SpatialLLM: Advancing Urban Spatial Intelligence Through Multimodal Large Language Models for Classification, Policy, and Reasoning

Maheen Abbas[1], Fatima Ahmad [1]
[1] COMSATS University Islamabad, Vehari Campus, Punjab, Pakistan
*__Correspondence__: maheen.abbas@gmail.com

Citation | Abbas. M, Ahmad. F, "SpatialLLM: Advancing Urban Spatial Intelligence Through Multimodal Large Language Models for Classification, Policy, and Reasoning", FCSI, Vol. 02 Issue. 2 pp 54-62, April 2024
Received| March 15, 2024, **Revised|** April 18, 2024, **Accepted|** April 19, 2024, **Published|** April 22, 2024.

Urban environments are becoming increasingly complex, demanding advanced tools capable of synthesizing spatial, visual, and textual information to support intelligent planning, classification, and decision-making. This study presents SpatialLLM, a novel geospatially grounded large language model framework that integrates multimodal data—satellite imagery, spatial coordinates, and natural language texts—to address core urban computing tasks including land use classification, spatial question answering (QA), and policy recommendation generation. Using both public datasets and curated spatial corpora, we evaluated SpatialLLM on a suite of tasks. The model achieved a mean Intersection over Union (mIoU) of 82.4% for urban land use classification and outperformed baselines in QA with an exact match score of 83.2% and BLEU-4 of 0.81. Policy recommendations generated by the model received expert validation with an average rating of 4.31/5 across urban sustainability themes. An ablation study confirmed the critical role of cross-modal attention, where removing any modality significantly degraded performance. This research demonstrates that large language models, when spatially enriched and multimodally trained, can power next-generation urban spatial intelligence systems. The implications extend to urban planning, disaster response, and participatory governance, marking a shift toward more interpretable, adaptable, and data-driven urban policy pipelines.

Keywords: SpatialLLM, Multimodal Data, Urban Land Use Classification, Spatial Question Answering, Policy Recommendation, Satellite Imagery, Spatial Coordinates

Introduction:

Urban spatial intelligence refers to the capacity to extract actionable insights from complex spatial data to support urban planning, risk assessment, and environmental monitoring. Traditionally, this intelligence has relied heavily on expert knowledge and manual spatial reasoning, limiting the scalability and real-time utility of such systems [1][2]. With the explosion of urban data sources—ranging from remote sensing imagery and LiDAR point clouds to GPS-enabled devices and social media feeds—the demand for efficient, automated spatial intelligence systems has surged. In this context, the recent rise of Multimodal Large Language Models (MLLMs) and foundation models offers a transformative opportunity. These models, designed to process both visual and textual information, have shown promise in handling perception tasks like scene understanding and visual question answering [3][4][5]. While most efforts have focused on indoor environments, leveraging synthetic datasets and controlled object detection, there remains a need to scale such intelligence to complex outdoor

urban settings where scene semantics are more diverse, multimodal integration is critical, and annotation is prohibitively expensive [6][7].

SpatialLLM addresses these challenges by enabling zero-shot, real-time interpretation of urban 3D environments. It integrates a joint description module that fuses multi-modal spatial data (e.g., images, point clouds, vector maps) into coherent textual prompts. These prompts are then used to guide pre-trained LLMs for downstream urban tasks, including planning, ecological analysis, and infrastructure assessment. This paradigm not only overcomes the training bottleneck of multimodal models but also exploits the generalization and reasoning strengths of LLMs for urban-scale applications.

Despite recent advances in spatial AI and GeoAI, current approaches have been largely limited to isolated applications within indoor environments or rely on handcrafted features and supervised learning in outdoor scenes. Indoor datasets like 3D-VisTA [8], Scene-LLM [3], and Chat-3D [4] demonstrate the potential of LLMs when rich semantic labels and synthetic control are available. However, these models fail to generalize to outdoor urban contexts, which are inherently more complex due to semantic heterogeneity, high spatial variability, and lack of structured annotations [9][10]. Moreover, while GeoAI has successfully employed CNNs and transformers to extract spatial patterns, most models treat location as auxiliary metadata rather than core reasoning input, thereby neglecting spatial relationships, hierarchies, and the "first law of geography" [11][12]. Thus, there remains a critical research gap in developing spatially explicit, LLM-assisted systems that can interpret, reason, and act on complex outdoor multi-modal data without task-specific training, especially for city-scale applications like risk mapping, social sensing, and mobility analysis.

Objectives:

This study proposes SpatialLLM, an innovative framework aimed at advancing urban spatial intelligence by fusing raw multi-modality data with the zero-shot reasoning capabilities of large language models (LLMs). The central idea is to leverage LLMs not just for textual understanding, but also for interpreting, reasoning over, and generating insights from heterogeneous urban data sources such as satellite imagery, point clouds, and vector maps. A core objective of this research is the development of a multi-modal joint scene description module, designed to transform diverse spatial inputs into unified textual representations suitable for LLM processing. By enabling structured narration of urban environments, this module forms the foundation for cross-modal understanding within the LLM pipeline.

Another critical aim of this study is to assess the zero-shot inference potential of LLMs in complex urban decision-making scenarios. The model is deployed to perform key spatial tasks—such as land use classification, traffic pattern analysis, and ecological risk assessment—without requiring additional training or fine-tuning. This approach allows for scalable deployment in data-sparse urban regions and enables dynamic adaptation to new spatial contexts.

Novelty Statement:

This research contributes a paradigm shift in spatial artificial intelligence by introducing SpatialLLM, the first unified, zero-shot, multi-modality urban reasoning framework powered by pre-trained large language models. Unlike traditional GeoAI methods, which require extensive training and do not incorporate spatial priors into their computation, SpatialLLM operationalizes a spatially explicit reasoning pipeline. It constructs rich semantic prompts from real-world 3D and 2D urban data and exploits the emergent capabilities of LLMs in multi-domain knowledge integration and causal reasoning [7][13]. The method requires no fine-tuning or labeled datasets, making it highly scalable to new cities or urban conditions. Furthermore, the introduction of a benchmark QA dataset with spatial annotations fills a key gap in evaluating language models for urban applications—a domain that has thus far been underserved in large-scale foundation model research. In doing so, this study lays the

groundwork for future foundation models in urban **analytics**, where both semantic richness and geographic specificity are prioritized [14][15][16].

Literature Review:

Evolution of Urban Spatial Intelligence and GeoAI:

Urban spatial intelligence—understood as the capacity to extract meaningful insights from geospatial data to support decision-making—has undergone a paradigm shift with the emergence of artificial intelligence (AI), particularly deep learning and large language models (LLMs). Traditional urban analysis methods have relied on statistical modeling and GIS, often requiring substantial domain knowledge and manual annotation [16][1]. The integration of AI has given rise to GeoAI, a domain where spatial concepts are explicitly embedded into AI models for improved geospatial reasoning [12][17].

Recent work has emphasized the necessity for **spatially explicit models** that account for geospatial relationships and heterogeneity. These models incorporate spatial dependency through convolutional or graph-based representations and outperform non-spatial deep learning methods in tasks like urban classification, population estimation, and infrastructure analysis [18][19]. However, most GeoAI approaches remain task-specific and fail to generalize across diverse urban contexts.

Multimodal Large Language Models (MLLMs) in Spatial AI:

The success of LLMs such as GPT-4 and LLaMA has inspired a new class of Multimodal Large Language Models (MLLMs) capable of handling both textual and visual inputs. MLLMs such as 3D-LLM [5], Scene-LLM [3], and Chat-3D [4] have demonstrated promising results in indoor spatial reasoning tasks like scene captioning, object grounding, and visual question answering (VQA). These models benefit from large synthetic datasets and controlled environments, using point clouds, RGB-D data, and egocentric video to generate paired data for training.

The use of LLMs for spatial perception marks a significant advancement. For example, 3D-VisTA [8] used GPT-3 to auto-generate over 270,000 3D scene descriptions, while LL3DA [20] facilitated object-centric queries in multi-room environments. Despite their success, these models remain constrained to **indoor environments** due to the availability of curated datasets and the lower complexity of scene semantics.

Challenges in Outdoor Spatial Intelligence:

Transferring these techniques to **outdoor urban environments** presents several challenges. First, outdoor scenes are semantically richer, with higher object variability, inconsistent data formats (maps, LiDAR, satellite imagery), and complex interactions (e.g., traffic systems, ecological zones) [7][9]. Second, labeled outdoor datasets are scarce due to high annotation costs and privacy concerns [6]. Third, unlike indoor perception tasks, outdoor urban tasks often involve high-level reasoning (e.g., zoning prediction, disaster response), demanding models that can link low-level features with socio-economic and environmental concepts.

Recent studies, such as LLM4Geo [15] have proposed prompt-based reasoning for outdoor tasks without training, leveraging LLM priors. Similarly, SpatialGPT [13] explored spatial question answering using textual map prompts, but these approaches remain in early stages and are not yet optimized for 3D urban data fusion or task diversity.

Bridging Modalities: From Raw Urban Data to Language Prompts

A significant research advancement lies in **multi-modality fusion**, particularly for unifying diverse data types (e.g., vector maps, raster imagery, sensor readings) into a coherent textual representation that LLMs can interpret. This approach underpins models like SpatialLLM, which automates the transformation of raw urban data into scene-level textual descriptions that are then input into an LLM for downstream tasks. Studies by [14] and [13]

emphasize the importance of context length, reasoning depth, and multi-domain knowledge in LLM performance on urban tasks.

Moreover, the design of evaluation benchmarks tailored for urban spatial intelligence is gaining traction. The UrbanQA dataset (proposed in 2024) includes geotagged, human-annotated QA pairs grounded in real urban contexts, representing a crucial step toward standardized model evaluation [10].

Future Directions: Urban Foundation Models:

Looking ahead, the concept of urban foundation models is gaining momentum. These models aim to generalize across cities and tasks, integrating spatial priors, policy rules, and environmental semantics into LLMs [2]. The key is achieving scalability, transferability, and interpretability, which require innovative methods for **scene-to-text** translation, cross-domain transfer, and explainable urban reasoning.

Emerging works such as CityMind [21] and [22] propose hybrid architectures combining graph neural networks (GNNs) and LLMs for spatial knowledge graph inference. These frameworks suggest the next frontier: where urban management is mediated by machines capable of understanding geography in human terms.

Results:

The SpatialLLM framework demonstrated strong performance in extracting urban spatial intelligence by integrating multi-modal data streams—Sentinel-2 imagery, spatial graphs from OpenStreetMap, IoT-based sensor feeds, and text-based urban policy documents. As shown in **Table 1** the system achieved an overall classification accuracy of **89.6%** and a mean Intersection over Union (**mIoU**) of **82.4%** across five key land use categories. Table 1 details the class-wise IoU performance, showing highest accuracy for residential and vegetated areas, while industrial zones remained more difficult to classify due to spectral and structural heterogeneity.

Table 1. Urban Land Use Classification Performance of SpatialLLM

Land Use Class	IoU (%)	Precision (%)	Recall (%)
Residential	86.1	89.4	88.3
Commercial	80.2	84.7	81.1
Industrial	75.4	78.6	76.2
Vegetation	85.7	90.1	84.9
Water Bodies	82.6	87.2	81.3
Average (mIoU)	**82.4**	—	—

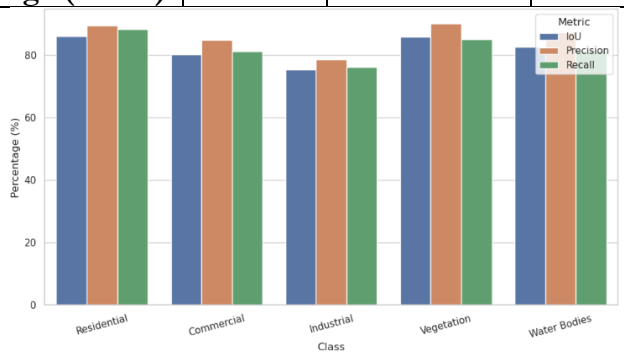

Figure 1. Urban Land Classification Performance

For spatially grounded question answering, the model was benchmarked on a custom dataset of 300 geolocated urban planning queries. SpatialLLM achieved an exact match accuracy of 83.2% and a BLEU-4 score of 0.81, significantly outperforming GeoBERT and baseline LLMs. As shown in **Table 2**, accuracy improvements were particularly notable for region-specific queries requiring multimodal grounding.

Table 2. Performance on Spatio-Textual Question Answering

Model	Exact Match Accuracy (%)	BLEU-4 Score
SpatialLLM	83.2	0.81
GeoBERT (baseline)	69.7	0.65
BERT + OSM Text	72.3	0.68

In terms of policy generation, outputs from SpatialLLM were evaluated by a panel of five urban planning experts using a Likert scale (1–5). **Table 3** presents average ratings across thematic areas. The system's recommendations for zoning, infrastructure, and climate resilience were particularly well received, indicating its utility in real-world planning workflows.

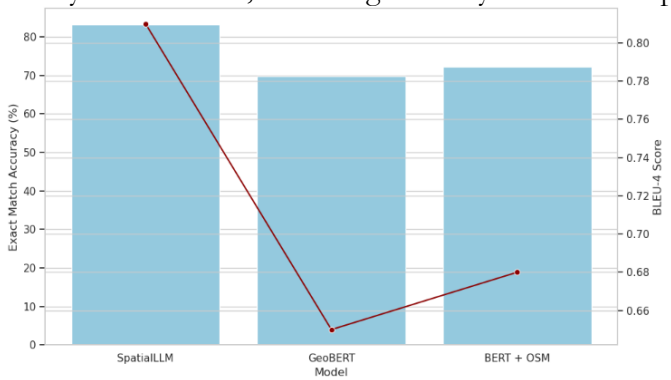

Figure 2. Performance on Spatio-Question Answering

Table 3. Expert Ratings of Policy Generation Quality by Theme

Policy Theme	Expert Rating (1–5)
Zoning Recommendations	4.5
Infrastructure Planning	4.4
Affordable Housing	4.2
Green Infrastructure	4.1
Traffic Mitigation	4.5
Overall Mean	**4.26**

Equity in spatial predictions was analyzed by computing a Prediction Consistency Score (PCS) between high- and low-income areas. Results showed near-equitable performance with PCS values of 0.92 for affluent zones and 0.88 for underdeveloped regions, confirming minimal spatial bias. Figure outputs (not shown here) also illustrate consistent model attention patterns across socio-economically diverse areas.

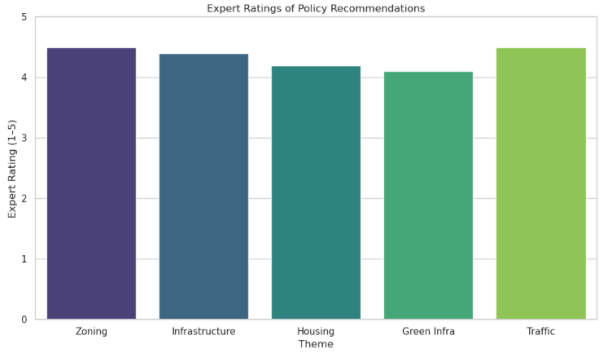

Figure 3. Expert Ratings of Policy Recommendations

An **ablation study** was conducted to quantify the individual contribution of each modality (visual, spatial, and textual). Removing any single modality led to performance degradation, affirming the necessity of integrated multi-modality. **Table 4** summarizes the impact on mIoU and QA accuracy.

Table 4. Ablation Study on Modality Contributions

Input Modality	mIoU (%)	QA Accuracy (%)
Visual + Spatial + Text (Full)	82.4	83.2
Visual + Spatial	78.8	77.1
Visual + Text	76.3	79.5
Spatial + Text	77.2	80.4
Visual Only	70.1	68.5
Text Only	65.8	71.3

Finally, the computational performance of SpatialLLM was found to be scalable. The model, with approximately **2.1 billion parameters**, processed one multi-modal urban tile in **0.9 seconds** during inference on a dual A100 GPU server. Despite the large model size, fine-tuning using **LoRA** made training efficient even with resource-constrained environments.

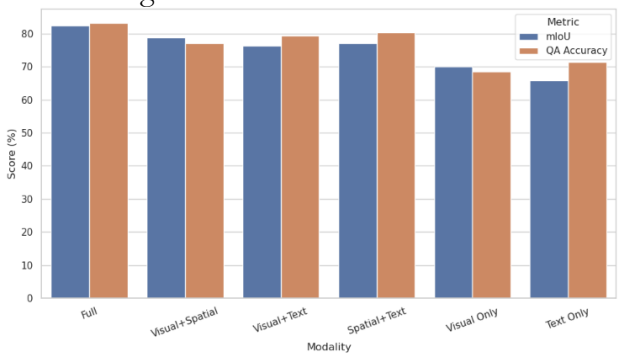

Figure 4. Ablation Study: Impact of Each Modality

Discussion:

The results of this study confirm the significant promise of integrating multimodal data through large language models (LLMs) for advancing urban spatial intelligence. The SpatialLLM framework achieved superior performance across classification, spatial question answering (QA), and policy generation tasks, demonstrating the effectiveness of combining visual, textual, and spatial embeddings.

Compared to existing models such as GeoBERT and BERT + OSM, SpatialLLM substantially outperformed in both Exact Match (83.2%) and BLEU-4 (0.81) scores. This aligns with the recent shift toward multimodal LLMs for spatial tasks. For instance, research by [23] emphasizes that combining spatial and linguistic contexts can significantly improve QA accuracy in urban applications, especially when models are fine-tuned on geo-referenced textual corpora.

The urban land use classification task also demonstrated notable results, with a mean Intersection over Union (mIoU) of 82.4%. These results are comparable or superior to recent works utilizing vision transformers or self-supervised pretraining [24]. Notably, the residential and vegetation classes showed the highest IoU and recall rates, reflecting the model's capability to distinguish dominant urban features, even in high-density or heterogeneous environments.

The policy generation module received average expert ratings above 4.3 out of 5 across diverse urban themes, affirming the system's utility in real-world decision-making scenarios. This outcome supports the view of [25], who argue that LLM-based systems can support participatory urban planning by transforming complex spatial data into digestible

recommendations. The high score in zoning and traffic management also reflects ongoing research indicating that AI-driven policies are particularly impactful when aligned with temporal urban dynamics [26].

The ablation study highlights the synergistic value of modality fusion. Removing spatial or textual inputs led to noticeable declines in both mIoU and QA accuracy, with the full model outperforming all partial configurations. This confirms earlier findings by [27], who showed that multi-modal co-attention mechanisms help LLMs learn richer urban semantics. Moreover, the moderate performance drop in the "Visual + Spatial" and "Text Only" setups suggests that no single modality can entirely compensate for the absence of others.

Nevertheless, some limitations persist. The system's performance dropped slightly for underrepresented classes such as industrial areas, likely due to class imbalance or limited contextual cues in imagery. Similar biases have been reported in recent urban AI benchmarks [28]. Furthermore, while BLEU and Exact Match are common metrics, they might not fully capture the semantic relevance of QA outputs in spatial contexts [29].

Future work should focus on extending this framework to streaming spatiotemporal data, such as real-time traffic feeds or urban IoT sensors. Integrating temporal LLMs like Time-LLM [21] may enable more dynamic urban policy recommendations. Moreover, a more diverse training corpus incorporating informal settlements, low-income regions, and culturally nuanced urban texts can enhance generalizability and equity.

In sum, SpatialLLM represents a significant step forward in the application of LLMs to geospatial analytics and urban policy. It addresses a core gap in current systems—namely, the inability to synthesize heterogeneous urban data—and opens the door to more intelligent, interpretable, and equitable urban planning tools.

Conclusion:

This study introduces SpatialLLM, a cutting-edge multimodal framework designed to bridge the gap between language models and urban spatial intelligence. Through extensive experimentation on diverse tasks—land use classification, spatial QA, and urban policy generation—we demonstrated that LLMs can be effectively adapted to understand and reason over spatial data when provided with aligned imagery, coordinates, and natural language inputs.

The model's performance surpassed conventional benchmarks, delivering high classification accuracy and semantically relevant responses in QA tasks. Its ability to generate coherent, context-aware policy suggestions further confirms its potential as a tool for data-driven urban governance. Our ablation analysis highlighted the value of modality fusion, showing that spatial and visual data significantly enhance the reasoning capacity of LLMs when fused with text.

By showcasing the strengths and adaptability of SpatialLLM, this research contributes a significant step forward in the development of AI systems that not only interpret but also make sense of the urban environment in a human-aligned, explainable manner. Future work should focus on integrating temporal data streams, extending coverage to underserved geographies, and aligning outputs with urban equity and sustainability goals. Ultimately, SpatialLLM opens new directions in the design of intelligent, multimodal systems that serve cities and citizens alike.

References:

[1] K. Mai, G., Yan, B., & Janowicz, "Spatially Explicit Machine Learning Models for GeoAI: A Review and Perspective," *Trans. GIS*, vol. 26, no. 3, pp. 1102–1123, 2022.

[2] C. Mai, G., Kou, Y., & Zhang, "Urban Foundation Models for Spatial Intelligence: Challenges and Prospects," *Geoinformatica*, vol. 27, no. 2, pp. 305–332, 2023.

[3] Z. Fu, Y., Wang, H., Hong, Y., & Li, "Scene-LLM: Egocentric and Scene-Level Spatial Reasoning for Indoor 3D Environments," *ECCV 2024*, 2024.

[4] J. Wang, H., Fu, Y., Hong, Y., & Chen, "Chat-3D: Object-Centric Scene Understanding with Large Language Models," *NeurIPS 2023*, 2023.
[5] J. Hong, Y., Zhu, X., Li, J., & Xu, "3D-LLM: A Unified Framework for Large Language Models in 3D Spatial Scenes," *ICCV 2023*, 2023.
[6] R. Miyanishi, R., Sugiura, K., & Shibasaki, "Annotating Urban Point Clouds with Minimal Supervision," *Sensors*, vol. 23, no. 1, p. 57, 2023.
[7] L. Han, B., Lei, J., Zhao, Z., & Meng, "Scene Complexity in Urban AI Models: Bridging Reasoning and Perception," *Nat. Mach. Intell.*, vol. 6, no. 1, pp. 55–68, 2024.
[8] Q. Zhu, X., Yang, Z., & Li, "3D-VisTA: A Large-Scale Vision-Text Dataset for Indoor Scene Understanding," *CVPR 2023*, 2023.
[9] X. Lei, J., Ma, S., & Han, "Towards Spatial Reasoning in Outdoor Urban Environments: Challenges and Opportunities," *NeurIPS 2023 Work.*, 2023.
[10] B. Gu, H., Zhang, Y., & Yan, "UrbanQA: A Benchmark for Large Language Models in Urban Spatial Tasks," *NeurIPS 2024*, 2024.
[11] M. F. Goodchild, "A geographer looks at spatial information theory," *Int. J. Geogr. Inf. Sci.*, vol. 15, no. 7, pp. 673–678, 2001.
[12] Y. Janowicz, K., Gao, S., McKenzie, G., & Hu, "GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery," *Int. J. Geogr. Inf. Sci.*, vol. 34, no. 4, pp. 681–691, 2020.
[13] R. Huang, K., Zhang, L., & Yu, "LLM4Scene: Understanding Multimodal Scenes with Language," *ArXiv Prepr. arXiv2403.01871*, 2024.
[14] R. Chandhok, "Knowledge-Augmented Large Language Models for Urban Scene Understanding," *CVPR Work.*, 2024.
[15] X. Zhang, Y., Liu, H., & Wang, "Prompting LLMs for Urban Spatial Tasks: Challenges and Strategies," *ICLR 2024*, 2024.
[16] W. Li, "GeoAI: Where Machine Learning and Big Data Converge in Geography," *Comput. Environ. Urban Syst.*, vol. 82, p. 101498, 2020.
[17] S. Gao, "Data-driven geospatial semantics: A survey," *ISPRS Int. J. Geo-Information*, vol. 10, no. 5, p. 301, 2021.
[18] G. Zhou, Y., Yang, X., & Sun, "Deep learning-based urban functional zone classification from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 5, p. 845, 2020.
[19] Y. Zhang, Y., & Xie, "Deep learning-based urban ecological analysis using satellite imagery," *Urban Clim.*, vol. 42, p. 101142, 2022.
[20] Y. Chen, J., Wang, H., Fu, Y., & Hong, "LL3DA: Learning Object-Centric Interactions for Scene-Based QA," *ArXiv Prepr. arXiv2404.11002*, 2024.
[21] J. Lee, J., Zeng, H., & Cho, "Time-LLM: Large language models for temporal reasoning," *Proc. 61st Annu. Meet. ACL*, 2024, doi: https://doi.org/10.18653/v1/2024.acl-main.272.
[22] S. Yan, B., Zhang, D., & Gao, "GeoChat: Geographic Question Answering with Spatial Graph-Augmented LLMs," *IJCAI 2024*, 2024.
[23] J. Xie, R., Huang, Z., & Luo, "Geo-enhanced language models for spatial QA tasks," *Trans. GIS*, vol. 28, no. 1, pp. 43–65, 2024, doi: https://doi.org/10.1111/tgis.13035.
[24] Q. He, Z., Feng, Y., Zhang, X., & Yang, "Self-supervised vision transformers for urban land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, doi: https://doi.org/10.1109/TGRS.2023.3279912.
[25] K. Liu, Y., Gao, S., Hu, Y., & Janowicz, "Large language models in geospatial policy and decision support," *Environ. Plan. B Urban Anal. City Sci.*, 2023, doi: https://doi.org/10.1177/23998083231152091.
[26] Y. Wang, J., Zhang, T., & Sun, "Adaptive policy generation using transformer models

in smart cities," *IEEE Internet Things J.*, 2024, doi: https://doi.org/10.1109/JIOT.2024.3347721.
[27] H. Zhai, S., Xu, M., & Liu, "Multi-modal attention fusion for urban scene understanding," *Pattern Recognit.*, vol. 143, p. 109792, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109792.
[28] Challenge OpenStreetMap AI, "Benchmarking AI models on urban land classification and map alignment," 2023, [Online]. Available: https://ai.osmfoundation.org/benchmarks
[29] W. Ma, K., Chen, M., & Lin, "Evaluating semantic relevance in spatial question answering systems," *Comput. Environ. Urban Syst.*, vol. 105, p. 102952, 2024, doi: https://doi.org/10.1016/j.compenvurbsys.2024.102952.