

Toward networked cognitive traffic monitoring: Language-driven multi-camera perspective

Senyun Kuang^{1,†}, Minhui Zhang^{2,†}, Xiaoyu Wen³, Haitao Wang¹, Yang Liu¹, Yintao Wei^{1,✉}

 Cite this article: Kuang S, Zhang M, Wen X, et al. *J Intel Connect Veh* 2026, 9(3): 9210072. <https://doi.org/10.26599/JICV.2026.9210092>

1 Introduction

Every large city now runs on video. Major metropolitan areas in China operate camera networks with hundreds of thousands of streams. These cameras produce far more footage than any control room can watch, so modern intelligent transportation systems depend on automated perception (Gao et al., 2025). However, the way we use this footage has barely changed in a decade. Most deployed systems still rely on a fixed catalog of pretrained detectors. Every new event type needs its own data, labels, and retraining, and any question outside the predefined list ends with an operator scrubbing through hours of recordings.

This design is passive by nature. The system reacts to triggers it was built to recognize, and it cannot answer the open questions that operators actually ask. A vehicle that leaves one camera and enters another is treated as a new object. A description that any person would use without thinking, such as "the white van with red lettering on its rear", has no place in a pipeline built around numeric features. We see this as the real ceiling for smart city monitoring, and we believe that recent progress in vision foundation models and multimodal large language models offers a way past it.

2 Three structural barriers

Three limitations keep current systems from improving. The first is identity that breaks across cameras. Reidentification methods learn appearance features from large labeled datasets, but viewpoint and lighting change quickly across a city, and the learned features drift with them. The second is a rigid query interface. Operators reason at very different levels, from how many wrong-way vehicles passed a corridor in the last five minutes to a request to segment the car that just ran the red light to a question about its make and plate region. One detector trained for one event type cannot cover that range. The third is cost. Running large models on every stream of a city network is not feasible in bandwidth, energy, or budget. Pushing all inference to the edge instead caps model capacity and rules out reasoning that needs context from several cameras.

3 Language-driven, networked paradigm

We argue for a shift from single-camera passive detection toward

a networked cognitive paradigm, one in which operators query the city in natural language and perception stays consistent across cameras. Fig. 1 contrasts the two views. Consider how a routine investigation would unfold under this paradigm. An operator looking for a vehicle involved in a minor collision could type a plain request. The system searches the relevant cameras and time window, returns a short list of candidates, and allows the operator to confirm the correct one. What takes hours of manual review today would take minutes, and every step would stay in language that a nonspecialist can read and check. Three ideas turn the shift into an engineering plan.

The first idea is to use language as a cross-camera anchor. A multimodal model can describe a target through attributes that stay fairly stable, such as vehicle type, dominant color, visible markings, and body shape (Liu et al., 2023). Cross-camera perception then becomes a question of whether a language-specified target appears in each view rather than a brittle match between numeric features. Open-vocabulary recognition and promptable segmentation make this practical because a description from one camera can be tested against any tracked object downstream (Kirillov et al., 2023). Since the anchor is written in plain language, an operator can read it, correct it, and keep it as an auditable record of how a decision was reached. Language anchors will fail in predictable ways, and those failures should be treated as normal operating conditions. Occlusion can hide the attributes that distinguish a target, low light washes out color and markings, dense traffic fills a view with near-identical candidates, and the same vehicle can look different across viewpoints. The right response is not to force a single match. The system should return a ranked short list with confidence scores, narrow it with camera topology and plausible travel times, and ask the operator for one more distinguishing detail when ambiguity remains. When the evidence is too thin, the honest answer is to decline rather than guess, leaving the final confirmation to the operator.

The second idea is to build the system around the natural hierarchy of operator intent. Queries fall into three levels. Area-level queries ask for counts, flows, and anomaly rates over a region and a time window. Instance-level queries point to a single object that must be pinned down. Attribute-level queries then interrogate that object for its make, color, or plate region. The

† Senyun Kuang and Minhui Zhang contributed equally to this work.

¹ School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. ² College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge UB8 3PH, UK. ³ Guangzhou Xiaopeng Motors Technology Co., Ltd., Guangzhou 510630, China.

✉ Corresponding author. E-mail: weiyt@tsinghua.edu.cn

Received: June 7, 2026; Revised: July 20, 2026; Accepted: July 29, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

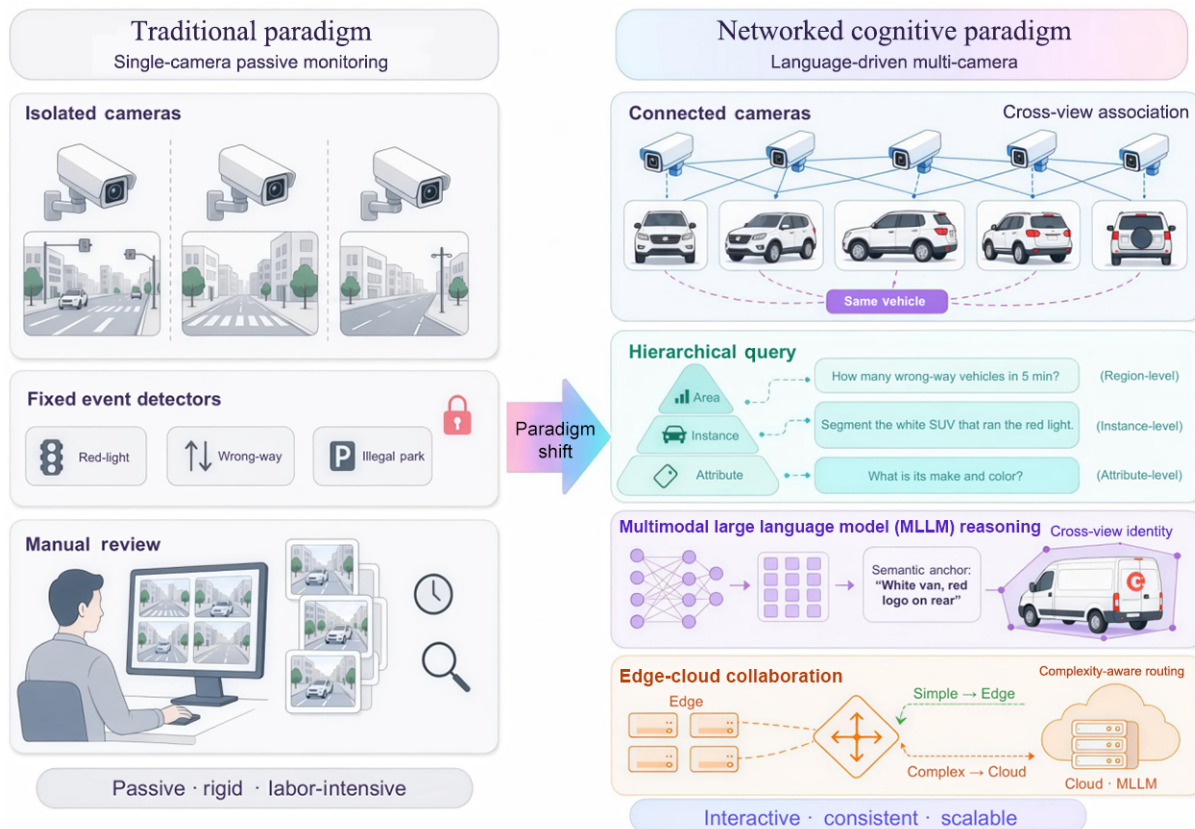


Fig. 1 From single-camera passive monitoring to a networked, language-driven cognitive paradigm. Conventional systems rely on isolated cameras, fixed event detectors, and manual review. The proposed paradigm couples cross-view identity, hierarchical natural-language queries, multimodal reasoning, and edge and cloud routing to support interactive, consistent, and scalable traffic monitoring

three levels should work as one flow, so an operator can find events, select a target, and ask about it, all in natural language.

The hierarchy is not a one-way funnel. A plate region read on one vehicle may widen the instance search to every camera in the district, and an attribute answer about a damaged bumper may spawn an area-level count over the corridor. The interface should therefore carry the anchored target, the active time window, and earlier answers as session states so that each follow-up refines the investigation rather than restarting it.

The third idea is to schedule computation across the edge and cloud by the complexity of each query. Light models at the camera or a local node handle continuous work such as scene change detection, candidate filtering, and simple single-view questions. The cloud is reserved for harder cases that span several cameras or longer time windows, and only a short clip is sent up when deeper analysis is needed. None of this requires waiting for a single breakthrough model. The building blocks already exist across separate communities, and the real work is integration. The harder questions are practical. How should a query be decomposed and routed when the latency budget is tight? When a description fits several objects at once, how many candidates are worth an operator's attention? How should confidence be calibrated and displayed so that an operator knows when to trust a result? Treating these as first-class design problems, rather than afterthoughts, is what will move language-driven monitoring from demonstrations to dependable infrastructure.

4 Agenda for the community

Turning this vision into practice is a research and engineering program, not a single model (Chen et al., 2025), and we close with four priorities. First, the field needs open, city-scale, multicamera datasets that pair natural-language queries with cross-view identity

labels. A traffic video question-answering benchmark is a useful step in this direction (Zhou et al., 2025), but resources at the scale of real deployments are still missing. Second, governance must be designed from the start. Query authorization, data minimization, and audit logging belong in the architecture rather than being added later because a system that can find any described person or vehicle across a city carries clear risks. These principles map directly onto the edge and cloud design. Continuous raw footage never leaves the edge node that captured it and is kept only for a short rolling window set by local regulation. The cloud receives only language anchors, detection metadata, and brief clips of query triggers, and that material is purged when its case closes. Authorization can follow the three query levels, from broadly available area-level statistics to case-authorized instance tracking and role-restricted attribute requests, and every query enters an append-only audit log recording that asks what and when. Third, this is an interdisciplinary effort. Progress will come from joint work across computer vision, natural language processing, networked systems, and urban governance.

Fourth, evaluation deserves attention. The field still lacks agreed ways to measure how well a system understands a query, how consistently it holds an identity across views, and how gracefully it fails. Accuracy on a fixed set of events says little about open-ended language use, so new metrics are needed that reward correct refusals and calibrated uncertainty as much as correct detections. A correct refusal has a concrete meaning here. When two similar white sport-utility vehicles (SUVs) pass the same junction in poor light, the right output is a statement that no confident match exists, not a forced choice. The consequences are not symmetric. A confident wrong answer can send officers to the wrong vehicle, and every later step inherits the error, while a refusal simply returns the case to a human reviewer. Benchmarks should

therefore score an abstention as correct when the ground truth is genuinely ambiguous and penalize a confident error more heavily than a refusal. Shared evaluation of this kind would allow groups to compare their methods honestly.

The shift we describe is not a better detector. It is a change in what a monitoring system is for. Today's systems watch the city and wait for the events they already know. A language-driven and networked system lets operators ask and let the city answer. That is the capability worth building toward.

Author contributions

Senyun Kuang: Conceptualization, Methodology, Investigation, Writing – original draft. **Minhui Zhang:** Methodology, Investigation, Writing – original draft. **Xiaoyu Wen:** Resources, Investigation, Writing – review & editing. **Haitao Wang:** Investigation, Writing – review & editing. **Yang Liu:** Data curation, Validation, Writing – review & editing. **Yintao Wei:** Supervision, Funding acquisition, Project administration.

Acknowledgements

This work was supported by the "U35 Pei Miao Funding Program" of Changping District, Beijing (No. CHPU35202504003), the National Natural Science Foundation of China (Nos. 52572334, 52221005, and 5220105001), the Independent Research Project of the State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University (No. ZZ-GG-20250403), and Tsinghua University (State Key Laboratory of Intelligent Green Vehicle and Mobility) – Hangzhou Airport Economic Demonstration Zone Joint Research Center for Integrated Transportation.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

- Chen, L., Dong, T., Li, X., Xu, X., 2025. Logistics engineering management in the platform supply chain: An overview from logistics service strategy selection perspective. *Engineering*, 47, 236–249.
- Gao, Z., Jia, B., Xie, D., Wang, W., Wu, J., 2025. A discussion on the complexity and transit mechanisms of urban traffic systems. *Engineering*, 44, 24–29.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al., 2023. Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 3992–4003.
- Liu, H., Li, C., Wu, Q., Lee, Y. J., 2023. Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, 34892–34916.
- Zhou, X., Larintzakis, K., Guo, H., Zimmer, W., Liu, M., Cao, H., et al., 2025. TUMTraffic-VideoQA: A Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes. <https://doi.org/10.48550/arXiv.2502.02449>



Senyun Kuang received the B.S. degree in automation from Southwest Jiaotong University, China, in 2019. She is currently a Ph.D. candidate in mechanical engineering at the School of Vehicle and Mobility, Tsinghua University, China. Her research interests include multimodal large language models, intelligent transportation systems, and challenging traffic scene understanding.



Minhui Zhang is an undergraduate student majoring in mathematics for Data Science at the College of Engineering, Design and Physical Sciences, Brunel University London, UK. Her research interests include computer vision and intelligent transportation systems.



Xiaoyu Wen received the B.S. degree in armored vehicle engineering from Nanjing University of Science and Technology, China, in 2018, and the M.E. degree in engineering management from Tsinghua University, China, in 2025. She is currently with Guangzhou Xiaopeng Motors Technology Co., Ltd., where she works on AI-powered chassis technologies. Her research interests include vehicle dynamics and suspension control.



Haitao Wang received the B.S. degree in intelligent science and technology from Qingdao University, China, in 2021. He is currently an M.S. student majoring in engineering management at the School of Vehicle and Mobility, Tsinghua University, China. His research interests include intelligent transportation systems and data-driven decision making.



Yang Liu earned the Ph.D. degree in transportation engineering from Southeast University, Nanjing, China, in 2021. He is currently an Associate Research Professor at the School of Vehicle and Mobility, Tsinghua University. His research interests include intelligent transportation systems, artificial intelligence techniques, and data mining, with applications in complex real-world problems such as autonomous vehicles and urban mobility. He is a recipient of the National High-Level Young Talents Program and a Marie Curie Fellow of the European Union. He serves as an Editorial Board Member of *Transportation Research Part E*, Associate Editor of *IEEE Transactions on Intelligent Vehicles*, and Co-Chair of the WTC Technical Committee on Intelligent Driving and Mobility Services. He has achieved multiple awards in premier AI competitions, including four championships/runner-ups in the KDD Cup, known as the "World Cup of Data Mining".



Yintao Wei received the B.S. and M.S. degrees in computational mechanics and the Ph.D. degree in composite materials from the Harbin Institute of Technology, Harbin, China, in 1992, 1994, and 1997, respectively. He is currently a Professor with tenure at the School of Vehicle and Mobility, Tsinghua University, and Chair of the Smart Chassis Division. He has managed more than 50 research projects, published over 100 articles, and been awarded more than 30 patents. His research interests include intelligent tires for road perception, smart materials and suspensions, and their industrial applications. He is currently an Associate Editor of *Tire Science and Technology* and a Reviewer for several international journals.