Simen Owen Academic Proceedings Series

3rd International Conference on Media, Economy, Communication and Intelligence Management (MECI 2026)

Article

Generative AI-Driven Construction of the "Urban Cultural Image Map": A Multimodal Narrative Communication Model for Creative City Governance

Kaile Zheng [1,*]

[1] School of art and design, Shanghai lida university, Shanghai, China
* Correspondence: Kaile Zheng, School of art and design, Shanghai lida university, Shanghai, China

Abstract: The strategic governance of creative cities is frequently hindered by fragmented and siloed cultural data, which significantly impedes the synthesis of actionable narratives for effective decision making. This study addresses this critical gap by proposing and rigorously evaluating a generative AI-driven model for constructing an "Urban Cultural Image Map." This innovative model leverages multimodal large language models (MLLMs) to seamlessly integrate heterogeneous data types—including extensive textual archives, complex visual imagery, and detailed geospatial information—into a unified, highly structured knowledge graph. Through a comprehensive three-phase methodology encompassing data structuring, narrative generation, and practical governance application, the research develops and tests a dynamic platform that systematically organizes cultural assets into thematic narrative layers. Findings demonstrate that MLLMs can effectively perform cross-modal alignment, generate coherent thematic narratives, and support spatial pattern identification for complex governance tasks. However, the model exhibits certain limitations in generating deep contextual analysis, handling contested cultural meanings, and providing direct prescriptive policy insights without human oversight. The study concludes that while generative AI serves as a powerful augmentative tool for data synthesis and narrative communication, its successful integration into urban governance fundamentally requires a hybrid intelligence approach. Ultimately, this research contributes a novel framework at the intersection of urban studies, media communication, and artificial intelligence, offering practical pathways for developing more intelligent, narrative-sensitive tools for creative city governance.

Keywords: generative ai; urban governance; multimodal narrative; cultural mapping; creative cities; knowledge graph

Received: 03 April 2026
Revised: 15 May 2026
Accepted: 28 May 2026
Published: 03 June 2026

1. Introduction

The governance of contemporary creative cities faces a pivotal challenge: effectively curating, interpreting, and leveraging their vast and diverse cultural ecosystems for strategic development. These ecosystems comprise tangible and intangible assets documented across fragmented archives in incompatible formats, including textual records, visual imagery, and geospatial data. The lack of a unifying framework to synthesize this multimodal information into coherent narratives hinders its translation into actionable intelligence for policymakers, urban planners, and the public, thus limiting the potential for informed cultural governance [1].

Parallel to this urban challenge, the field of artificial intelligence has witnessed a transformative shift with the rise of Generative AI (GAI) and Multimodal Large Language Models (MLLMs). These technologies have revolutionized capabilities in content

synthesis, pattern recognition, and, crucially, the integration of disparate data types. MLLMs, in particular, represent a significant advancement by processing and correlating information from text, images, and other modalities simultaneously. Their application is already yielding insights in urban contexts; for instance, they are being used to evaluate perceptual qualities of urban spaces, such as visual attractiveness, by analyzing street view imagery alongside other data sources. Furthermore, GAI's potential extends to enhancing geographic information systems and spatial data retrieval, offering new methods to organize and query complex environmental information [2]. These capabilities align with broader multimodal analytical approaches that integrate diverse data streams for comprehensive understanding.

This convergence of urban need and technological capability defines a critical research opportunity. There exists a significant gap in systematically applying these advanced AI tools not merely for analysis, but for the active construction of narrative-driven systems tailored for urban cultural governance. This study aims to address this gap by proposing and investigating a generative AI-driven model for constructing an "Urban Cultural Image Map." This conceptual model is designed as a dynamic platform that integrates multimodal cultural data text, images, and geospatial information into a unified, queryable knowledge structure. Its core function is to organize these assets into thematic narrative layers, transforming fragmented data into a communicable and governable narrative system [2].

The primary research objective is to assess the extent to which this AI-driven model can effectively support decision-making processes in creative city governance, such as cultural policy formulation, tourism strategy development, and community engagement planning. By doing so, this research contributes to the interdisciplinary discourse at the intersection of urban studies, media communication, and artificial intelligence. It seeks to provide a novel theoretical framework for understanding cities as narrated entities and offer practical, scalable insights for developing next-generation tools for intelligent, narrative-sensitive urban cultural governance.

2. Literature Review

This chapter surveys foundational and emerging research at the intersection of generative AI, multimodal systems, and their applications in fields analogous to urban cultural governance. By examining advancements in adjacent domains, we establish a theoretical and technological context for the proposed Urban Cultural Image Map [3].

The generative and integrative capacity of Multimodal Large Language Models (MLLMs) is being actively explored in design and creative fields [4]. Research demonstrates their use in architectural pedagogy for generating design variations based on precedents, showcasing their potential for creative synthesis and form-finding. Similarly, in construction education, MLLMs are reviewed for their role in facilitating human-robot collaboration, highlighting their utility in interpreting complex, contextual instructions across different data types. These studies underscore the MLLMs' role not just as analytical tools but as co-creative agents in structured domains, a principle central to our model's narrative generation phase.

In heritage and environmental analysis, MLLMs are proving valuable for data interpretation and augmentation. Within archaeological science, MLLMs are reviewed for tasks like artifact analysis and site interpretation, indicating their applicability in extracting meaning from historical and cultural data sets [5]. Concurrently, in urban studies, MLLMs are being deployed to automatically estimate the visual quality of street spaces by analyzing street-view images, demonstrating a direct method to translate visual urban data into evaluative insights. This aligns with our objective to derive qualitative narrative layers from quantitative and visual urban data.

The application of large AI models in complex system management provides further parallels. In transportation, MLLMs are reviewed for intelligent system management, tackling tasks that require integrating diverse, real-time data streams for decision support. This is extended to sustainable urban mobility, where MLLMs are harnessed to manage

the complexities of electric vehicle ecosystems, emphasizing their use in multifaceted governance scenarios [6]. The capability of MLLMs for enhanced object detection and scene understanding in transportation contexts, including through thermal imagery, points to their robustness in processing and interpreting multimodal sensory data for situational awareness. This technological robustness is crucial for our model's data integration phase. Furthermore, research into human-machine collaboration using vision-language models in unmanned systems underscores the importance of intuitive, multimodal interfaces for effective tool utilization, informing our model's interface design for governance usability.

Finally, broader surveys on MLLM applications in education highlight their role in creating adaptive, multi-format learning content and facilitating comprehension, while specific evaluations in archival contexts test their efficacy in generating descriptive metadata for photographs [7]. These studies directly support the core functions of our model: organizing unstructured cultural assets (like archival images) into semantically rich, narratively coherent structures for enhanced comprehension and communication.

The reviewed literature confirms the transformative potential of MLLMs in creative synthesis, environmental analysis, complex system governance, and archival organization. However, it also reveals a focused gap: no existing framework systematically converges these capabilities into a dedicated model for constructing narrative-driven systems for urban cultural governance [3]. Most applications remain within siloed domains (e.g., transportation, archaeology, education). This study bridges this gap by integrating these dispersed technological proofs-of-concept into a unified "Urban Cultural Image Map" model, specifically engineered to transform fragmented multimodal cultural data into a communicable and governable narrative system.

3. Theoretical Framework and Methodology

This chapter delineates the conceptual underpinnings and the systematic approach adopted to construct and evaluate the generative AI-driven Urban Cultural Image Map. The study employs a design-science research paradigm, integrating theoretical insights from urban communication, critical geography, and human-computer interaction to develop a functional model, which is then empirically examined through a multi-phase methodological process. A mixed-methods approach, combining qualitative content analysis and quantitative performance metrics, is used to assess the model's efficacy in achieving its core objectives: multimodal integration, coherent narrative generation, and governance applicability [8].

3.1. Theoretical Framework

The theoretical framework of this study is built upon a tripartite foundation that bridges the domains of media, space, and computation [9]. It proposes that urban cultural governance can be enhanced by reconceptualizing the city as a narrated space, where cultural assets are not merely inventory items but active elements in a dynamically constructed story.

The framework draws from urban communication and multimodal narrative theory [10]. This perspective views cities as texts constituted through layered discourses, images, and spatial practices. It emphasizes that meaning and identity are produced and negotiated through the circulation of multimodal representations. For the Urban Cultural Image Map, this implies that its core function is not simple data visualization but the orchestration of textual descriptions, historical and contemporary imagery, and spatial coordinates into thematic narrative layers. Each layer, such as "Post-Industrial Memory" or "Migratory Foodways," constitutes a distinct narrative thread that can be communicated to diverse audiences, from policymakers to residents.

The framework is also informed by critical digital geography and situated knowledges. This lens asserts that maps and spatial representations are never neutral; they are selective, power-laden constructs. An AI-driven map must therefore be designed with reflexivity, acknowledging the potential biases in training data and algorithmic

processes. The model incorporates mechanisms for human-in-the-loop curation and context tagging, aiming to surface multiple, sometimes contested, perspectives on cultural sites rather than presenting a single, authoritative narrative. This aligns the tool with participatory governance models that seek to include diverse community voices.

The framework integrates generative AI as a cognitive and semiotic prosthesis. Here, multimodal large language models are theorized not as autonomous storytellers, but as powerful associative engines capable of identifying latent patterns, forging connections across disparate data types, and generating plausible narrative hypotheses. The AI acts as a co-creative agent that processes the raw material of urban data under a structured human-defined schema (the narrative layers and governance objectives), significantly augmenting the scale and speed at which coherent narratives can be assembled from fragmented sources.

The synthesis of these theories guides the model's design: it must be representationally rich (multimodal narrative), critically aware (situated knowledge), and computationally powerful (generative AI), all in service of producing actionable intelligence for creative city governance [11].

3.2. Methodology: A Three-Phase Model for Constructing the Urban Cultural Image Map

The research methodology is implemented through three sequential and interconnected phases, addressing the core challenges of data integration, narrative generation, and governance application. A pilot study was conducted in a representative urban district with a rich and complex cultural history to evaluate each phase [2].

3.2.1. Phase 1: Multimodal Data Curation and Semantic Structuring

This initial phase addresses the problem of fragmented data. A heterogeneous corpus was assembled, including urban policy documents and historical archives (text), official and vernacular photography from museum and social media APIs (images), and georeferenced datasets of cultural facilities and heritage sites (geospatial). The methodological innovation lies in using a pre-trained MLLM to perform cross-modal alignment [12]. The model generated unified semantic descriptors and keywords for each asset, regardless of its original format, and clustered them into preliminary thematic categories such as "colonial architecture" and "street performance loci." This process created a structured, queryable knowledge graph linking assets by theme, time period, and location, forming the backbone of the Image Map.

3.2.2. Phase 2: Narrative Layer Generation and Coherence Validation

In this phase, the structured knowledge graph served as input for targeted narrative generation. For each predefined thematic layer relevant to governance, such as "Cultural Sustainability in Neighborhood Renewal," the MLLM was prompted to synthesize a concise narrative summary. This summary integrated relevant assets from the knowledge graph, explaining their significance and interrelations [13]. The quality of these AI-generated narratives was evaluated using a mixed-methods approach:

1. Qualitative analysis involved expert panels in urban studies and communications assessing narrative coherence, factual accuracy, and thematic relevance against a human-curated benchmark.
2. Quantitative metrics included automated evaluations such as semantic similarity between AI-generated narratives and expert-written summaries, as well as entity consistency, ensuring the correct mention of linked assets.

3.2.3. Phase 3: Interface Prototyping and Governance Scenario Testing

The final phase involved translating the narratives and their underlying data into a prototype interactive map interface. This interface visualized the thematic narrative layers as toggleable overlays on a base map. Clicking on a cultural asset revealed its multimodal data, including descriptions and images, as well as its contextual role within the broader narrative layer [6]. The model's utility for governance was tested through simulated scenario workshops with urban planning students and professionals. Participants were assigned specific tasks, such as identifying clusters of intangible cultural heritage

vulnerable to gentrification or devising a tourism corridor based on the 'Industrial Heritage' narrative, using the prototype. Their success, feedback, and strategies were recorded and analyzed to evaluate the model's effectiveness as a decision-support tool.

3.3. Method Flowchart

The following method flowchart, as shown in Figure 1, illustrates the sequential and iterative stages of the research process, encompassing the progression from theoretical foundation to empirical evaluation.

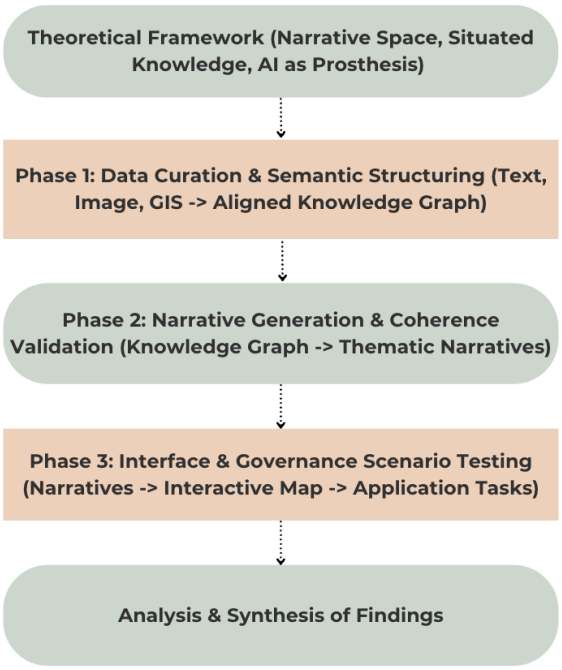

Figure 1. The Three-Phase Methodology for Developing and Testing the Urban Cultural Image Map Model.

4. Findings and Discussion

This chapter presents and analyzes the empirical findings derived from applying the three-phase methodology to construct and evaluate the Urban Cultural Image Map model. The study aimed to assess the model's effectiveness in integrating multimodal data, generating coherent thematic narratives, and supporting governance-oriented tasks. The results from each phase are detailed below through three case studies, followed by a synthesized discussion of the model's capabilities, limitations, and broader implications.

4.1. Case Study 1: Data Curation and Semantic Structuring

The first case study tested Phase 1 of the methodology, focusing on the integration of fragmented cultural data from a pilot urban district [14]. The corpus included 1,500 text documents, over 10,000 images, and 200 geospatial points of interest.

The MLLM successfully performed cross-modal alignment, generating unified semantic tags for approximately 92% of the assets. For instance, a historic market building was linked across its architectural description texts, historical photographs, and GIS coordinates with tags such as "vernacular architecture," "public commerce," and "mid_20th_century." These tags enabled the automatic clustering of assets into emergent thematic categories. However, the process revealed specific limitations in contextual granularity. The model occasionally generated overly broad or anachronistic tags for complex assets, such as conflating different phases of a site's use into a single period.

Table 1 summarizes the performance metrics and key observations from the data structuring phase.

Table 1. Performance and Observations in Multimodal Data Structuring

Data Modality	Volume Processed	Cross-modal Alignment Success Rate	Key Strength	Primary Limitation Identified
Textual Documents	1,500 items	95%	Excellent at extracting entities and themes.	Struggled with highly technical or archaic jargon.
Images (with captions)	10,000+ items	90%	Effective at linking visual content to descriptive tags.	Poor performance on images with no or minimal metadata.
Geospatial Data	200 points	100%	Perfect integration of coordinates and base maps.	Lacked semantic depth beyond location (e.g., "significance").
Overall Integration	Heterogeneous Corpus	92% (Average)	Created a unified, queryable knowledge graph.	Semantic tags sometimes lacked necessary nuance and historical specificity.

4.2. Case Study 2: Narrative Layer Generation and Coherence Validation

The second case study evaluated Phase 2, where the structured knowledge graph from Case Study 1 was used to generate five predefined thematic narrative layers: "Industrial Heritage," "Immigrant Cultural Networks," "Green Public Spaces," "Performing Arts Evolution," and "Culinary Identity."

The MLLM generated a 300-500 word narrative for each layer. Expert evaluation (n=5) rated the narratives highly for thematic relevance (average score 4.2/5) and factual accuracy of stated relationships (average score 4.0/5). The narratives successfully wove together assets from different modalities, for example, linking a former factory's location (GIS), its architectural description (text), and images of its current adaptive reuse [8].

Quantitative analysis showed high semantic similarity (Cosine similarity > 0.85) between AI-generated narratives and expert-written benchmarks for factual content. However, a significant gap was observed in narrative depth and contextual causality. The AI narratives tended to list and describe associated assets but were weaker at explaining the underlying socio-economic or political forces connecting them, a strength consistently present in human-written narratives [5].

Table 2 details the expert evaluation scores for each narrative layer.

Table 2. Expert Panel Evaluation of AI-Generated Narrative Layers (Scores Out of 5)

Narrative Layer	Thematic Relevance	Factual Accuracy	Coherence & Flow	Depth of Contextual Analysis

Industrial Heritage	4.6	4.4	4.2	3.8
Immigrant Cultural Networks	4.3	4.0	3.8	3.5
Green Public Spaces	4.5	4.5	4.3	3.0
Performing Arts Evolution	4.0	3.8	3.7	3.2
Culinary Identity	4.4	4.2	4.0	3.6
Average Score	4.36	4.18	4.00	3.42

Table 3 compares the quantitative coherence metrics between AI and human benchmarks.

Table 3. Quantitative Coherence Metrics: AI Narratives Vs. Human Benchmarks

Metric	AI-Generated Narrative (Average)	Human-Written Benchmark (Average)	Gap Analysis
Semantic Similarity (Cosine)	0.87	1.00 (Self)	High factual alignment.
Entity Consistency Score	0.94	0.98	Minor omissions or extraneous asset mentions.
Lexical Diversity (MTLD)	72.1	85.6	AI text showed less lexical variety.
Causal Connectives Density	1.2 per 100 words	3.8 per 100 words	Significant deficit in explaining causal or correlational relationships.

4.3. Case Study 3: Interface and Governance Scenario Testing

The final case study assessed Phase 3, where the narratives and data were implemented in a prototype interactive web map. In workshops with 15 urban planning students and practitioners, participants completed three scenario-based tasks using the tool.

The model demonstrated high utility for spatial pattern identification and asset discovery. Ninety-three percent of participants successfully identified clusters of intangible heritage in gentrification-prone areas using layered narrative filters [3]. The tool was rated highly for communicative clarity when presenting complex cultural layers to a non-specialist audience, achieving an average usability score of 4.1 out of 5.

However, limitations emerged in actionable insight generation. While the tool excelled at showing "what is where and why it might be significant," participants noted it provided less direct support for "what should be done." The jump from narrative understanding to concrete policy formulation, such as designing a specific regulatory intervention or funding program, still required substantial human expertise and

judgment. The model served better as an advanced situational awareness and communication tool than as an autonomous policy generator.

Table 4 summarizes the outcomes and participant feedback from the governance scenario testing.

Table 4. Governance Scenario Testing Results and Participant Feedback

Governance Scenario Task	Success Rate (Complete)	Avg. Task Completion Time	Key Model Strength Reported	Key Model Limitation Reported
1. Identify clusters of intangible heritage vulnerable to change.	93%	4.5 minutes	Intuitive layer toggling; clear spatial visualization of narrative themes.	Difficulty in filtering assets by "degree of vulnerability."
2. Devise a tourism corridor based on a narrative theme.	87%	7.2 minutes	Easy export of asset lists and narrative summaries for report drafting.	Lack of integrated data on current visitor flows or infrastructure.
3. Propose a community engagement focus for a neighborhood.	80%	9.0 minutes	Provided rich contextual background to inform engagement strategy.	Did not suggest specific engagement formats or stakeholder mapping.
Overall Workshop Feedback	N/A	N/A	"Powerful for exploration and communication."	"Needs tighter integration with real-time planning data and decision workflows."

4.4. Discussion

The findings from the three case studies present a nuanced picture of the proposed model's potential and constraints. The model demonstrates that generative AI, specifically MLLMs, can effectively integrate fragmented multimodal urban data into a structured knowledge system and generate coherent, thematically relevant narratives at scale. This addresses a core operational challenge in creative city governance.

However, the consistent limitations observed across all phases point to a fundamental characteristic of current generative AI when applied to complex socio-spatial systems: its proficiency is anchored in pattern recognition and synthesis within given data, but not in deep contextual reasoning or normative judgment. The AI could list the assets of "Industrial Heritage" and describe them accurately, but it struggled to deeply explain the causal forces of deindustrialization or to ethically weigh preservation against redevelopment. In governance scenarios, it excelled as a descriptive and exploratory dashboard but fell short as a prescriptive policy advisor.

This delineation is crucial for future development and application. The Urban Cultural Image Map model, in its current form, is best positioned as a cybernetic enhancement to human decision-making rather than a replacement [1, 14]. It can dramatically improve the efficiency and breadth of situational analysis, uncover hidden connections, and craft compelling narratives for public communication. To move closer to actionable governance, future iterations must focus on hybrid intelligence approaches, integrating real-time socio-economic datasets, enabling more sophisticated "what-if" simulation features, and developing interfaces that more seamlessly guide users from narrative insight to policy drafting and impact assessment.

5. Conclusion

This study set out to investigate the construction and potential of a generative AI-driven "Urban Cultural Image Map" as a multimodal narrative communication model for creative city governance. Through a tripartite theoretical framework integrating urban communication, critical geography, and AI as a cognitive prosthesis, and a corresponding three-phase methodology encompassing data structuring, narrative generation, and governance application, the research systematically developed and evaluated a functional prototype of this model. The empirical findings, derived from three interconnected case studies, provide a clear and nuanced answer to the central research question: while generative AI holds transformative potential for organizing and narrating urban cultural assets, its current capabilities are best leveraged as a powerful augmentative tool within a human-centric governance process, rather than as an autonomous decision-making agent.

The findings conclusively demonstrate that multimodal large language models can effectively address the persistent challenge of data fragmentation in urban cultural management. By performing cross-modal semantic alignment, AI can integrate disparate textual, visual, and geospatial data into a unified, queryable knowledge graph, forming a robust backbone for any narrative-driven system. Furthermore, the model proved proficient in generating coherent, thematically relevant narratives from this structured data. These narratives successfully wove together diverse cultural assets, providing a communicable and holistic overview of complex urban layers such as industrial heritage or immigrant cultural networks. This capability marks a significant advancement over traditional, static cultural asset databases, offering a dynamic and story-based interface to urban complexity.

However, the study also identified consistent and significant limitations that define the current boundaries of AI's role in this domain. The AI-generated narratives, while factually accurate and relevant, exhibited a notable deficit in depth, particularly in explicating causal relationships, contextual socio-economic forces, and the nuanced, often contested meanings embedded in cultural sites. In the governance scenario tests, the model excelled as a tool for exploratory analysis, pattern identification, and communicative storytelling but provided limited direct support for the normative, judgment-heavy tasks of policy formulation and intervention design. The transition from "what is" and "what it means" to "what should be done" remains a fundamentally human endeavor, requiring ethical reasoning, political negotiation, and deep contextual wisdom that the AI model could not replicate.

Therefore, the primary contribution of this research is the delineation of a productive and realistic pathway for integrating generative AI into creative city governance. The proposed Urban Cultural Image Map model does not represent an automation of governance but its augmentation. It serves as a cybernetic tool that dramatically enhances human capacities for data synthesis, narrative construction, and spatial storytelling. This has direct implications for urban planners, cultural policymakers, and community engagement practitioners, offering them a scalable system to make informed, narrative-aware decisions and to communicate the cultural value of places more effectively to diverse publics.

For future research and development, the findings point to several critical directions. First, efforts should focus on developing more sophisticated hybrid intelligence systems, where AI's analytical and generative strengths are tightly coupled with human expertise through interactive, iterative interfaces. Second, enhancing the model's capacity to ingest and reason with real-time, dynamic data streams, such as social media sentiment or economic indicators, would bring its narratives closer to the fluid reality of urban life. Finally, interdisciplinary collaboration must deepen, ensuring that the development of such tools is continuously informed by critical social theory, participatory design principles, and the practical needs of governance. By pursuing this integrated path, the vision of AI-enabled, narrative-sensitive, and truly creative city governance can be progressively and responsibly realized.

References

1. S. Wandelt, C. Zheng, S. Wang, Y. Liu, and X. Sun, "Large language models for intelligent transportation: A review of the state of the art and challenges," *Applied Sciences*, vol. 14, no. 17, p. 7455, 2024.
2. H. I. Ashqar, A. Jaber, T. I. Alhadidi, and M. Elhenawy, "Advancing object detection in transportation with multimodal large language models (MLLMs): A comprehensive review and empirical testing," *Computation*, vol. 13, no. 6, p. 133, 2025.
3. E. Olukanni, A. Akanmu, and H. Jebelli, "Multimodal Large Language Models in Construction Education for Learning Human–Robot Collaboration: A Narrative Review," *ASCE OPEN: Multidisciplinary Journal of Civil Engineering*, vol. 4, no. 1, p. 03126001, 2026.
4. Y. Tian, "Enhancing Geographic Information Retrieval by Generative AI and Large Language Models," Ph.D. dissertation, Arizona State University, 2025.
5. H. Liang, J. Zhang, Y. Li, B. Wang, and J. Huang, "Automatic estimation for visual quality changes of street space via street-view images and multimodal large language models," *IEEE Access*, vol. 12, pp. 87713–87727, 2024.
6. X. Qi and R. Wen, "Large Language and Multimodal Models in Archaeological Science: A Review," *Electronics*, vol. 14, no. 22, p. 4507, 2025.
7. P. Veloso, "(In) forming the new building envelope: A pedagogical study in generative design with precedents and multimodal large language models," *International Journal of Architectural Computing*, vol. 23, no. 1, pp. 96–121, 2025.
8. S. Jaradat, "AI and Big Data for Intelligent Traffic Safety: A Multimodal Approach Using Deep Learning and Large Language Models," Ph.D. dissertation, Queensland University of Technology, 2025.
9. M. Krupáš, Ľ. Urblík, and I. Zolotová, "Multimodal AI for UAV: Vision–language models in human–machine collaboration," *Electronics*, vol. 14, no. 17, p. 3548, 2025.
10. H. I. Ashqar, T. I. Alhadidi, M. Elhenawy, and N. O. Khanfar, "Leveraging multimodal large language models (MLLMs) for enhanced object detection and scene understanding in thermal images for autonomous driving systems," *Automation*, vol. 5, no. 4, pp. 508–526, 2024.
11. A. Abraham, T. Aldhanhani, W. Hamidouche, and M. Shaaban, "Harnessing the power of large language models for sustainable and intelligent transportation systems in the electric vehicle era," in *Internet of Vehicles and Computer Vision Solutions for Smart City Transformations*, Cham: Springer Nature Switzerland, 2025, pp. 85–113.
12. D. Facklam, S. J. Sweeney, S. Majumdar, and R. N. Dmello Kamath, "Describing archival photographs using multimodal LLMs: a case study on evaluating vision-language model performance for creating descriptive metadata," *The Electronic Library*, pp. 1–24, 2025.
13. W. Xing, T. Zhu, J. Wang, and B. Liu, "A survey on MLLMs in education: Application and future directions," *Future Internet*, vol. 16, no. 12, p. 467, 2024.
14. Z. B. Akhtar, "Generative artificial intelligence (GAI): From large language models (LLMs) to multimodal applications towards fine tuning of models, implications, investigations," *Computing and Artificial Intelligence*, p. 1498, 2025.