Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Perspective:

TransRAG for parallel transportation: toward reliable and trustworthy transportation systems via retrieval-augmented generation*

Jing YANG[1], Xingyuan DAI[1], Yisheng LV[1], Levente KOVÁCS[2], Fei-Yue WANG‡[1,3]

[1]*Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

[2]*John von Neumann Faculty of Informatics, Obuda University, Budapest H-1034, Hungary*

[3]*Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China*

E-mail: yangjing2020@ia.ac.cn; xingyuan.dai@ia.ac.cn; yisheng.lv@ia.ac.cn; kovacs@uni-obuda.hu; feiyue.wang@ia.ac.cn

Received Sept. 13, 2024; Revision accepted Dec. 4, 2024; Crosschecked Dec. 16, 2024; Published online Dec. 26, 2024

https://doi.org/10.1631/FITEE.2400800

Parallel transportation serves as a holistic paradigm for achieving intelligent traffic management and control, focusing on addressing the complexity of human and social factors. Recently, the emergence and development of foundational models (FMs) have ushered in a new era for the realization of parallel transportation. However, the inherent issues of "hallucinations," outdated knowledge, and the "black-box" nature of FMs render their generated decisions unreliable and untrustworthy. To address these issues, we propose a TransRAG framework for parallel transportation based on retrieval-augmented generation and chain-of-thought (CoT) prompting. TransRAG is composed of three interacting layers, storage, management, and execution, which work together to deliver personalized and diverse traffic services to users. The external knowledge from the storage layer is incorporated to augment the FM in management layers for computational experiments. The real–virtual interaction between artificial and actual transportation systems is used to continuously optimize the decisions from the management layer. Therefore, TransRAG can incrementally update knowledge and adjust strategies to adapt to the evolving and dynamic traffic environment. Additionally, the integration of blockchain, smart contracts, and caching into TransRAG is expected to address a range of challenges, such as single point of failure, potential privacy breaches, and delays in data access, thereby advancing the transition to "6S" Transportation 5.0.

1 Introduction

A transportation system is a complex cyber–physical–social system (CPSS) (Wang X et al., 2022, 2024; Yang et al., 2023a), seamlessly integrating physical infrastructure, information processing, and social behaviors, all of which are intricately interconnected and interact with each other. However, existing models address each transportation subproblem (e.g., path planning, traffic flow prediction, and decision recommendation) independently from different perspectives, and no universal model offers comprehensive management of the entire transportation system. Additionally, the uncertainty, diversity, and complexity of human and social factors

‡ Corresponding author

* Project supported by the Science and Technology Development Fund of Macao SAR, China (No. 0093/2023/RIA2) and the National Natural Science Foundation of China (No. U1811463)

ORCID: Jing YANG, https://orcid.org/0000-0002-5918-2991; Xingyuan DAI, https://orcid.org/0000-0001-7517-5049; Yisheng LV, https://orcid.org/0000-0002-7565-4979; Levente KOVÁCS, https://orcid.org/0000-0002-3188-0800; Fei-Yue WANG, https://orcid.org/0000-0001-9185-3989

further exacerbate the difficulty of effective system management.

To address these issues, the concept and framework of parallel transportation (Wang FY, 2010) were proposed to enhance the agility, focus, and convergence in system management and control. Its core principle is artificial transportation systems, computational experiments, and parallel execution (ACP) (Wang FY, 2010; Zhu et al., 2020); that is, the extensive real traffic data collected by ubiquitous terminals alongside virtual data generated by artificial models are used to build artificial transportation systems, where computational experiments are conducted to analyze traffic patterns and their underlying causes, ultimately enabling the implementation of optimal management solutions in real-world systems.

Recently, the advancement of FMs, such as ChatGPT (Zhou et al., 2023), GPT-4, and Sora, has elevated the implementation of parallel transportation systems (Jin et al., 2021; Dai et al., 2024; Gan et al., 2024; Zhang et al., 2024) to a new level because of their exceptional understanding, generation, and reasoning abilities. However, FMs' inherent flaws hinder their application in actual transportation systems and could even jeopardize the safety of passengers and drivers as follows: (1) "Hallucination" phenomena and outdated knowledge within FMs contribute to the unreliability of their inferences; (2) The fact that FMs are neural network based "black-box" models makes their generation results inexplicable and untrustworthy.

Retrieval-augmented generation (RAG) (Lewis et al., 2020; Chen et al., 2024) has garnered widespread attention for enhancing generated content by integrating external knowledge into prompts. The design of CoT prompts (Wei et al., 2022; Feng et al., 2024) is also regarded as an effective method for improving the interpretability and understandability of results by elaborating on the intermediate reasoning steps. Therefore, incorporating RAG and CoT can enable FMs to access real-time knowledge and information without the need for re-pretraining, steering them toward generating more accurate and interpretable outcomes, which, in turn, ensures the reliability of traffic decisions. To this end, we propose a unified parallel transportation framework based on RAG and CoT to guarantee the dependability and trustworthiness of FMs' decision-making in different traffic scenarios and thus the safety of traffic system operations.

2 Parallel transportation

Parallel transportation (Wang FY, 2010) aims at intelligent transportation management and control, whose core is ACP, and infrastructure is CPSS, as shown in Fig. 1. It primarily involves creating one or more artificial transportation systems that mirror an actual transportation system. This includes not only simulating vehicle movements but also modeling human behaviors and activities in the entire societal context. Subsequently, a series of computational experiments are performed within these artificial systems to analyze historical traffic scenarios, particularly severe situations such as traffic accidents and congestion, and to assess the effectiveness of various decisions under different conditions. Finally, the optimal decision is executed in the actual systems and continuously refined through real–virtual interactions with artificial systems.

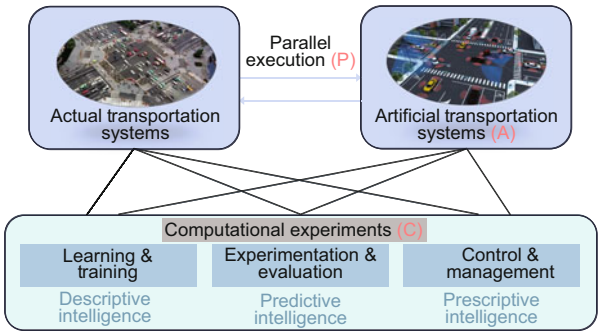

Fig. 1 Parallel transportation

There are three typical modes of connection and operation in parallel transportation (Wang FY, 2008), that is, learning and training, experimentation and evaluation, and control and management. In the learning-and-training mode, operators and administrators interact with various artificial systems to swiftly learn operational procedures for descriptive intelligence, especially gaining experience in handling extreme traffic scenarios. In the experimentation-and-evaluation mode, the systematic, continuous conduction of computational experiments in artificial systems allows us to analyze and predict the behavior of actual systems in various scenarios and thus to estimate and evaluate the performance of different decision-making

for applications, known as predictive intelligence. In the control-and-management mode, artificial systems and actual systems maintain a real-time online connection, where operational parameters are identified and feedback control is generated by using the differences in their behaviors. This allows the rolling-horizon optimization of control strategies, thus enabling the generation of prescriptive intelligence.

3 TransRAG

To build reliable, trustworthy, and safe transportation systems, TransRAG is proposed as an RAG-based parallel transportation framework, as illustrated in Fig. 2, consisting of three layers: storage, management, and execution. The functionality and design details of these layers are as follows.

3.1 Storage layer

The storage layer is the lowest tier, primarily responsible for storing actual data and knowledge and supporting Web-based retrieval functions. The stored data can be updated in real time, aiming to provide more accurate data support for the higher tiers, which can be divided into five types according to their sources and contents: real-time traffic data, traffic engineering knowledge, traffic scenario library, historical traffic data, and social media & survey data.

The real-time traffic data refer to the current and up-to-date traffic and environmental information from diverse terminal sensors and monitoring devices, including traffic flow, vehicle speed, road conditions, congestion, and weather. The traffic engineering knowledge involves professional theories

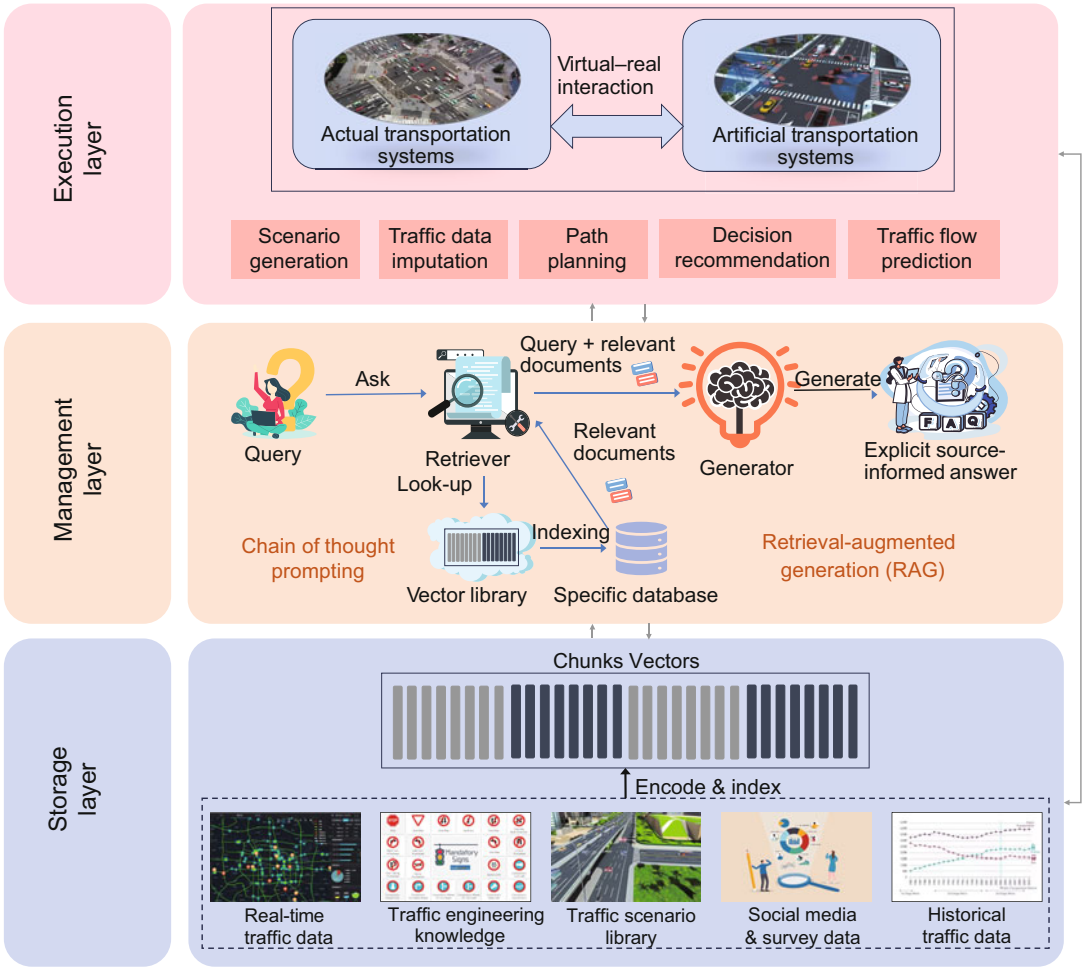

Fig. 2 Framework of TransRAG

and practical experience in areas such as traffic flow analysis, road design, and traffic control and management. The traffic scenario library stores extensive map data, area classifications, and activity details, aiding in the analysis of traffic patterns within specific geographic and social contexts, as well as in the construction of various artificial transportation systems. The historical traffic data include past traffic statistics and essential scenario data, including the number of traffic accidents on specific road segments and traffic volume during particular periods, such as holidays. The social media & survey data refer to the traffic information and knowledge derived from social media and search engines, such as recent traffic policies and rules, the reported details of traffic congestion and accidents, and public reactions to traffic conditions and their needs.

These multi-source data are selected based on some criteria such as reliability, relevance, and timeliness. They are then integrated and stored in relational databases (such as MySQL, SQLite, and PostgreSQL) or non-relational databases (such as MongoDB, Redis, and Cassandra) using data integration platforms and extract–transform–load (ETL) tools (e.g., Stitch, Talend, and Informatica). The weight assigned to the information from each source is adjusted and determined by factors including the source's credibility, task-specific metrics, and user feedback. However, the vast volume of data makes retrieval more time-consuming and resource-intensive. To speed up retrieval and meet FMs' input token limits, all the texts in the database are broken down into several smaller and more approachable chunks. Subsequently, an embedding model is used to encode these chunks as corresponding vector representations, and thus an index is created, where chunks are stored as values and vector representations as keys. This method enables efficient, agile, and scalable search from the storage layer.

3.2 Management layer

The management layer is the middle tier and leverages RAG and CoT to guide FMs in making reliable and interpretable decisions through computational experiments. The entire decision-making process is framed around RAG and can be divided into two stages, retrieval and generation, so the layer includes at least one retriever and one generator. In the retrieval stage, after a query is asked, the same embedding model as the storage layer is leveraged to encode it as a query vector. The similarity between the query vector and each chunk vector from the storage layer is computed to choose the top-K similar chunk vectors. Based on these chunk vectors, the corresponding chunks are indexed and integrated as the relevant information of the query.

In the generation stage, the query, its relevant information, and task-specific prompts are combined as a coherent contextual prompt, which thus is fed into the generator to formulate a response. To enhance the interpretability of generated content, the task-specific prompts are designed in the form of CoT, which compels an FM as the generator to output a series of intermediate steps to reach the desired answer. Depending on the different needs and input modalities (e.g., texts, images, and videos), various FMs or their combinations can be used as generators, such as ChatGPT (Wang FY et al., 2023a) for text-to-text tasks, GPT-4 (Sanderson, 2023) for image/text-to-text tasks, and Sora (Wang FY et al., 2024) for text-to-video tasks.

3.3 Execution layer

The execution layer is the highest tier, where the decisions from the management layer are executed in actual transportation systems in parallel with artificial transportation systems. A variety of artificial systems are built using scenario data from the storage layer, where computational experiments are performed by the management layer. Specifically, RAG is used to create diverse and plausible scenarios based on the existing data, which are then employed in artificial systems where data are generated for specific tasks to support training and testing. During parallel execution, the differences between the two types of systems are fed back into the management layer to continuously optimize decision-making. Additionally, generated intermediate data and tested case studies are transmitted to the storage layer for storage, thereby accumulating experience for future strategy formulation. By this means, the results from solving a range of problems can be tested, executed, and visualized within the layer, such as scenario generation, traffic data imputation, path planning, decision recommendation, and traffic flow prediction, thereby providing diverse personalized services to users.

3.4 Opportunities and challenges

Transportation systems continually enable the movement of people and goods, which are closely intertwined with the economic development of society. Therefore, it is crucial to improve the intelligence of transportation systems through the combination of telecommunications and computer technologies. TransRAG, as a typical intelligent transportation framework, holds promise in achieving more autonomous and reliable transportation management and control, propelling transportation systems toward Transportation 5.0 (Wang FY et al., 2023b). In particular, we expect that ongoing progress in FM technology could allow TransRAG to truly achieve "6S" in Transportation 5.0: Safe in physical spaces, Secure in cyberspaces, Sustainable in ecology, Sensitive in individual privacy and rights, Service for all, and Smartness of all. However, before reaching this goal, there are several challenges to overcome, as shown in Fig. 3.

Personalized transportation systems require an increasing amount of individual data, but regulations such as General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) restrict centralized data collection. Concerns about privacy breaches and data loss for centralized data storage diminish user enthusiasm for data sharing. Additionally, centralized systems are inherently vulnerable to single point of failure, which could jeopardize personal safety. Therefore, a decentralized TransRAG needs to be explored. Blockchain (Yang et al., 2023b; Li et al., 2024), as a decentralized, immutable, and distributed ledger technology, offers a secure, transparent, and automated environment for data storage and system operations. Smart contracts (Wang S et al., 2018; Ouyang et al., 2022), as a piece of code that runs on a blockchain, define the rights, obligations, and triggering conditions for the involved parties, which can be executed automatically without requiring third-party intervention, thereby reducing transaction costs and trust risks. Therefore, the incorporation of blockchain and smart contracts into TransRAG is expected to construct a decentralized, secure, and personalized transportation system.

During the operation of TransRAG, the storage layer transmits the retrieved data to the management or execution layer through the network, which

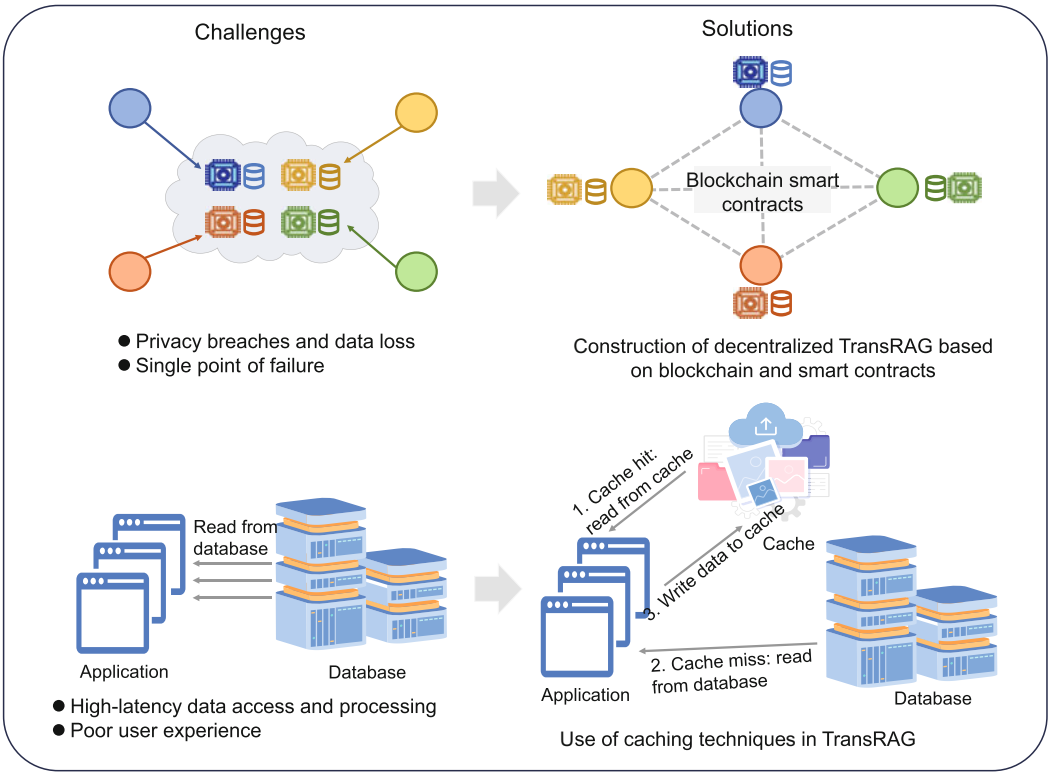

Fig. 3 Challenges faced by TransRAG and their solutions

results in a certain amount of delay and thus a reduction in traffic efficiency. In particular, as data volume increases and transmission frequency rises, insufficient Internet bandwidth leads to high-latency data processing and poor user experience. Therefore, there is an urgent need to develop a method for optimizing the data storage and transmission process to reduce access latency and enhance response speed. Caching (Lai et al., 2001; Zeng et al., 2004) is a technique that stores the results of a request in a location different from the original one or in a temporary storage area to avoid performing the same operations. Obviously, it contributes to the fulfillment of data access demands and the reduction in user-perceived network latency by optimizing the storage location of the data based on the user's preferences. Therefore, caching can be regarded as a promising method for TransRAG's efficient operations.

4 Conclusions

Transportation is a multifaceted and interconnected system involving traffic lights, individuals, roads, vehicles, regulations, and navigation systems. The evolution from Transportation 4.0 to Transportation 5.0 requires a unified approach that seamlessly integrates all these elements, especially human and societal factors, instead of addressing them as separate issues. Considering the general knowledge embedded in FMs and the adaptability shown by RAG technology, the proposed TransRAG for parallel transportation shows promise in advancing the shift toward Transportation 5.0 as a holistic framework. In TransRAG, RAG can integrate knowledge from the external traffic databases into CoT prompts to enhance the accuracy, reliability, and interpretability of FMs in problem-solving, thereby preventing errors that could result in traffic accidents. Finally, the shortcomings of TransRAG, such as privacy breaches, single point of failure, and delays in data access, are analyzed, and the corresponding solutions are offered, including the use of blockchain, smart contracts, and caching. We hope that TransRAG can inspire more work on further advancements in Transportation 5.0.

Contributors

Jing YANG drafted the paper. Xingyuan DAI, Yisheng LV, Levente KOVÁCS, and Fei-Yue WANG helped organize the paper. Jing YANG revised and finalized the paper.

Conflict of interest

Fei-Yue WANG is an executive associate editor-in-chief of *Frontiers of Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

References

Chen JW, Lin HY, Han XP, et al., 2024. Benchmarking large language models in retrieval-augmented generation. Proc 38th AAAI Conf on Artificial Intelligence, p.17754-17762. https://doi.org/10.1609/aaai.v38i16.29728

Dai XY, Guo C, Tang Y, et al., 2024. VistaRAG: toward safe and trustworthy autonomous driving through retrieval-augmented generation. *IEEE Trans Intell Veh*, 9(4):4579-4582. https://doi.org/10.1109/TIV.2024.3396450

Feng GH, Zhang BH, Gu YT, et al., 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. Proc 37th Int Conf on Neural Information Processing Systems, p.70757-70798.

Gan L, Chu WB, Li GF, et al., 2024. Large models for intelligent transportation systems and autonomous vehicles: a survey. *Adv Eng Inform*, 62:102786. https://doi.org/10.1016/j.aei.2024.102786

Jin K, Wi J, Lee E, et al., 2021. TrafficBERT: pre-trained model with large-scale data for long-range traffic flow forecasting. *Expert Syst Appl*, 186:115738. https://doi.org/10.1016/j.eswa.2021.115738

Lai GP, Liu MK, Wang FY, et al., 2001. Web caching: architectures and performance evaluation survey. Proc IEEE Int Conf on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace, p.3039-3044. https://doi.org/10.1109/ICSMC.2001.971982

Lewis P, Perez E, Piktus A, et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proc 34th Int Conf on Neural Information Processing Systems, p.9459-9474.

Li JJ, Qin R, Guan ST, et al., 2024. Blockchain intelligence: intelligent blockchains for Web 3.0 and beyond. *IEEE Trans Syst Man Cybern Syst*, 54(11):6633-6642. https://doi.org/10.1109/TSMC.2023.3348449

Ouyang LW, Zhang WW, Wang FY, 2022. Intelligent contracts: making smart contracts smart for blockchain intelligence. *Comput Electr Eng*, 104:108421. https://doi.org/10.1016/j.compeleceng.2022.108421

Sanderson K, 2023. GPT-4 is here: what scientists think. *Nature*, 615(7954):773. https://doi.org/10.1038/d41586-023-00816-5

Wang FY, 2008. Toward a revolution in transportation operations: AI for complex systems. *IEEE Intell Syst*, 23(6):8-13. https://doi.org/10.1109/MIS.2008.112

Wang FY, 2010. Parallel control and management for intelligent transportation systems: concepts, architectures, and applications. *IEEE Trans Intell Transp Syst*, 11(3):630-638. https://doi.org/10.1109/TITS.2010.2060218

Wang FY, Li JJ, Qin R, et al., 2023a. ChatGPT for computational social systems: from conversational applications to human-oriented operating systems. *IEEE Trans Comput Soc Syst*, 10(2):414-425. https://doi.org/10.1109/TCSS.2023.3252679

Wang FY, Lin YL, Ioannou PA, et al., 2023b. Transportation 5.0: the DAO to safe, secure, and sustainable intelligent transportation systems. *IEEE Trans Intell Transp Syst*, 24(10):10262-10278. https://doi.org/10.1109/TITS.2023.3305380

Wang FY, Miao QH, Li LX, et al., 2024. When does Sora show: the beginning of TAO to imaginative intelligence and scenarios engineering. *IEEE/CAA J Autom Sin*, 11(4):809-815. https://doi.org/10.1109/JAS.2024.124383

Wang S, Yuan Y, Wang X, et al., 2018. An overview of smart contract: architecture, applications, and future trends. IEEE Intelligent Vehicles Symp, p.108-113. https://doi.org/10.1109/IVS.2018.8500488

Wang X, Yang J, Han JP, et al., 2022. Metaverses and DeMetaverses: from digital twins in CPS to parallel intelligence in CPSS. *IEEE Intell Syst*, 37(4):97-102. https://doi.org/10.1109/MIS.2022.3196592

Wang X, Wang YT, Yang J, et al., 2024. The survey on multi-source data fusion in cyber-physical-social systems: foundational infrastructure for industrial metaverses and Industries 5.0. *Inform Fus*, 107:102321. https://doi.org/10.1016/j.inffus.2024.102321

Wei J, Wang XZ, Schuurmans D, et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Proc 36$^{\text{th}}$ Int Conf on Neural Information Processing Systems, p.24824-24837.

Yang J, Wang X, Tian YL, et al., 2023a. Parallel intelligence in CPSSs: being, becoming, and believing. *IEEE Intell Syst*, 38(6):75-80. https://doi.org/10.1109/MIS.2023.3284694

Yang J, Ni QH, Luo GY, et al., 2023b. A trustworthy Internet of Vehicles: the DAO to safe, secure, and collaborative autonomous driving. *IEEE Trans Intell Veh*, 8(12):4678-4681. https://doi.org/10.1109/TIV.2023.3337345

Zeng D, Wang FY, Liu MK, 2004. Efficient web content delivery using proxy caching techniques. *IEEE Trans Syst Man Cybern C Appl Rev*, 34(3):270-280. https://doi.org/10.1109/TSMCC.2004.829261

Zhang KP, Zhou F, Wu L, et al., 2024. Semantic understanding and prompt engineering for large-scale traffic data imputation. *Inform Fus*, 102:102038. https://doi.org/10.1016/j.inffus.2023.102038

Zhou J, Ke P, Qiu XP, et al., 2023. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, early access. https://doi.org/10.1631/FITEE.2300089

Zhu FH, Lv YS, Chen YY, et al., 2020. Parallel transportation systems: toward IoT-enabled smart urban traffic control and management. *IEEE Trans Intell Transp Syst*, 21(10):4063-4071. https://doi.org/10.1109/TITS.2019.2934991