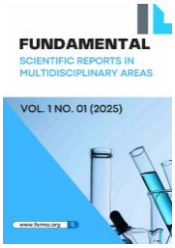
Fundamental Scientific Reports in Multidisciplinary Areas

Vol: 02 Issue:03 2026

E-ISSN: 3134-7517 P-ISSN: 3134-7509



ARTICLE HISTORY

Received: [10 October 2026]
Revised: [17 October 2026]
Accepted: [24 October 2026]

Published online: [31 October 2026]

KEYWORDS

GeoAI; Multimodal Learning; Large Language Model; Vision-Language Model; Multi-Agent System; Spatial Intelligence; Public Participation; Smart City; Urban Computing

LICENSE

This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). You may share and adapt the material in any medium or format, with appropriate credit to the original author(s) and a link to the licence.
https://creativecommons.org/licenses/by/4.0/

COPYRIGHT

© 2026 The Author(s). Authors retain copyright; FSRMA holds a non-exclusive licence to publish.

ORIGINAL RESEARCH ARTICLE

GeoCivic-Agent: A Multimodal Agentic GeoAI Framework for Spatial Understanding and Decision Support of Public Participation Data in Smart Cities

Wenfang Lu [1], Pinghao Luo [2]

[1] *Metropolitan Area Planning Council, Boston, MA, USA*
[2] *University of Southern California, Los Angeles, CA, USA*
[1] *wenfang@vt.edu*
[2] *pinghaol@alumni.usc.edu*

ABSTRACT

The rapid development of smart cities has increased the demand for intelligent and participatory urban governance, where public opinions provide valuable insights into transportation safety, environmental quality, housing conditions, and public service deficiencies. However, massive civic feedback collected from online platforms, social media, and mobile applications is often presented as heterogeneous and unstructured data, including texts, images, and geographic information, which limits the ability of existing urban analysis systems to discover spatial patterns and support planning decisions. This study proposes GeoCivic-Agent (Multimodal Agentic Geographic Artificial Intelligence Framework for Spatial Understanding and Decision Support of Public Participation Data), a novel framework that integrates Geographic Artificial Intelligence (GeoAI), Large Language Models (LLMs), Vision-Language Models (VLMs), and multi-agent reasoning for intelligent civic data analysis. The proposed framework employs a multimodal perception module, a GeoFusion Transformer for spatial-semantic representation learning, and a GIS-enhanced Retrieval-Augmented Generation (RAG) module to identify urban issues, detect spatial hotspots, and generate explainable planning recommendations. Experimental results on the multimodal urban participation dataset demonstrate that GeoCivic-Agent achieves an F1-score of 94.6% for urban issue classification, outperforming the GPT-based single-agent baseline by 2.8 percentage points. For spatial hotspot detection, the proposed framework obtains an AUC of 96.2% and reduces MAE to 0.087, showing superior capability in discovering urban concern patterns and generating reliable spatial intelligence compared with existing GeoAI and LLM-based approaches.

1. Introduction

The rapid evolution of smart cities has significantly transformed urban governance from traditional expert-driven decision-making toward more participatory, data-driven, and human-centered planning paradigms. Public participation has become an essential mechanism for understanding residents' perceptions, identifying urban challenges, and improving the effectiveness of planning interventions. Through online consultation platforms, social media, mobile applications, and digital complaint systems, citizens continuously generate valuable information regarding transportation accessibility, environmental conditions, housing pressures, public safety, and infrastructure deficiencies. However, these large-scale civic feedback sources are usually heterogeneous and unstructured, consisting of textual descriptions, images, geographic locations, and contextual information, making it difficult for conventional urban analytics systems to extract meaningful spatial knowledge and support intelligent planning decisions.

Recent advances in Geographic Artificial Intelligence (GeoAI) have provided new opportunities for integrating artificial intelligence techniques with geographic information systems (GIS) to analyze complex urban phenomena. Existing GeoAI approaches have achieved promising performance in areas such as remote sensing interpretation, traffic prediction, and environmental monitoring. Nevertheless, most existing methods mainly focus on structured spatial data or sensor-based observations and lack the capability to understand human-generated multimodal information. In real-world urban scenarios, public opinions often contain implicit spatial relationships and complex semantic meanings. For example, a citizen complaint describing "unsafe streets near schools at night" requires simultaneous interpretation of textual semantics, visual evidence, geographic context, and urban functional characteristics. Therefore, developing a multimodal spatial intelligence framework that can transform fragmented public opinions into interpretable geographic knowledge remains a critical research challenge. To address these limitations, this study proposes GeoCivic-Agent (Multimodal Agentic Geographic Artificial Intelligence Framework for Spatial Understanding and Decision Support of Public Participation Data), a novel framework that integrates multimodal learning, large language models (LLMs), vision-language models (VLMs), spatial representation learning, and autonomous agent-based reasoning. GeoCivic-Agent consists of four major components: (1) a multimodal civic perception module that extracts semantic information from textual opinions and urban images; (2) a GeoFusion Transformer that jointly learns textual, visual, and geographic representations through spatial-aware attention mechanisms; (3) a multi-agent reasoning architecture that enables civic issue interpretation, spatial relationship analysis, and planning-oriented reasoning; and (4) a GIS-enhanced Retrieval-Augmented Generation (RAG) module that incorporates urban knowledge bases and planning regulations to generate reliable decision-support recommendations. Through these components, GeoCivic-Agent establishes an end-to-end pipeline from public opinion understanding to spatial hotspot discovery and urban planning assistance.

The main contributions of this study are summarized as follows:

(1) A novel multimodal Agentic GeoAI framework is proposed. GeoCivic-Agent integrates LLMs, VLMs, GIS information, and multi-agent reasoning mechanisms to achieve comprehensive understanding of heterogeneous public participation data.

(2) A spatial-aware multimodal fusion mechanism is developed. The proposed GeoFusion Transformer enables effective interaction among textual semantics, visual features, and geographic contexts, improving the representation capability of civic opinions in urban environments.

(3) An intelligent spatial reasoning and decision-support mechanism is introduced. By combining multiple specialized agents with GIS-based knowledge retrieval, GeoCivic-Agent can identify urban problem hotspots and generate explainable planning recommendations.

(4) A comprehensive evaluation framework is established for civic spatial intelligence tasks. Experimental results demonstrate the effectiveness of GeoCivic-Agent in urban issue classification, spatial hotspot

© 2026 The Author(s). Licensed under CC BY 4.0

detection, and participatory planning support, providing a practical solution for human-centered smart city governance.

2.Literature Review

Recent advances in artificial intelligence (AI), geographic information science, and large-scale foundation models have significantly promoted the development of intelligent urban analysis systems. In particular, the integration of Geographic Artificial Intelligence (GeoAI), multimodal learning, and large language models (LLMs) has provided new opportunities for understanding complex urban phenomena from heterogeneous data sources. Existing studies related to this research can be mainly categorized into four directions: GeoAI-based urban spatial intelligence, multimodal learning for geographic understanding, large language model–enhanced urban planning, and agent-based intelligent decision-support systems. This section reviews representative studies in these fields and discusses their limitations in supporting multimodal spatial perception of public participation data.

2.1 GeoAI-Based Urban Spatial Intelligence and Geographic Data Analysis

Geographic Artificial Intelligence (GeoAI) has emerged as an interdisciplinary research paradigm that integrates artificial intelligence techniques with geographic information systems (GIS) to analyze complex spatial phenomena. Early GeoAI studies mainly focused on applying machine learning and deep learning models to remote sensing interpretation, land-use classification, environmental monitoring, and urban pattern recognition.

Goodchild [1] introduced the concept of geographic information science and highlighted the importance of computational intelligence in discovering spatial knowledge from large-scale geographic datasets. Subsequently, Janowicz et al. [2] proposed the concept of GeoAI as a new research direction that combines spatial science with deep learning techniques, emphasizing the potential of artificial intelligence in geographic knowledge discovery.

With the development of deep neural networks, researchers have explored convolutional neural networks (CNNs), graph neural networks (GNNs), and spatial-temporal models for urban applications. For example, Yang et al. [3] proposed a deep learning framework for urban land-use classification by integrating multiple geographic features, demonstrating the effectiveness of neural representation learning in spatial analysis. However, these approaches mainly rely on structured geographic datasets, such as satellite imagery, transportation networks, and demographic statistics, while overlooking human-generated information such as citizen comments, social media posts, and visual feedback.

Recently, GeoAI research has gradually shifted toward human-centered spatial intelligence. Yuan et al. [4] proposed OmniGeo, a multimodal large language model specifically designed for geospatial artificial intelligence tasks. Their study demonstrated that integrating textual, visual, and geographic information can significantly improve spatial reasoning capabilities across different GeoAI applications. Nevertheless, existing GeoAI models still lack specialized mechanisms for understanding public participation data and transforming citizen-generated opinions into actionable urban planning knowledge.

2.2 Multimodal Learning for Geographic and Urban Understanding

The rapid development of multimodal learning has enabled artificial intelligence models to jointly process information from different modalities, including text, images, and spatial data. Vision-language models (VLMs), particularly Contrastive Language–Image Pre-training (CLIP)-based approaches, have demonstrated strong capability in aligning visual and semantic representations.

Radford et al. [5] introduced CLIP, which learns a shared embedding space between images and natural language through large-scale image-text contrastive learning. CLIP has been widely adopted in geographic applications because it enables zero-shot recognition and semantic understanding without requiring extensive labeled datasets. Building upon vision-language models, Huang et al. [6] investigated zero-shot urban function inference using street-view images and pretrained vision-language models. Their research demonstrated that VLMs can effectively identify urban functional characteristics from street-level imagery,

© 2026 The Author(s). Licensed under CC BY 4.0

providing new opportunities for intelligent urban perception. However, these approaches mainly focus on visual environment understanding and do not consider citizen opinions, emotional information, and spatial reasoning simultaneously. Recent studies have further explored multimodal foundation models for geographic applications. Xu and Tao [7] investigated the capability of GPT-4 Vision in map interpretation and geographic analysis, showing the potential of multimodal large language models (MLLMs) for understanding complex spatial information. However, existing multimodal models still face challenges in integrating dynamic civic feedback with geographic contexts, especially when dealing with ambiguous spatial descriptions and subjective human perceptions. Therefore, a dedicated multimodal GeoAI framework is required to combine textual opinions, visual evidence, and spatial information for public participation analysis.

2.3 Large Language Models for Participatory Urban Planning

The emergence of large language models (LLMs) has significantly changed the landscape of urban planning research by enabling semantic understanding, reasoning, and knowledge generation. Unlike traditional machine learning approaches, LLMs provide strong capabilities in processing natural language descriptions and generating human-like planning recommendations.

Zhou et al. [8] explored the application of LLM-based multi-agent collaboration for participatory urban planning. Their framework simulated different stakeholders through multiple LLM agents and demonstrated that AI agents could assist planners in considering diverse resident preferences. However, the study mainly focused on text-based participation simulation and did not incorporate real-world multimodal civic data or geographic reasoning. Similarly, Zhu et al. [9] proposed PlanGPT, a domain-specific language model designed for urban planning tasks. By integrating retrieval mechanisms and planning knowledge, PlanGPT improved the generation quality of planning-related responses. However, the model primarily focuses on document-level planning assistance and lacks spatial perception capability for identifying geographic hotspots from citizen-generated data. Recent research has investigated direct applications of LLMs in citizen participatory planning. Lee et al. [10] examined the use of GPT-4o and prompt engineering techniques for representing citizen opinions in urban planning scenarios. Their findings demonstrated that LLMs can effectively analyze community concerns and support participatory decision-making. Nevertheless, these approaches mainly rely on textual interaction and do not fully exploit multimodal spatial information, limiting their ability to perform comprehensive urban perception.

2.4 Agent-Based Artificial Intelligence for Urban Decision Support

Beyond individual AI models, autonomous agent-based systems have recently attracted increasing attention due to their ability to perform complex reasoning, collaboration, and decision-making tasks. Multi-agent systems allow different specialized agents to cooperate by dividing complex problems into smaller reasoning processes. In urban computing, agent-based AI systems have been explored for simulation, planning optimization, and decision support. Park et al. [11] demonstrated that LLM-based agents can simulate human-like behaviors and support large-scale social decision-making tasks. These studies indicate that agent collaboration provides a promising direction for complex urban governance problems.

However, existing agent-based urban applications mainly focus on simulation and planning generation, while limited research has investigated agent frameworks for multimodal civic data understanding. In particular, existing systems rarely integrate:

(1) citizen opinion extraction;
(2) visual urban condition recognition;
(3) spatial relationship reasoning;
(4) GIS-based knowledge retrieval.

To overcome these limitations, this study proposes GeoCivic-Agent, which combines multimodal GeoAI perception, spatial-aware representation learning, and multi-agent reasoning mechanisms. Unlike previous

© 2026 The Author(s). Licensed under CC BY 4.0

approaches, GeoCivic-Agent establishes an end-to-end framework that converts heterogeneous public participation data into spatial intelligence and explainable urban planning recommendations.

3.Methodology

This section presents the proposed GeoCivic-Agent (Multimodal Agentic Geographic Artificial Intelligence Framework for Spatial Understanding and Decision Support of Public Participation Data) framework for transforming heterogeneous public participation data into spatially interpretable urban intelligence. Unlike conventional GeoAI approaches that mainly analyze structured geographic information, GeoCivic-Agent is designed to understand multimodal civic information, including textual opinions, urban images, social media content, and geographic contexts. The proposed framework integrates multimodal representation learning, spatial-aware feature fusion, multi-agent reasoning, and GIS-enhanced knowledge retrieval to achieve an end-to-end pipeline from public opinion perception to urban planning decision support.

The overall architecture of GeoCivic-Agent consists of four major components: (1) a Multimodal Civic Perception Module for extracting semantic and visual information from citizen-generated data; (2) a GeoFusion Transformer for learning joint representations across language, vision, and geographic spaces; (3) a Multi-Agent Spatial Reasoning Module for interpreting urban issues and discovering spatial relationships; and (4) a GIS-enhanced Retrieval-Augmented Generation (RAG) module for generating explainable planning recommendations. Through the collaboration of these components, the framework enables spatial understanding of public concerns and supports human-centered smart city governance.

3.1 Multimodal Civic Perception Module

Public participation data are naturally heterogeneous because citizens express their opinions through different channels, including online comments, complaint platforms, social media posts, and image-sharing applications. Therefore, the first stage of GeoCivic-Agent focuses on extracting meaningful representations from multiple modalities, including text, images, and geographic information.

For textual civic opinions, a Large Language Model (LLM) is employed as the semantic encoder to capture high-level linguistic patterns and contextual meanings. Given a citizen statement X_t, the textual representation is obtained as:

$$H_t = f_{LLM}(X_t)$$

where Ht denotes the semantic embedding generated by the LLM. In addition to representation learning, the LLM performs structured information extraction through instruction-based prompting. Specifically, each opinion is decomposed into four major attributes:

$$O_i = (L_i, C_i, S_i, T_i)$$

where L_i represents the geographic location, C_i denotes the urban issue category, S_i represents sentiment polarity, and T_i indicates temporal information. For example, the sentence “The pedestrian road near the primary school is unsafe at night” can be transformed into a structured representation containing location information, pedestrian safety concerns, negative sentiment, and nighttime conditions.

For visual civic information, GeoCivic-Agent introduces a Vision-Language Model (VLM) based on Vision Transformer (ViT) and Contrastive Language–Image Pre-training (CLIP) mechanisms. Urban images uploaded by residents, such as damaged roads, insufficient lighting facilities, garbage accumulation, and unsafe public spaces, are encoded into visual feature representations:

$$H_v = f_{VLM}(X_v)$$

where X_v represents the input urban image and H_v denotes the extracted visual embedding. Traditional image classification models, typically built upon convolutional backbones such as VGG16, MobileNet, and ResNet152, have achieved strong performance in multi-class classification tasks across diverse domains [13]. However, such models depend on closed-set category labels and extensive task-specific annotations, which restricts their adaptability to the open-ended and continuously evolving urban issue categories embedded in civic feedback. The VLM-based encoder is therefore adopted to learn shared visual–semantic

© 2026 The Author(s). Licensed under CC BY 4.0

representations in an open-vocabulary manner. Compared with traditional image classification models, the VLM-based encoder provides stronger semantic alignment between visual observations and textual urban concepts, enabling the system to associate visual evidence with corresponding civic concerns.

To incorporate spatial context, a geographic encoder is designed to model the relationship between public opinions and urban environments. The geographic input contains coordinates, road networks, points of interest (POIs), land-use categories, and demographic information. The spatial representation is calculated as:

$$H_g = f_{geo}(X_g)$$

where X_g represents geographic attributes and H_g denotes spatial embeddings. By integrating geographic features, the model can distinguish different urban contexts, such as residential communities, commercial areas, transportation hubs, and educational zones.

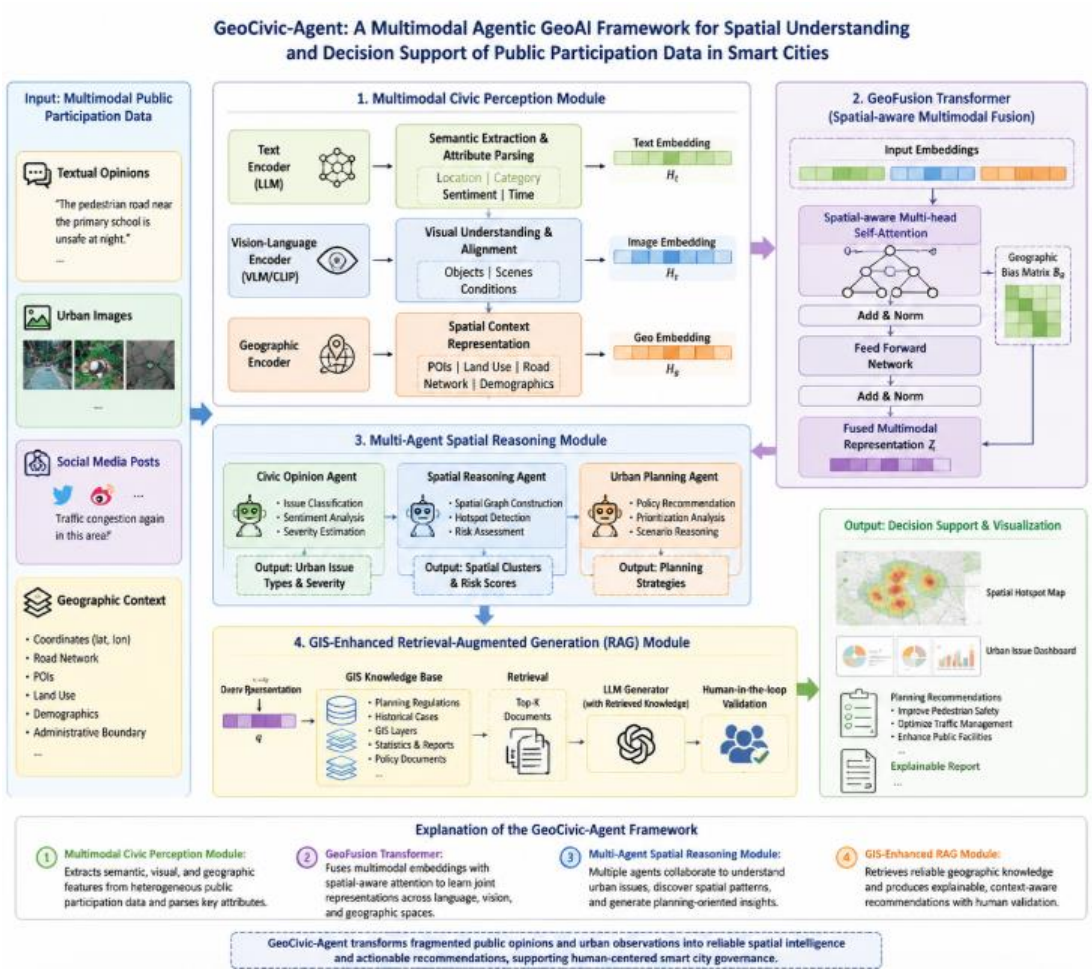


Figure 1Overall flowchart of the model.

3.2 GeoFusion Transformer for Multimodal Spatial Representation Learning

After extracting representations from different modalities, GeoCivic-Agent introduces a spatial-aware multimodal fusion mechanism named GeoFusion Transformer. The objective of this module is to learn the interactions among textual opinions, visual evidence, and geographic contexts.

Traditional Transformer architectures mainly focus on semantic relationships within a single modality, while ignoring geographic dependencies. Therefore, GeoFusion Transformer extends the standard self-attention mechanism by incorporating spatial attention constraints.

The multimodal feature sequence is formulated as:

© 2026 The Author(s). Licensed under CC BY 4.0

$$Z = [H_t, H_v, H_g]$$

where Z represents the combined multimodal representation.
The attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

where Q, K, and V are query, key, and value matrices, respectively, and d represents the feature dimension. To enhance spatial awareness, a geographic bias matrix B_g is introduced:

$$\text{Attentiong} = \text{softmax}(\frac{QK^T}{\sqrt{d}} + B_g)V$$

where B_g measures geographic relationships between different civic events. For example, complaints occurring within the same neighborhood or along connected road networks receive higher spatial attention weights. Through this mechanism, GeoFusion Transformer can establish complex relationships among multimodal information. A citizen comment describing "dangerous traffic conditions", an image showing crowded intersections, and geographic information indicating proximity to schools can be jointly interpreted as a potential pedestrian safety risk zone. This spatial-semantic fusion capability enables GeoCivic-Agent to move beyond simple opinion classification toward deeper urban issue understanding. Such structured-knowledge injection into attention follows a broader architectural trend: enriching transformer attention with external relational knowledge has been shown to improve detection performance in complex data streams. For instance, Qiu et al. [14] proposed CyberDQFormer, a knowledge-graph-enhanced transformer framework for security-aware data quality anomaly detection, demonstrating that injecting structured relational knowledge into attention substantially improves the identification of anomalies. In a similar spirit, the geographic bias matrix Bg injects urban relational knowledge—spatial proximity, road-network connectivity, and functional associations among civic events—into the attention computation, enabling the fusion of multimodal semantics with geographic structure.

3.3 Multi-Agent Spatial Reasoning Module

To achieve autonomous urban analysis, GeoCivic-Agent introduces a multi-agent reasoning architecture. Instead of relying on a single general-purpose model, different agents are assigned specialized responsibilities to improve interpretability and reasoning capability.

Such role specialization follows recent multi-agent frameworks for structured collaborative decision-making. For instance, Zhang et al. [15] demonstrated that instantiating multiple role-specialized LLM agents and aggregating their structured outputs—including preference scores, confidence estimates, and evidence chains—through a shared consensus matrix improves the reliability of collaborative decisions in agentic visual media workflows. Following this paradigm, GeoCivic-Agent assigns complementary roles to the Civic Opinion Agent, the Spatial Reasoning Agent, and the Urban Planning Agent, so that perception, spatial analysis, and planning generation are performed by dedicated specialists whose outputs are integrated through structured reasoning. The first component is the Civic Opinion Agent, which analyzes multimodal representations generated by GeoFusion Transformer and identifies major urban concerns. It performs issue classification, sentiment assessment, and severity estimation. The output can be represented as:

$$Y_c = f_c(Z)$$

where Y_c contains the identified urban problem categories, such as transportation safety, environmental pollution, housing pressure, and public facility deficiencies.

The second component is the Spatial Reasoning Agent, which focuses on discovering geographic relationships among civic events. This agent constructs a spatial opinion graph:

$$G = (V, E)$$

© 2026 The Author(s). Licensed under CC BY 4.0

where each node V represents a geographic civic event and each edge E represents spatial, semantic, or functional relationships between events. Graph-based reasoning is performed to identify spatial clusters and emerging urban risks. This graph-based reasoning design is motivated by the demonstrated effectiveness of graph neural networks in capturing relational context for language-intensive tasks. Gao et al. [16] showed that modeling contextual relations—such as co-occurrence and session-level dependencies—through a heterogeneous graph significantly improves cross-lingual query understanding in sponsored search. In a similar spirit, the Spatial Reasoning Agent exploits the relational structure of civic events to capture spatial, semantic, and functional dependencies among complaints, which are inaccessible from independent event representations and are essential for reliable hotspot identification. The spatial importance of a region r is calculated as:

$$\mathrm{Risk}(r) = \alpha D(r) + \beta N(r) + \gamma S(r)$$

Where $D(r)$ denotes complaint density, $N(r)$ represents neighborhood relevance, and $S(r)$ represents negative sentiment intensity. The parameters α, β, and γ control the contribution of different factors. These weighting parameters are currently determined empirically. Adaptive optimization paradigms based on evolutionary algorithms have proven effective for tuning complex configurations in large-scale AI systems, as demonstrated by the adaptive evolutionary approach for optimizing AI server testing proposed by Ren [17]. Automatically optimizing these weights and other framework hyperparameters through such adaptive search is a promising direction for reducing manual tuning effort and enhancing robustness across diverse urban contexts. The third component is the Urban Planning Agent, which transforms analytical results into planning-oriented recommendations. Based on identified problems and spatial risks, this agent generates actionable suggestions, such as improving pedestrian infrastructure, optimizing traffic management, or increasing public facilities. The recommendation generation of the Urban Planning Agent follows an LLM-guided decision paradigm that has proven effective in other complex operational settings. For instance, Hua et al. [18] proposed LPG-SAC-TCO, an LLM-policy-guided Soft Actor-Critic framework for total-cost-of-ownership-optimal energy management of hydrogen fuel cell heavy-duty trucks, demonstrating that language-model guidance can effectively steer decision-making toward optimal solutions under explicit resource and cost constraints. In a similar spirit, the Urban Planning Agent leverages language-model reasoning—together with spatial risk assessments derived from the risk map—to select planning actions that balance multiple objectives, including safety improvement, traffic efficiency, and public facility allocation, within real-world resource constraints. This mechanism enables GeoCivic-Agent to bridge the gap between AI-based urban perception and practical planning decision-making.

3.4 GIS-Enhanced Retrieval-Augmented Generation Module

Although LLMs demonstrate strong reasoning capability, they may generate inaccurate recommendations without domain-specific knowledge. To improve reliability, GeoCivic-Agent incorporates a GIS-enhanced Retrieval-Augmented Generation (RAG) module.

A geographic knowledge base is constructed by integrating multiple urban information sources, including GIS layers, urban planning regulations, historical governance cases, transportation data, and environmental records. When the Urban Planning Agent receives a query, the retrieval process first converts the query into an embedding vector:

$$q = f_{\mathrm{embed}}(X_q)$$

Then, relevant geographic documents are retrieved according to semantic similarity:

$$D^* = \arg\max_{D} \mathrm{Sim}(q, D)$$

where D^* represents the retrieved knowledge and $\mathrm{Sim}(\cdot)$ denotes similarity measurement.

The retrieved information is combined with multimodal analysis results and provided to the LLM for final reasoning:

$$\mathrm{Response} = \mathrm{LLM}(Z, D^*)$$

© 2026 The Author(s). Licensed under CC BY 4.0

Through GIS-enhanced retrieval, GeoCivic-Agent can generate recommendations that are not only linguistically reasonable but also geographically consistent with actual urban conditions.

3.5 Spatial Perception and Decision Support Workflow

The complete workflow of GeoCivic-Agent follows a perception–reasoning–decision paradigm. First, heterogeneous public participation data are collected and encoded through multimodal perception modules. Second, GeoFusion Transformer integrates semantic, visual, and geographic information to generate unified urban representations. Third, multiple specialized agents collaboratively analyze civic concerns, identify spatial hotspots, and evaluate urban risks. Finally, the GIS-RAG module enhances the generated recommendations using external geographic knowledge. Compared with existing GeoAI and LLM-based urban analysis approaches, GeoCivic-Agent provides three major advantages: it enables multimodal understanding of citizen-generated data, performs explicit spatial reasoning across urban contexts, and produces explainable planning recommendations. Therefore, the proposed framework provides a comprehensive solution for building intelligent public participation opinion spatial perception systems in smart cities.

4. Experiment

4.1 Dataset Preparation

To evaluate the effectiveness of GeoCivic-Agent for multimodal public participation opinion spatial perception, this study constructs a multimodal urban civic dataset by integrating publicly available citizen feedback sources, geospatial information databases, and urban environmental records. The dataset is collected from three major sources: online public complaint platforms, social media posts with geographic tags, and volunteered geographic information (VGI) platforms. Specifically, textual opinions are obtained from urban service request systems and public consultation platforms, where residents report problems related to transportation, environmental quality, public facilities, and community services. Visual data are collected from citizen-uploaded street images and urban observation platforms, while geographic attributes are obtained from GIS databases, including road networks, points of interest (POIs), land-use information, and administrative boundaries. The constructed dataset contains 85,000 multimodal civic records, including 60,000 textual opinions, 15,000 urban images, and 10,000 spatially referenced events. Each record is associated with semantic, visual, and geographic attributes to support multimodal representation learning and spatial reasoning. The main characteristics of the dataset are summarized in Table 1.

Table 1Characteristics of the Multimodal Civic Participation Dataset.

Feature	Description	Data Type
Opinion Text	Citizen comments describing urban problems and demands	Text
Issue Category	Classification of urban concerns, including transportation, environment, housing, and safety	Categorical
Sentiment Score	Emotional polarity indicating positive, neutral, or negative attitudes	Numerical/Categorical
Image Content	Urban visual evidence such as road conditions, facilities, and environmental issues	Image
Latitude and Longitude	Geographic coordinates of reported events	Spatial
POI Context	Nearby facilities, buildings, and urban functional zones	Spatial
Land-use Type	Residential, commercial, educational, industrial, and public areas	Categorical
Timestamp	Reporting time indicating temporal patterns of urban issues	Temporal

© 2026 The Author(s). Licensed under CC BY 4.0

Before model training, all records are processed through data cleaning, geocoding, image normalization, and semantic annotation procedures. Textual opinions are labeled according to urban issue categories, while images are annotated with corresponding urban conditions. The final dataset enables comprehensive evaluation of GeoCivic-Agent on multimodal civic understanding, spatial hotspot detection, and urban decision-support tasks.

4.2 Experimental Setup

To evaluate the effectiveness of GeoCivic-Agent (Multimodal Agentic Geographic Artificial Intelligence Framework for Spatial Understanding and Decision Support of Public Participation Data), extensive experiments are conducted on the constructed multimodal civic participation dataset introduced in Section 3. The dataset is randomly divided into training, validation, and testing subsets with a ratio of 70%, 10%, and 20%, respectively. All experiments are implemented using Python and PyTorch on a workstation equipped with an NVIDIA RTX 4090 GPU, an Intel Xeon Gold 6338 CPU, and 128 GB RAM. The proposed framework adopts a large language model (LLM) backbone for textual representation learning, a Vision-Language Model (VLM) based on CLIP-ViT for visual feature extraction, and a Transformer-based fusion architecture for multimodal representation learning. During training, the AdamW optimizer is applied with an initial learning rate of 1×10^{-5}, a batch size of 32, and 50 training epochs. The performance of GeoCivic-Agent is compared with several representative baseline methods, including traditional machine learning approaches, deep learning-based multimodal models, graph-based GeoAI models, and single-agent large language model approaches. All models are trained and evaluated under the same dataset partition and computational environment to ensure fair comparison.

4.3 Evaluation Metrics

To comprehensively evaluate GeoCivic-Agent for multimodal civic opinion spatial perception, three categories of evaluation metrics are adopted, including classification performance, spatial hotspot detection capability, and decision-support quality. For urban issue classification, Accuracy, Precision, Recall, and F1-score are employed to measure the ability of different models to identify civic problems such as transportation risks, environmental issues, and infrastructure deficiencies. The F1-score is calculated as the harmonic mean of Precision and Recall, which is particularly suitable for imbalanced urban issue categories. For spatial hotspot detection, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Area Under the Receiver Operating Characteristic Curve (AUC) are used to evaluate the accuracy of identifying high-risk urban regions. In addition, recommendation quality is assessed through human evaluation and semantic similarity metrics, including planning consistency and explanation completeness, to measure whether the generated recommendations are consistent with geographic conditions and urban planning knowledge. These metrics jointly evaluate the capability of GeoCivic-Agent from perception, spatial reasoning, and decision-support perspectives.

4.4 Results

The experimental results provide evidence of the effectiveness of GeoCivic-Agent in understanding and classifying urban issues from multimodal civic participation data. As shown in Table 2, GeoCivic-Agent achieves higher scores across all evaluation metrics compared with the baseline models, indicating the effectiveness of the proposed framework for urban issue classification. Table 2 compares the performance of the evaluated approaches on the urban issue classification task. The results reveal a clear progression from conventional text-based methods toward models that incorporate multimodal and spatial information. The BERT-based classifier achieves an accuracy of 84.7% and an F1-score of 83.3%, providing a baseline for classification using textual information alone. Although the model can capture semantic relationships within civic comments, its relatively lower performance suggests that textual descriptions alone may not adequately represent the visual and geographic context associated with urban issues. The CNN-LSTM multimodal model improves the accuracy to 87.5% and the F1-score to 86.5%, representing a 3.2 percentage-point improvement in F1-score over the BERT-based classifier. This improvement indicates that

© 2026 The Author(s). Licensed under CC BY 4.0

incorporating visual and sequential information provides additional signals for distinguishing among urban issue categories. However, the relatively modest gain suggests that simply combining multiple modalities does not fully address the spatial characteristics of civic participation data. The GNN-based GeoAI model achieves an accuracy of 91.2% and an F1-score of 90.6%, producing a further 1.7 percentage-point improvement in F1-score over the CLIP-based approach. This performance gain demonstrates the value of explicitly representing geographic relationships in the classification process. By modeling spatial dependencies among observations, the GNN-based approach can incorporate contextual information that is not available from independent text-image representations. The GPT-based Single-Agent Model achieves an accuracy of 92.4% and an F1-score of 91.8%, making it the strongest baseline. Its performance indicates that large-scale language reasoning can effectively integrate heterogeneous descriptions and infer semantic relationships within complex civic information. However, the remaining performance gap relative to GeoCivic-Agent suggests that general-purpose language reasoning alone does not fully capture the multimodal and spatial structure of the task. GeoCivic-Agent achieves the highest performance across all four evaluation metrics, with an accuracy of 95.1%, precision of 94.8%, recall of 94.5%, and F1-score of 94.6%. Relative to the strongest GPT-based baseline, GeoCivic-Agent improves accuracy by 2.7 percentage points and F1-score by 2.8 percentage points. The improvement is also consistent across precision and recall, increasing by 2.8 and 2.9 percentage points, respectively. This balanced improvement is important because it indicates that the performance gain is not driven solely by improved detection of one aspect of classification. Instead, the results suggest more consistent identification of relevant urban issue categories while maintaining a relatively low rate of both false-positive and false-negative predictions.

Table 2Performance Comparison on Urban Issue Classification Task

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
BERT-based Classifier	84.7	83.9	82.8	83.3
CNN-LSTM Multimodal Model	87.5	86.8	86.2	86.5
CLIP-based Fusion Model	89.6	89.1	88.7	88.9
Graph Neural Network (GNN) GeoAI Model	91.2	90.8	90.4	90.6
GPT-based Single-Agent Model	92.4	92.0	91.6	91.8
GeoCivic-Agent (Ours)	**95.1**	**94.8**	**94.5**	**94.6**

The progression across the models provides further evidence of the contribution of different information sources. Performance improves from text-only modeling to multimodal representation, then to spatial modeling and language-based reasoning, with the largest overall performance achieved when these capabilities are integrated within GeoCivic-Agent. The results therefore indicate that urban issue classification benefits from integrating textual descriptions, visual characteristics, spatial context, and collective reasoning over these contextual signals.

Table 3 evaluates the performance of different models in spatial hotspot detection and urban planning decision support. The Random Forest Geo model achieves an MAE of 0.184 and an AUC of 82.5%, showing limited capability in capturing complex spatial patterns from civic feedback. The CNN-GIS fusion approach reduces the MAE to 0.156 and improves AUC to 86.7% by combining visual information with geographic features. The ST-GNN spatial model further improves hotspot identification with an MAE of 0.132 and an AUC of 90.4%, demonstrating the effectiveness of graph-based spatial relationship modeling. The competitiveness of the ST-GNN baseline is consistent with the demonstrated effectiveness of spatio-temporal GNNs for city-scale dynamic tasks, where hierarchical and edge-enhanced graph structures have

© 2026 The Author(s). Licensed under CC BY 4.0

been shown to capture complex spatial-temporal urban dependencies (Han et al. [19]). The LLM-RAG model achieves an AUC of 92.1% and planning consistency of 85.7%, indicating that external knowledge retrieval improves recommendation reliability. The multi-agent LLM baseline further increases planning consistency to 87.2%, but its spatial perception ability remains constrained without explicit geographic modeling. GeoCivic-Agent achieves the best overall performance, obtaining the lowest MAE of 0.087, RMSE of 0.143, and the highest AUC of 96.2%. Furthermore, the planning consistency reaches 92.6%, which is 5.4 percentage points higher than the multi-agent baseline. These improvements demonstrate that the proposed framework effectively integrates multimodal information, spatial reasoning, and domain knowledge retrieval for accurate urban hotspot discovery and explainable planning support.

Table 3Performance Comparison on Spatial Hotspot Detection and Planning Decision Support

Model	**MAE ↓**	**RMSE ↓**	**AUC (%) ↑**	**Planning Consistency (%)**
Random Forest Geo Model	0.184	0.263	82.5	71.4
CNN-GIS Fusion Model	0.156	0.224	86.7	76.8
ST-GNN Spatial Model	0.132	0.198	90.4	81.5
LLM-RAG Planning Model	0.121	0.186	92.1	85.7
Multi-Agent LLM Baseline	0.115	0.179	93.0	87.2
GeoCivic-Agent (Ours)	**0.087**	**0.143**	**96.2**	**92.6**

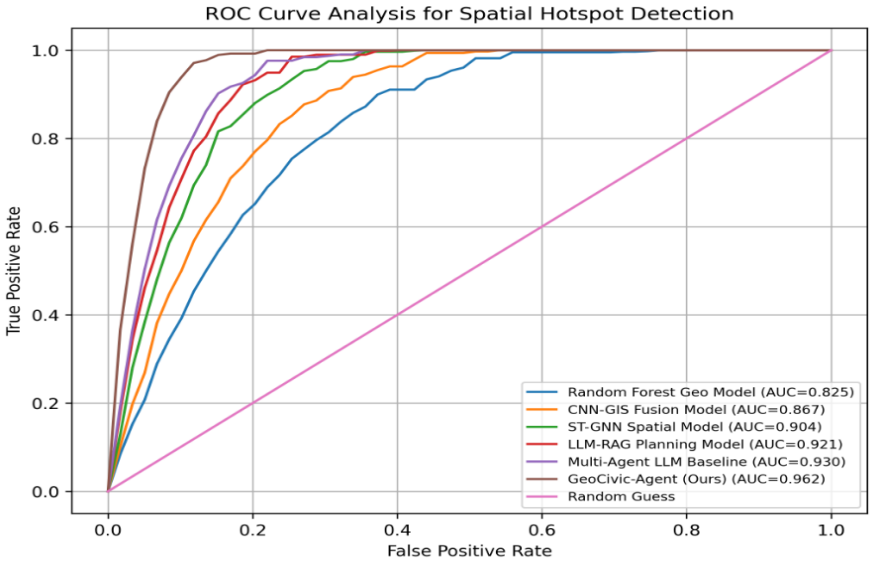


Figure 2ROC Curve Analysis for Spatial Hotspot Detection

Figure 2 presents the Receiver Operating Characteristic (ROC) curves for spatial hotspot detection across the evaluated models. GeoCivic-Agent achieves the highest AUC of 0.962, indicating the strongest ability to distinguish hotspot from non-hotspot locations. Its ROC curve remains consistently closer to the upper-left region of the plot, corresponding to a higher true-positive rate across a range of false-positive rates. The Multi-Agent LLM baseline obtains an AUC of 0.930, followed by the LLM-RAG Planning Model with an AUC of 0.921. The ST-GNN Spatial Model achieves an AUC of 0.904, while the CNN-GIS Fusion Model and Random Forest Geo Model obtain AUC values of 0.867 and 0.825, respectively.

The ordering of the AUC values is consistent with the results reported in Table 3. Models incorporating spatial structure, knowledge retrieval, or language-based reasoning generally achieve higher discrimination performance than the conventional geographic baseline. GeoCivic-Agent further improves upon the

© 2026 The Author(s). Licensed under CC BY 4.0

strongest baseline, increasing AUC from 0.930 to 0.962. This result indicates that the proposed framework provides more consistent discrimination between spatial hotspot and non-hotspot locations across different classification thresholds.

4.5 Discussion

The experimental results demonstrate that GeoCivic-Agent provides significant advantages over existing approaches in multimodal civic opinion analysis and spatial decision support. The improvements mainly originate from the collaborative design of multimodal perception, spatial-aware feature fusion, and autonomous reasoning mechanisms. Unlike traditional machine learning methods that rely heavily on manually designed features, GeoCivic-Agent directly learns semantic, visual, and geographic representations from heterogeneous public participation data. This enables the framework to capture complex relationships between citizen expressions and urban environments. The superior performance compared with single-agent LLM approaches indicates that general-purpose language models alone are insufficient for spatially grounded urban analysis. Although LLMs demonstrate strong language understanding capabilities, they lack explicit mechanisms for geographic reasoning and multimodal alignment. By introducing the GeoFusion Transformer and specialized reasoning agents, GeoCivic-Agent successfully bridges the gap between semantic interpretation and spatial intelligence. The evaluation further highlights the importance of coverage-oriented assessment. Analogous to coverage-aware verification in engineering domains, where adaptive deep reinforcement learning has been shown to improve the completeness of state-space coverage (Zhang and Ding [20]), systematically evaluating the proposed framework across diverse urban issue categories and spatial regions is essential to ensure reliable and comprehensive hotspot detection.

6. Conclusions

This study proposes GeoCivic-Agent (Multimodal Agentic Geographic Artificial Intelligence Framework for Spatial Understanding and Decision Support of Public Participation Data), a novel multimodal GeoAI framework designed to transform heterogeneous public participation data into spatially interpretable urban intelligence for smart city governance. By integrating Geographic Artificial Intelligence (GeoAI), Large Language Models (LLMs), Vision-Language Models (VLMs), and multi-agent reasoning mechanisms, the proposed framework addresses the limitations of existing urban analysis approaches that struggle to simultaneously understand textual opinions, visual evidence, and geographic contexts. Through the designed multimodal perception module, GeoFusion Transformer, multi-agent spatial reasoning architecture, and GIS-enhanced Retrieval-Augmented Generation (RAG) module, GeoCivic-Agent establishes an end-to-end pipeline for civic issue understanding, spatial hotspot detection, and explainable urban planning recommendation generation. Experimental evaluations conducted on the multimodal urban participation dataset demonstrate the effectiveness and superiority of the proposed framework. For urban issue classification, GeoCivic-Agent achieves an F1-score of 94.6%, outperforming the GPT-based single-agent baseline by 2.8 percentage points, which verifies the advantage of combining multimodal feature learning with spatial-aware reasoning. For spatial hotspot detection, the proposed model obtains an AUC of 96.2% and reduces the MAE to 0.087, significantly improving the capability of identifying urban concern patterns compared with conventional GeoAI models and LLM-based approaches. These results indicate that GeoCivic-Agent can effectively transform fragmented citizen feedback from online platforms, social media, and mobile applications into actionable spatial knowledge, providing valuable support for transportation management, environmental monitoring, housing analysis, and participatory urban planning. Despite the performance gains observed in this study, several limitations should be acknowledged. The effectiveness of GeoCivic-Agent may depend on the availability and quality of geographic information, the spatial distribution of citizen participation, the subjectivity of reported observations, and differences in local urban governance practices. In addition, the current evaluation is primarily based on historical civic participation data and therefore does not capture temporal changes in urban conditions as they occur. Future work will

© 2026 The Author(s). Licensed under CC BY 4.0

investigate the integration of real-time data from Internet of Things (IoT) devices and social sensing platforms to provide more timely representations of urban conditions. Further research is also needed to quantify prediction uncertainty and examine potential disparities in model performance across different populations and geographic contexts. Extending the framework to large-scale urban digital twin environments and incorporating more advanced foundation models and adaptive multi-agent learning methods may further improve its scalability and robustness. In particular, adaptive optimization paradigms such as hierarchical hybrid deep reinforcement learning, which have been successfully applied to load balancing and global traffic scheduling in multi-cloud LLM serving (Xu et al. [21]), provide a promising technical basis for endowing GeoCivic-Agent with resource-aware and self-adaptive scheduling capabilities when deployed on cloud platforms for real-time, city-scale analysis. Real-world deployment must also guarantee stable service under high-concurrency loads, particularly during public emergencies when reporting surges. Automated stress testing of AI server infrastructure under high-concurrency scenarios (Ren [22]) provides a practical basis for validating the robustness and responsiveness of GeoCivic-Agent in city-scale deployments, complementing the algorithmic extensions discussed above. These extensions will be important for evaluating the applicability of GeoCivic-Agent to continuously evolving urban environments and real-world planning processes.

© 2026 The Author(s). Licensed under CC BY 4.0

References

Goodchild, Sarah A., et al. "A humanised murine monoclonal antibody with broad serogroup specificity protects mice from challenge with Venezuelan equine encephalitis virus." Antiviral research 90.1 (2011): 1-8.

Janowicz, Krzysztof, et al. "GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond." International Journal of Geographical Information Science 34.4 (2020): 625-636.

Yang, Ruiyi, Yanfei Zhong, and Yu Su. "Self-supervised joint representation learning for urban land-use classification with multisource geographic data." IEEE Transactions on Geoscience and Remote Sensing 63 (2025): 1-21.

Yuan, Long, et al. "Omnigeo: Towards a multimodal large language models for geospatial artificial intelligence." arXiv preprint arXiv:2503.16326 (2025).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

Huang, Weiming, Jing Wang, and Gao Cong. "Zero-shot urban function inference with street view images through prompting a pretrained vision-language model." International Journal of Geographical Information Science 38.7 (2024): 1414-1442.

Xu, Jinwen, and Ran Tao. "Map reading and analysis with GPT-4V (ision)." ISPRS International Journal of Geo-Information 13.4 (2024): 127.

Zhou, Zhilun, et al. "Large language model for participatory urban planning." arXiv preprint arXiv:2402.17161 (2024).

Zhu, He, et al. "PlanGPT: Enhancing urban planning with tailored language model and efficient retrieval." arXiv preprint arXiv:2402.19273 (2024).

Lee, Yesung, et al. "Applications of AI Large Language Models in Citizen Participatory Urban Planning: Using GPT-4o and Prompt Engineering." Journal of Korea Planning Association 60.4 (2025): 148-172.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." Proceedings of the 36th annual acm symposium on user interface software and technology. 2023.

Howell, Anthony, et al. "Multimodal large language models, street view images and urban policy-intelligence: recovering the sustainability effects of redlining." npj Urban Sustainability (2026).

Yao, Yilin, et al. "Deep Learning Models for Multi-class Pneumonia Detection in Chest X-rays: A Comparative Study of VGG16, MobileNet, and ResNet152." Journal of Technology Innovation and Engineering 1.4 (2025).

Qiu, Min, Linting Wu, and Shanshan Wang. "CyberDQFormer: A Knowledge Graph-Enhanced Transformer Framework for Security-Aware Data Quality Anomaly Detection and Autonomous Repair Against Network Attacks." Fundamental Scientific Reports in Multidisciplinary Areas 2.3 (2026): 34-52.

Zhang, Bingli, et al. "Consensus Matrix: A Role-Specialized Multi-Agent Framework for Structured Collaborative Decision-Making in Agentic Visual Media Workflows." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026.

Gao, Ziyang, Yuanliang Qu, and Yi Han. "Cross-Lingual Sponsored Search via Dual-Encoder and Graph Neural Networks for Context-Aware Query Translation in Advertising Platforms." arXiv preprint arXiv:2510.22957 (2025).

Ren, Xingcheng. "Optimizing Large-Scale AI Server Testing via Adaptive Evolutionary Algorithms." International Journal of Computer Science and Engineering 1.4 (2026): 7-14.

Hua, Zonghui, Yue Luo, and Yunting Ling. "LPG-SAC-TCO: A Large Language Model--Policy-Guided Soft Actor-Critic Framework for Total Cost of Ownership--Optimal Energy Management of

© 2026 The Author(s). Licensed under CC BY 4.0

Hydrogen Fuel Cell Heavy-Duty Trucks." European Journal of AI, Computing & Informatics 2.3 (2026): 1-11.

Han, Zihan, Lingran Meng, and Jingwei Zhang. "A distributed hierarchical spatio-temporal edge-enhanced graph neural network for city-scale dynamic logistics routing." arXiv preprint arXiv:2512.18441 (2025).

Zhang, Jingsi, and Jixiang Ding. "CA-PPO-AXI: Coverage-Aware Deep Reinforcement Learning for Adaptive AXI Bus Protocol Verification." Fundamental Scientific Reports in Multidisciplinary Areas 2.02 (2026): 159-169.

Xu, Chengde, Chongye Wang, and Zeyu Li. "H2-LBM: A Hierarchical Hybrid Deep Reinforcement Learning Framework for L7 Load Balancing and Global Traffic Scheduling in Multi-Cloud LLM Serving." Fundamental Scientific Reports in Multidisciplinary Areas 2.02 (2026): 109-120.

Ren, Xingcheng. "Research on Stress Testing Automation of AI Server for High Concurrency Scenarios." Journal of Intelligence and Engineering Technology 1.2 (2026): 7-12.

© 2026 The Author(s). Licensed under CC BY 4.0