation according to the amount and difficulty of perceptual information present in the scene. The proposed framework is evaluated on Cityscapes, BDD100K, and nuScenes camera data to examine whether scene-adaptive token allocation can provide a more reliable efficiency–accuracy trade-off than fixed-budget pruning. The evaluation considers object detection and semantic perception together with foreground-token retention, inference latency, and recall for distant pedestrians and cyclists. Performance is further analyzed across traffic density, illumination, object scale, and other challenging driving conditions rather than being represented only by dataset-level averages. Deployment experiments on an NVIDIA Orin platform are used to determine whether the reduction in visual computation produces measurable real-time benefits on vehicle-oriented hardware. Through this design, the study aims to establish a closer connection between scene difficulty and Transformer computation. Its broader significance is to provide a perception architecture that saves computation when visual redundancy is high while preserving additional representational capacity when traffic conditions become uncertain or safety-critical, thereby improving the practical feasibility of adaptive Transformer

perception for resource-constrained autonomous-driving systems.

2. Materials and Methods

2.1. Datasets and Study Samples

The study used camera data from Cityscapes, BDD100K, and nuScenes. After data cleaning and label matching, 196,000 annotated frames were included. The data contained about 1.84 million object boxes and 25 semantic road classes. The selected frames covered clear daytime roads, crowded intersections, rainy conditions, nighttime traffic, and mixed-light scenes. The main road objects included vehicles, pedestrians, cyclists, traffic signs, lane regions, and road boundaries. Frames with missing labels, serious image damage, or incomplete camera records were removed. Frames from the same driving sequence were placed in only one data subset to avoid data overlap. The samples were divided into training, validation, and test sets at the sequence level. Each frame was also grouped by weather, illumination, traffic density, and foreground-object size for later analysis.

2.2. Experimental Design and Control Settings

STBT was compared with a full-token Transformer and a fixed-ratio token-pruning model. The full-token Transformer kept all visual tokens and served as the accuracy control. The fixed-ratio model removed the same proportion of tokens from every frame. It was used to compare fixed pruning with scene-based token control. Ablation tests were also performed by removing object density, lane-boundary uncertainty, foreground motion, and low-light information separately. All models used the same backbone, input size, optimizer, training schedule, image augmentation, and data split. STBT changed the token budget for each frame according to scene difficulty. Clear daytime scenes used fewer tokens. Crowded intersections, low-light scenes, and frames with strong foreground motion kept more tokens. This setting allowed the effect of scene-based token control to be tested under the same model and training conditions.

2.3. Measurement Methods and Quality Control

All frames were resized to the same input size and normalized before training. Random horizontal flipping, scaling, small shifts, brightness changes, and contrast changes were applied only to the training data. Object and semantic labels from the three datasets were mapped to the same class system before model training. The mapped labels were checked to reduce errors caused by different class names and annotation rules. Detection performance was measured by mAP, and semantic segmentation was measured by mIoU. Pedestrian and cyclist recall was calculated separately for objects more than 35 m from the camera. Model cost was measured by retained-token number, GPU memory use, and inference latency. All latency tests were performed on the same NVIDIA Orin platform with the same batch size, input size, and software settings. Warm-up runs were not included in the final timing results. The mean value from repeated runs was reported. Test frames were also grouped by daytime, nighttime, rain, traffic density, and small-object content to examine model performance under different road conditions.

2.4. Data Processing and Model Formulation

For each frame, STBT calculated a scene-complexity score C_s from object density, lane-boundary uncertainty, foreground motion, and low-light level. Each input term was normalized to

the same range before calculation. The scene-complexity score was defined as:

$$C_s = \alpha D_s + \beta U_s + \gamma M_s + \delta L_s$$

where D_s is object density, U_s is lane-boundary uncertainty, M_s is foreground-motion intensity, and L_s is the low-light score. The terms α , β , γ , and δ are weight parameters. The token budget for frame s was then calculated as:

$$N_s = N_{\min} + (N_{\max} - N_{\min}) \cdot \text{clip}(C_s, 0, 1)$$

where N_s is the number of retained tokens, and $N_{\min}$ and $N_{\max}$ are the minimum and maximum token budgets. A lower C_s gives a smaller N_s , while a higher C_s keeps more visual tokens. Foreground-related tokens were given higher priority than clear background tokens. The values of α , β , γ , δ , $N_{\min}$, and $N_{\max}$ were selected on the validation set and were not changed during final testing.

2.5. Model Evaluation and Statistical Analysis

Model performance was evaluated at both the full-dataset level and the scene level. The main measures were detection mAP, segmentation mIoU, pedestrian recall, cyclist recall, retained-token ratio, inference latency, and GPU memory use. Separate results were obtained for clear daytime roads, crowded intersections, rainy-night scenes, and frames containing small road users. The difference in mAP between the full-token model and each reduced-token model was used to measure the accuracy loss caused by token reduction. Long-distance pedestrian and cyclist recall was used to examine whether token reduction removed useful information from small foreground objects. Results from repeated runs were reported as mean $\pm$ standard deviation. Paired comparisons were made using the same test frames. The paired t-test was used when the differences followed an approximately normal distribution. Otherwise, the Wilcoxon signed-rank test was used. Statistical significance was set at $p < 0.05$. Accuracy and inference latency were considered together when comparing model efficiency.

3. Results and Discussion

3.1. Overall Perception Performance

STBT achieved detection accuracy close to that of the full-token Transformer while using fewer tokens in simple scenes. In clear daytime traffic, token computation decreased by 46.2%, while detection mAP dropped by only 0.3 points. This small loss indicates that many tokens from clear background regions contribute little to the final prediction. Fixed-ratio pruning showed a larger weakness because it used the same token ratio for every frame. As shown in Fig.1, STBT assigned fewer tokens to simple road scenes and increased the token budget when traffic density, motion, or scene uncertainty increased. Previous studies have also found that token importance changes across images and network stages (Esteves et al., 2025; Liang et al., 2026a). STBT differs by using traffic-scene conditions to set the token budget directly.

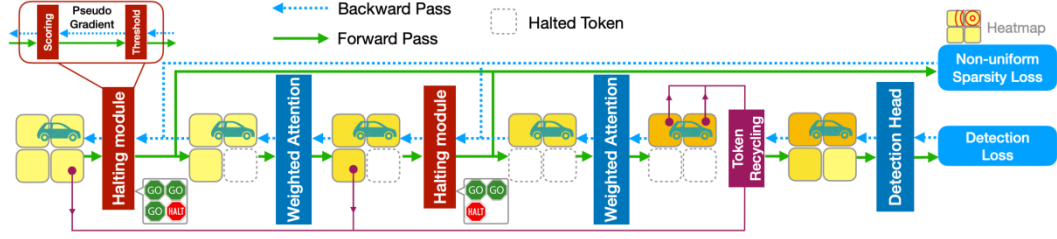


Fig. 1. Token use and detection performance of STBT under different traffic conditions.

3.2. Token Reduction and Inference Speed

The lower token count also reduced inference time on the NVIDIA Orin platform. Mean end-to-end latency decreased from 62.8 ms to 41.6 ms per frame, a reduction of 33.8%. This result is important because lower computational cost does not always produce the same change in hardware latency (Bao & Wang, 2026; Wang et al., 2026). STBT mainly reduces token processing in areas such as road surfaces, sky, buildings, and other clear background regions. More tokens are kept in crowded or unclear areas. Palanisamy et al. (2025) reported that less useful tokens can stop processing at earlier stages with only a small loss in detection accuracy. Yao et al. (2026) also found that image-based pruning performs better than a fixed pruning ratio in dense vision tasks. These findings support the use of different token budgets for traffic scenes with different levels of difficulty.

3.3. Rainy-Night and Low-Light Performance

The advantage of STBT was clearer under poor imaging conditions. In rainy-night scenes, the model retained 71.5% of foreground-related tokens and limited the mAP loss to 1.1 points compared with the full-token model. Low light reduces object contrast, while rain causes reflections and unclear boundaries. These effects increase the risk of removing useful tokens too early. STBT increases the token budget when low-light level, motion, or lane uncertainty becomes higher. As shown in Fig.2, more tokens were kept around vehicles, pedestrians, cyclists, and unclear road boundaries. Koh et al. (2025) also reported that tokens removed at early stages may still contain useful information for later detection. The present results show that one fixed pruning ratio is not suitable for both clear daytime roads and difficult nighttime scenes.

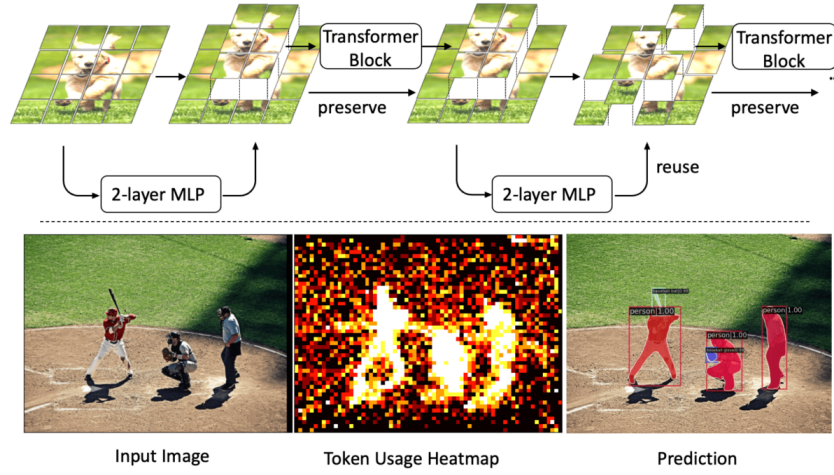


Fig. 2. Token retention around foreground objects and unclear regions in complex traffic scenes.

3.4. Long-Distance Pedestrian and Cyclist Detection

STBT showed a clear gain for small road users. For pedestrians and cyclists more than 35 m from the camera, recall increased by 3.4 points compared with fixed-ratio pruning. Small objects occupy only a few visual tokens. Removing several of these tokens can greatly reduce the information needed for detection. Fixed pruning does not separate these small foreground regions from large background areas. STBT gives higher priority to regions with foreground motion, high object density, or uncertain features. More tokens are therefore kept around distant pedestrians and cyclists. Previous studies have shown that dynamic token selection can reduce computation with only a small drop in average accuracy, but most work reports overall detection or segmentation results rather than distance-based recall for road users (Ishibashi & Meng, 2025; Yuan et al., 2026b). The present results show that scene-based token control can reduce processing cost while keeping more information for small road users. Further tests on more cameras, severe weather, and other vehicle hardware are still needed.

4. Conclusion

This study developed STBT for urban traffic perception under changing road conditions. The model sets the token budget according to object density, lane uncertainty, foreground motion, and low-light level. In clear daytime scenes, STBT reduced token computation by 46.2%, while detection mAP decreased by only 0.3 points. On the NVIDIA Orin platform, mean inference time decreased from 62.8 ms to 41.6 ms per frame. In rainy-night scenes, STBT retained 71.5% of foreground-related tokens and limited the mAP drop to 1.1 points. For pedestrians and cyclists more than 35 m from the camera, recall was 3.4 points higher than that of fixed-ratio pruning. These results show that different traffic scenes require different amounts of token processing. Scene-based token control can reduce computation while keeping useful information for small and hard-to-detect road users. STBT may be used in real-time traffic perception systems with limited computing resources. However, the experiments included only three datasets and one vehicle computing platform. Future studies should test the model with more cameras, cities, weather conditions, and embedded devices.

References

1. 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).
2. Bao, Y., Qiu, Y., & Wang, H. (2026). FLARE: A Real-Time Framework for Detecting Malicious Account Activities at Internet Scale. Available at SSRN 7185659.
3. Bozic, V., Slavkovic, I., Narnhofer, D., Metzger, N., Rozumny, D., Schindler, K., & Kalischek, N. (2026). Unified Panoramic Geometry Estimation via Multi-View Foundation Models. arXiv preprint arXiv:2605.26368.
4. Du, Y., Liu, Q., Dong, Y., & Chen, Z. (2026). Cross-Domain Transfer and Few-Shot Recognition of Defect Images in Complex Industrial Settings.
5. Esteves, C., Suhail, M., & Makadia, A. (2025, October). Spectral image tokenizer. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 17181-17190). IEEE.
6. Hong, Z., Liang, S., Xu, T., & Chen, H. (2026). A Study on Failover Verification and Recovery Objective Prediction for Cross-Region Cloud Services.
7. 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.
8. Ishibashi, R., & Meng, L. (2025). Automatic pruning rate adjustment for dynamic token reduction in vision transformer: R. Ishibashi and L. Meng. *Applied Intelligence*, 55(5), 342.
9. Jiao, Y., Shi, T., Zhao, B., & Wang, A. (2026). Retrieval-Guided Structured Reasoning and Interpretable Representation Learning for Large-Scale Video–Language Models.
10. Khan, A., Junaid, A., Iqbal, A., Husnain, G., Shin, W., & Lim, S. (2026). A Survey of Dynamic Token Computation in Transformers: Taxonomy, Stability and Budget-Aware Evaluation. *IEEE Access*.
11. Khatab, E., Fathy, F., AlKholy, A., & Shalash, O. (2026). Sensor Fusion and Perception for Autonomous Driving: A Critical Review of Modalities, AI Models, Algorithms, and Industry Configurations. *Machine Learning and Knowledge Extraction*, 8(7), 199.
12. Koh, J., Lee, Y., Kim, J., Lee, D., & Choi, J. W. (2025). OnlineBEV: Recurrent Temporal Fusion in Bird’s Eye View Representations for Multi-Camera 3D Perception. *IEEE Transactions on Intelligent Transportation Systems*.
13. Liang, S., Hsu, K. H., & Liu, Z. (2026a). Rolling Scheduling and Capacity Balancing for Server Assembly Under Engineering Change Disturbances.
14. Liang, S., Zhang, X., Du, Y., & Chen, W. (2026b). Supply Network Restructuring and Capacity Ramp-Up Limits in the Localized Expansion of High-Performance Computing Equipment Production. Available at SSRN 7339738.
15. Palanisamy, B., Hassija, V., Chatterjee, A., Mandal, A., Chakraborty, D., Pandey, A., ... & Kumar, D. (2025). Transformers for vision: A survey on innovative methods for computer vision. *Ieee Access*, 13, 95496-95523.

16. Pichappan, P. (2026). Context-Aware Scene Understanding for Autonomous Driving: Quantifying Semantic Dependencies, Complexity, and Perception Risk. *Journal of Electronic Systems* Volume, 16(1).
17. 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*.
18. Qi, C., & Qiao, X. (2026). Building and Operating a Large Scale Multi-Agent System: A Case Study from Industry. Available at SSRN 6795198.
19. 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.
20. 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.
21. Yang, J. (2026). Stage-Coupled Computational Framework for Stratified Accessibility and Equity Analysis in Community-Based Elderly Care Services.
22. Yang, Q., & Du, Y. (2026). Predicting Rollback Risks and Controlling Gradual Rollout Traffic for Online Ranking Feature Deployments.
23. 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).
24. You, S. (2026). Verifiable Audit Mechanisms in AI Compliance Automation: Scalability. Available at SSRN 6547458.
25. Yuan, Y., Huang, J., Yin, J., & Xu, T. (2026a). Confidence Calibration and Semantic Error Analysis for Ambiguous Coreference Resolution in User-generated Social Media Text. Available at SSRN 7393238.
26. Yuan, Y., Yin, J., Huang, J., & Xu, T. (2026b). Weakly Supervised Anomaly Detection and Privacy Risk Scoring in Community Message Threads.
27. 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.
28. Zhang, Z., Wang, J., Li, Z., Wang, Y., & Zheng, J. (2025). Anncoder: A multi-agent-based code generation and optimization model. *Symmetry*, 17(7), 1087.
29. 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.
30. Zhao, Z., & Welsch, R. E. (2024). Hierarchical reinforced trader (hrt): A bi-level approach for optimizing stock selection and execution. *arXiv preprint arXiv:2410.14927*.