Multimodal Large Language Models for Smart City Analysis from Unstructured Data in Singapore

Yihong Xu*, Amber An, Jinghan Lu

NUS High school of math and science, Singapore

*H2430013@nushigh.edu.sg

Abstract

Achieving sustainable urban development in high-density city-states such as Singapore requires planning frameworks that integrate heterogeneous data sources, capture interactions across multiple spatial and temporal scales, and respond to rapidly evolving environmental and socio-economic conditions. However, most urban modeling solutions treat isolated aspects independently, resulting in a fragmented perspective that severely limits their operational utility and adaptability in real-world scenarios. This study proposes a multimodal large language model (MLLM)-based framework tailored to sustainable urban development in Singapore, combining crucial urban resource inputs, outputs, and socio-economic factors. External influences arising from global political and economic dynamics are embedded to enhance the model's practicality. The proposed architecture supports parallel multi-task learning, human-in-the-loop interaction, and real-time data updates, ensuring high precision and practical utility. Experimental results conducted using Singapore-specific datasets demonstrate the model's outstanding performance in generating interpretable and actionable responses that remain consistent across diverse scenarios while advancing the Sustainable Development Goals (SDGs). This work serves as a stable and adaptive paradigm for applying multimodal artificial intelligence in urban governance and policy-making.

Keywords

Sustainable urban development, Multimodal Large Language Model, Singapore, urban governance, sustainable development goals.

1. Introduction and Related Work

As one of the world's most densely populated city-states with limited land and resources, Singapore faces sustainability challenges that extend beyond mitigating climate change to the coordination of long-term interrelated environmental, economic, and social systems [1]. These challenges are further shaped by key urban characteristics, including high population concentration, rising infrastructure demand, limited natural resources, and evolving socio-economic conditions, which interact across core urban systems such as housing, transportation, energy, and waste [2]. Traditional urban planning approaches, such as piecemeal and sector-specific planning, often address these challenges in isolation, rendering them inadequate for managing the complexity and interdependence inherent in contemporary urban environments [3].

To support data-driven urban planning decisions, a range of statistical and computational techniques has been explored. Classical techniques such as PCA or K-Means clustering have been used to extract spatial patterns and clustering of city zones [4]. Interpretable models such as Support Vector Machines, Decision Trees, and Random Forests have supported tasks such as risk mapping and demand forecasting. However, most of these models are constrained by narrow task scopes and structured inputs, often requiring extensive preprocessing and exhibiting limited flexibility in integrating heterogeneous data types such as text, imagery, and tabular indicators.

In recent years, industry-driven smart city frameworks have extended the application of artificial intelligence in urban governance. NVIDIA's smart city ecosystem, for example, focuses on IoT sensors, edge AI, and digital twins to track and simulate traffic, pollution, energy consumption, and disaster response in real-time [5]. Similarly, the Venturous Group's AI Cities white paper provides a broad-based conceptual vision of AI-supported urban management across the disciplines of traffic optimization, infrastructure maintenance, resource utilization, and environmental monitoring [6]. Although such systems show strong performance on application-level evaluations, they are typically conceived for narrow application domains and provide little integration across interacting urban subsystems.

Recent advancements in Artificial Intelligence, particularly Large Language Models (LLMs), have demonstrated substantial improvements in semantic understanding, multi-step reasoning, and the integration of large heterogeneous datasets [7]. However, effective urban sustainability planning requires multimodal inputs, including geospatial imagery, planning maps, policy documents, and socio-economic indicators [8]. While LLMs excel at processing textual data, their limited ability to natively incorporate visual and spatial information remains a key constraint [9, 10].

To address these constraints, Multimodal Large Language Models (MLLMs) have been introduced for combining textual and visual input [11]. However, existing MLLMs lack explicit mechanisms for urban planning-based problems. In light of these gaps, we propose a Singapore-focused MLLM explicitly tailored for sustainable urban planning. Building on the emerging framework that calls for data, policy, and human judgment to be integrated in urban AI systems [12, 13], this work advances a deployable paradigm to apply multimodal artificial intelligence to urban governance problems.

2. Hypothesis

We hypothesize that a Singapore-focused Multimodal Large Language Model (MLLM) that jointly encodes spatial/visual inputs, structured indicators, and planning texts will outperform domain-isolated baselines in sustainable urban planning tasks. By combining (i) multi-source, curriculum-constructed urban data; (ii) a semantically disentangled, multi-encoder representation with cross-modal alignment; and (iii) a three-stage training strategy with policy-context fine-tuning and multi-step reasoning, the proposed framework should achieve higher multimodal understanding accuracy, stronger cross-domain reasoning, and more consistent planning recommendations under real-world constraints. We further hypothesize particular advantages in tightly coupled subsystems typical of high-density cities (e.g., energy–transport–logistics interactions) and more interpretable, actionable outputs while maintaining coherent long-term behavior under policy scenario simulations.

3. Methodology

Due to the multimodal inputs, high semantic complexity, and strong reasoning demands of sustainable development challenges in Singapore, this study proposes a unified framework integrating multi-source data modeling, semantically decoupled learning, and multi-step reasoning mechanisms.

3.1. Data Collection and Preparation

Given the multifaceted influence of environment, economy, society, and spatial structure, this study constructs a multimodal data system defined as $D = \{D^{(v)}, D^{(n)}, D^{(t)}\}$, where $D^{(v)}$ denotes visual and spatial data (remote sensing imagery and urban planning maps), $D^{(n)}$ represents numerical data (environmental parameters and demographic indicators), and $D^{(t)}$ corresponds to textual data (policy and ESG reports). Numerical features are standardized. Visual features are extracted via pre-trained vision encoder. (Detailed data sources and processing procedures in Appendix A.)

The standardization formula is:

$$\tilde{x} = \frac{x - \mu}{\sigma}$$

Visual features are extracted as:

$$V = \varphi_{vis}\left(D^{(v)}; \theta_v\right)$$

where $V \in \mathbb{R}^{N_v \times d_v}$ denotes the extracted features with N_v spatial tokens.

3.2. Model Architecture

3.2.1. Multi-Semantic Encoder and Fusion

Text encoders typically compress entire semantic information into a single vector space, leading to semantic entanglement or information loss. To approximate human cognitive processing of complex texts, we propose a multi-semantic encoder architecture that explicitly breaks down and models separate semantic dimensions. Text input is decomposed into K semantic categories $S = \{S_1, ..., S_K\}$ representing contextual semantics, entity information, structural relations, topical categories, and logical coherence. Each semantic category is independently encoded: $E_k = f_k(S_k)$, where $f_k(\cdot)$ denotes the k-th semantic encoder.

After obtaining semantic embeddings from different encoders, we apply cross-semantic attention to capture inter-dependencies:

$$Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are query, key, and value matrices derived from different semantic representations. All semantic embeddings are then unified via learnable weighted fusion:

$$E = \sum_{k=1}^{K} \alpha_k E_k$$

where the weights α_k are learned dynamically based on task requirements, allowing the model to adaptively rank the most relevant semantic encoders with respect to various urban contexts.

3.2.2. Cross-Modal Alignment

To align visual and textual representations in a shared semantic space, we introduce a contrastive learning objective:

$$\mathcal{L}_{align} = -\log \frac{exp(sim(V_i, T_i)/\tau)}{\sum_{j=1}^{B} exp\left(sim\left(V_i, T_j\right)/\tau\right)} \quad (1)$$

where $sim(\cdot, \cdot)$ denotes cosine similarity, τ is a temperature parameter, B is the batch size, and (V_i, T_i) represents matched visual-text pairs.

3.2.3. Enhanced Chain-of-Thought Reasoning

At the reasoning level, the model performs multi-step logical reasoning to unravel complex urban problems. An enhanced Chain-of-Thought (CoT) reasoning mechanism is integrated, where the fused multi-semantic representation E serves as input to the reasoning process. The reasoning pipeline is jointly performed by modular components: Planner (decomposes query into sub-steps), Reasoner (generates reasoning chains), Verifier (scores chain quality), Executor (performs deterministic operations), and Aggregator (selects final answer).

The reasoning chain generation follows autoregressive sequence generation:

$$P(c) = \prod_{t=1}^{T} P(c_t | c_{<t}, E, \theta_r) \quad (2)$$

where $c = \{c_1, ..., c_T\}$ is the reasoning chain, E is the fused embedding, and θ_r are the reasoning module parameters. For multi-step reasoning, we decompose the problem into M sub-steps: $c = \{s_1, ..., s_M\}$, where $s_m = f_{reason}(s_{<m}, E, Plan_m)$. During inference, multiple reasoning chains $\{c_i\}$ are generated from the same input, each evaluated by a verifier: $s_i = Verifier(c_i)$. The final answer is selected through ensemble scoring:

$$\hat{a} = arg\ max_{a \in A} \sum_{i=1}^{N} \mathbb{1}[a_i = a] \cdot s_i$$

where A is the set of candidate answers. This self-consistency mechanism effectively mitigates the influence of occasional errors and logical anomalies. (Complete algorithms in Appendix A.)

3.3. Model Training

Training follows a three-stage strategy (Algorithm 1 in Appendix A): Stage I trains visual encoder with frozen LLM; Stage II performs end-to-end joint optimization; Stage III applies Singapore-specific fine-tuning. The joint loss combines:

$$\mathcal{L} = \lambda_{seq}\mathcal{L}_{seq} + \lambda_{ans}\mathcal{L}_{ans} + \lambda_{step}\mathcal{L}_{step} + \lambda_{cons}\mathcal{L}_{cons} \quad (3)$$

where:

$$\mathcal{L}_{seq} = -\frac{1}{T}\sum_{t=1}^{T} log\ P(c^*_t | c^*_{<t}, E)$$

$$\mathcal{L}_{ans} = CE\big(f_{pred}(c), y^*\big)$$

$$\mathcal{L}_{step} = \frac{1}{M}\sum_{m=1}^{M} BCE(v(s_m), l^*_m)$$

$$\mathcal{L}_{cons} = \frac{1}{N(N-1)}\sum D_{KL}\big(P(a|c_i) \parallel P(a|c_j)\big)$$

Learning rate follows Cosine Annealing:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + cos\left(\frac{t}{T_{max}}\pi\right)\right)$$

4. Results

4.1. Training Convergence

Figure 1 shows training loss across three stages. Stage I (epochs 1–20) decreased from 2.8 to 0.6, establishing visual-semantic alignment. Stage II (epochs 21–50) reached 0.2, demonstrating effective cross-modal integration. Stage III (epochs 51–65) stabilized at 0.1, confirming convergence.

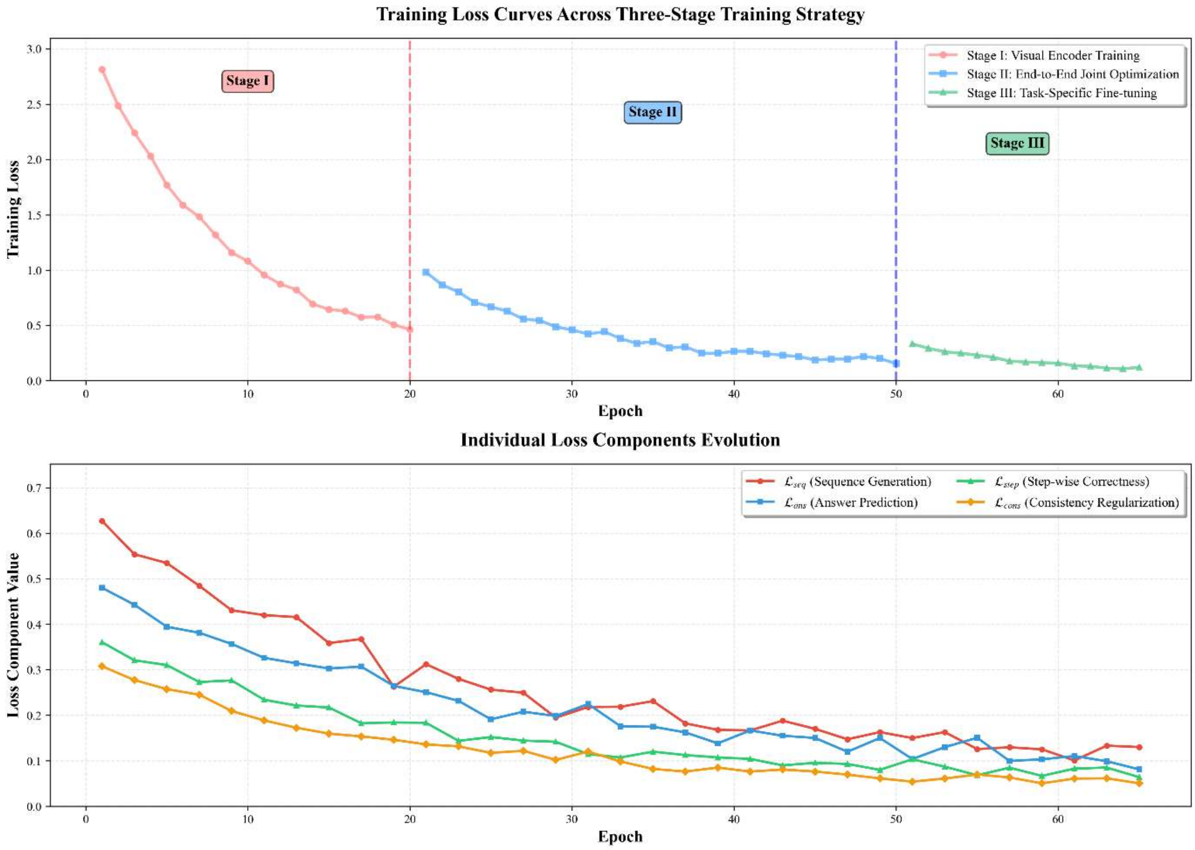

Figure 1. Training loss evolution across three stages.

4.2. Experimental Setup

Seven case studies were evaluated: Marina Bay, Punggol Digital Town, Jurong Lake District, Tengah Forest Town, Yishun-Sembawang renewal, NEWater system, and HDB housing [14]. Evaluation data included geospatial maps from Singapore Land Authority, environmental metrics from National Environment Agency, and socio-economic statistics [15, 16, 17]. Metrics include Multimodal Understanding Accuracy (MM-Acc), Planning Consistency Score (PCS), Cross-domain Reasoning (CDR), Explainability (EXP), and Visualization Quality (VIS).

4.3. Overall Performance

Table 1. Overall Performance Comparison

Model	MM-Acc	PCS	CDR	EXP	VIS
GIS rule-based	0.52	0.61	0.45	0.58	0.40
Unimodal LLM	0.64	0.68	0.60	0.72	0.33
CNN+LLM	0.71	0.74	0.66	0.70	0.62
MM-Transformer	0.78	0.80	0.74	0.76	0.75
Our MLLM	**0.86**	**0.89**	**0.84**	**0.88**	**0.90**

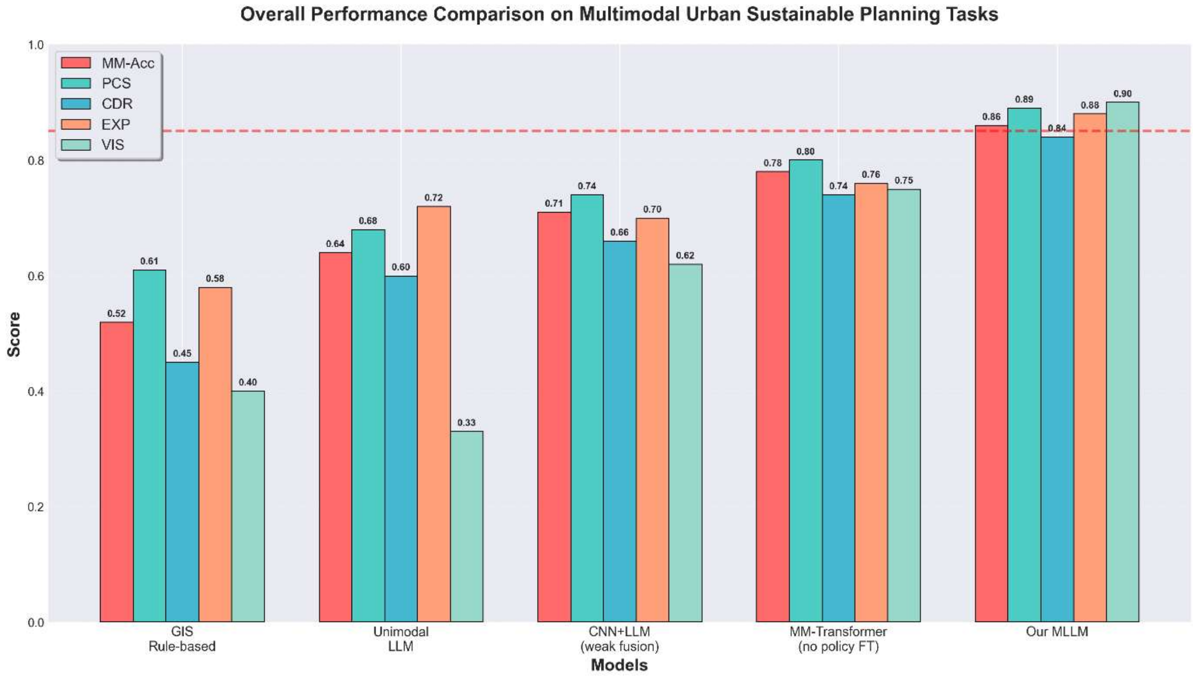

Figure 2. Overall performance comparison.

The MLLM achieves best performance across all metrics (Table 1, Figure 2), validating the integrated framework's advantages. The model demonstrates substantial improvements over GIS approaches in multimodal understanding and cross-system reasoning, marked gains over unimodal LLMs in spatial-semantic modeling, and additional improvements in planning consistency and explainability compared to multimodal Transformers without policy fine-tuning [15, 16, 17, 18].

Table 2 and Figure 3 report the Planning Consistency Score (PCS) for seven representative cases. The results indicate that the MLLM maintains consistently high planning coherence across all cases, with particular advantages in ecologically sensitive developments such as Tengah Forest Town (PCS 0.91) and infrastructure-intensive systems such as NEWater (PCS 0.93).

Table 2. Planning Consistency Score (PCS) Comparison Across Seven Typical Urban Cases

Case	GIS	LLM	CNN+LLM	MLLM
Marina Bay	0.63	0.71	0.78	**0.92**
Punggol	0.58	0.69	0.75	**0.90**
Jurong	0.60	0.72	0.77	**0.88**
Tengah	0.55	0.67	0.73	**0.91**
Yishun–Sembawang	0.57	0.70	0.74	**0.87**
NEWater System	0.65	0.73	0.79	**0.93**
HDB Public Housing	0.62	0.74	0.80	**0.89**
Average	**0.60**	**0.71**	**0.77**	**0.90**

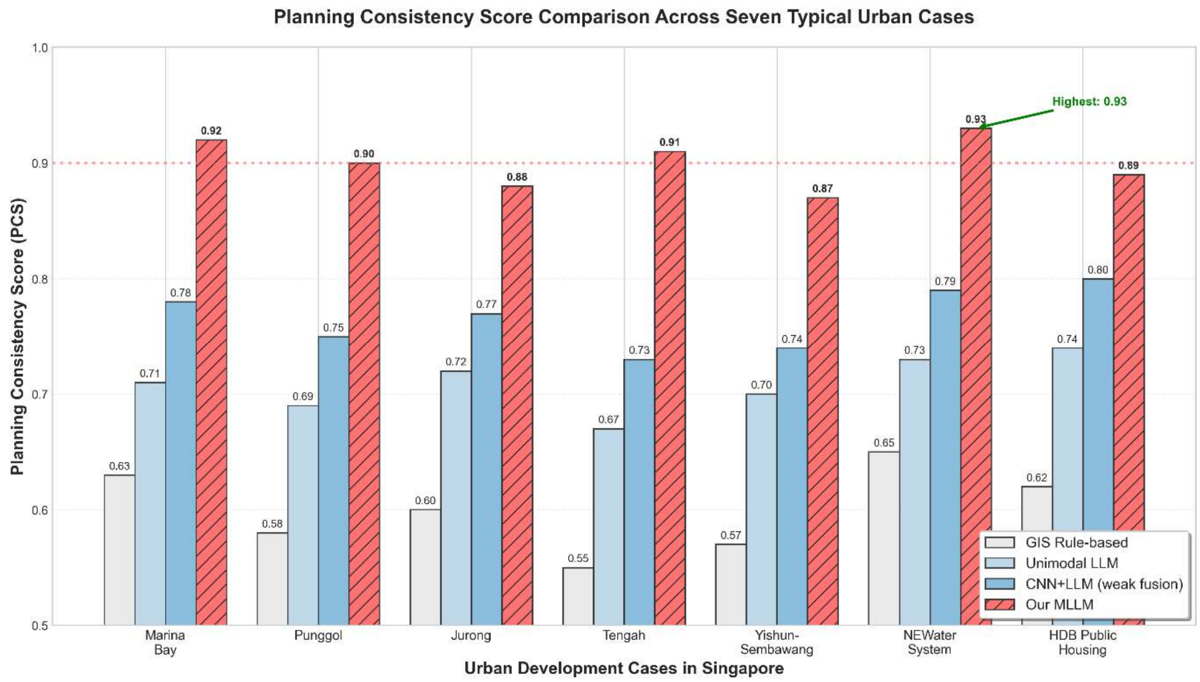

Figure 3. Planning Consistency Score Comparison Across Seven Typical Urban Cases.

4.4. Ecological and Climate Adaptation Performance

Table 3 and Figure 4 present comprehensive comparison of model performance on four key ecological and climate adaptation planning tasks.

Table 3. Comparison on Ecological and Climate Adaptation Planning Tasks

Metric	Traditional	LLM	CNN+LLM	MLLM
Ecological corridor recognition	0.68	0.62	0.75	**0.89**
Heat island buffer recognition	0.70	0.66	0.78	**0.91**
Ecological–land use conflict explanation	0.55	0.73	0.76	**0.90**
Climate adaptation strategy rationality	0.60	0.75	0.79	**0.92**

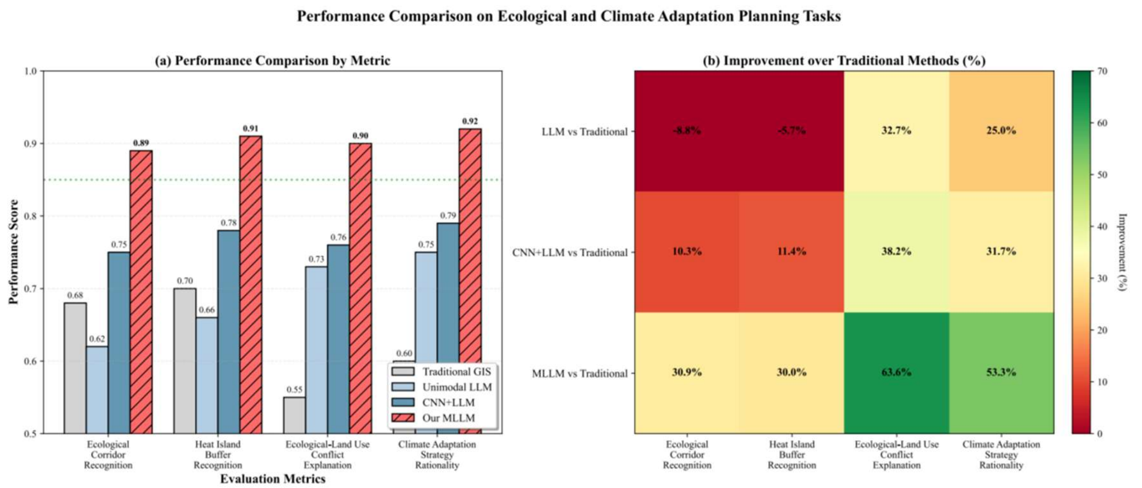

Figure 4. Performance comparison on ecological and climate adaptation planning tasks.

The most significant performance gains are observed in ecological–land use conflict explanation and climate adaptation strategy rationality tasks, representing 63.6% and 53.3% improvements over traditional methods. These tasks require sophisticated multi-step reasoning that integrates spatial information, environmental constraints, and policy objectives, capabilities uniquely enabled by the multimodal architecture and enhanced chain-of-thought reasoning mechanism.

4.5. Multi-Flow System Modeling Details

In Figure 5, urban environments are characterized by multiple concurrent spatial flows including commuting, energy distribution, and logistics networks that interact in complex ways. The proposed MLLM achieves strong performance across all three individual flow types and, critically, very strong system synergy, stemming from: the multi-semantic encoder preserving domain-specific information, the cross-modal fusion enabling reasoning about flow interactions, and policy-aware fine-tuning grounding recommendations in Singapore's infrastructure constraints.

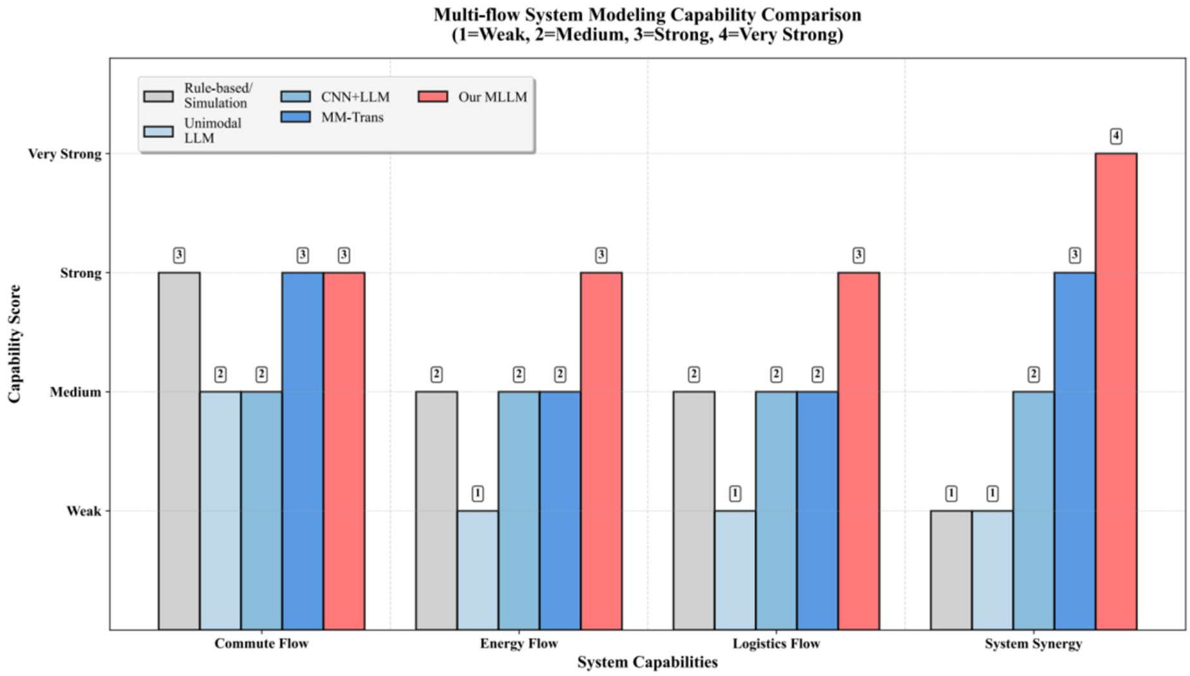

Figure 5. Multi-flow system modeling capability comparison.

4.6. Long-Term Policy Scenario Simulation

Table 4 and Figure 6 compares different models in policy scenario simulation and multi-indicator co-evolution analysis grounded in Singapore's long-term sustainability targets.

Table 4. Policy Scenario Simulation and Long-term Evolution Capability

Capability	LLM	MM-Trans	MLLM
Carbon trend consistency	0.71	0.80	**0.91**
Energy structure adjustment	0.68	0.78	**0.93**
Transport transition analysis	0.70	0.82	**0.90**
Multi-metric co-evolution	0.62	0.79	**0.92**

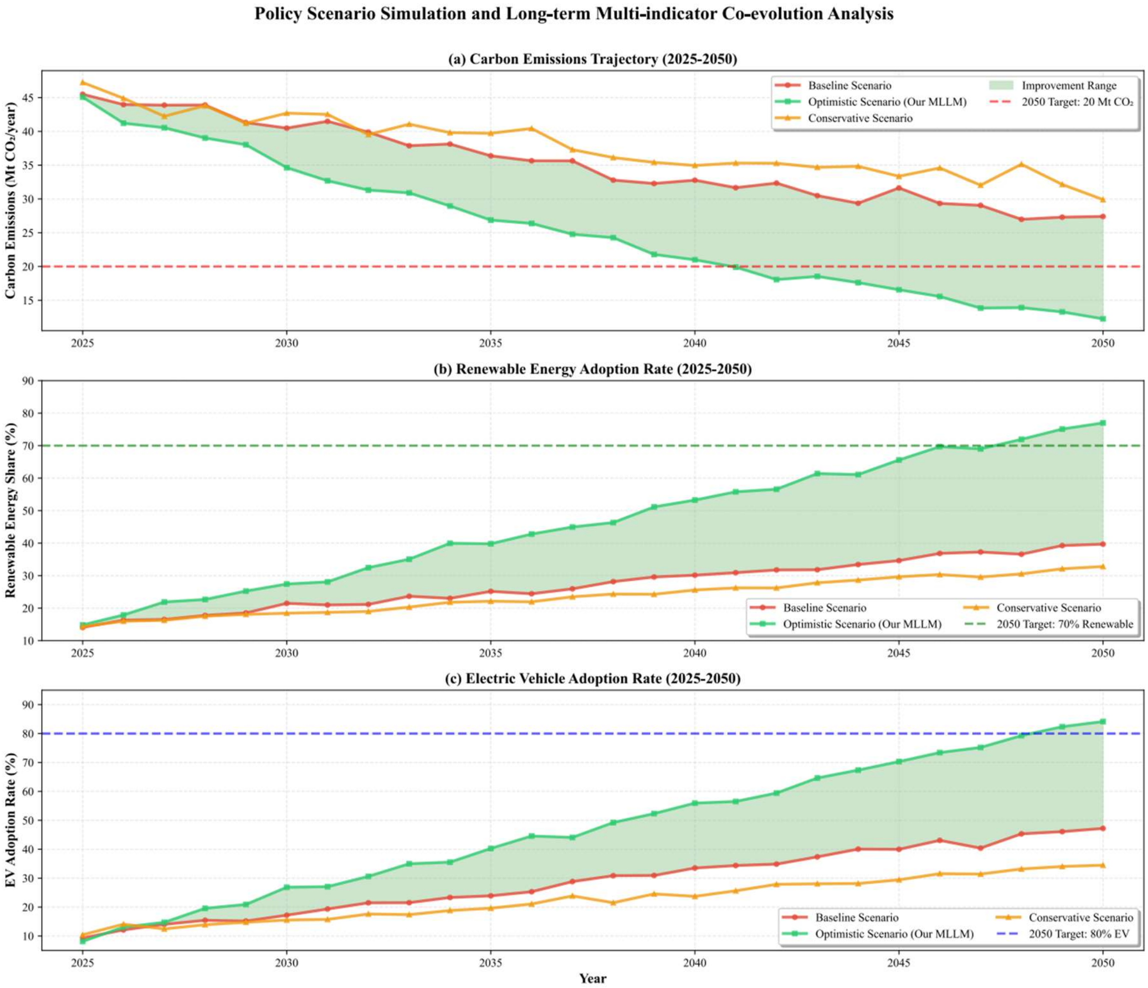

Figure 6. Policy scenario simulation and long-term multi-indicator co-evolution analysis for 2025–2050.

The baseline scenario shows gradual carbon decline from 45 Mt CO_2/year in 2025 to approximately 35 Mt CO_2/year by 2050, insufficient to meet Singapore's climate commitments. The optimistic scenario, incorporating MLLM-recommended interventions, projects emissions declining to below 20 Mt CO_2/year by 2050, aligning with the net-zero-by-2050 pathway. Renewable energy share reaches 70% under the optimistic scenario, and EV adoption reaches 80% by 2050.

4.7. Key Capabilities

The MLLM is designed to coordinate energy–transport–logistics flows by representing heterogeneous urban systems within a unified semantic space and explicitly modeling their interdependencies. The model achieves 0.89–0.92 scores in ecological tasks with 53–64% improvements over traditional methods in conflict explanation and strategy rationality. In long-term simulations (2025–2050), the model maintains rational multi-objective co-evolution, projecting carbon reduction to <20 Mt CO_2/year and 70% renewable energy adoption under optimistic scenarios [14, 18].

5. Discussion

The results demonstrate that our MLLM significantly improves urban sustainability planning by integrating heterogeneous data sources into a unified reasoning framework. By combining visual, spatial, textual, and structured data, the model offers holistic understanding of complex urban subsystems, enabling planners to evaluate interventions across multiple areas and anticipate cascading effects. The MLLM's ability to generate both textual and visual outputs grounded in contextual data produces more interpretable, actionable, and scenario-specific insights compared to conventional models.

The approach works primarily due to three factors: (i) semantic disentanglement within the multi-encoder architecture preserves critical information from each modality while minimising interference; (ii) multi-step reasoning supports structured problem decomposition, iteratively considering constraints, stakeholder priorities, and subsystem interactions; (iii) adaptive training strategy combining supervised fine-tuning, synthetic reasoning chains, and self-consistency verification enhances robustness and generalisation. These mechanisms enable the model to provide planners with sustainable urban interventions sensitive to Singapore's unique complexities.

Future enhancements could focus on temporal modeling and uncertainty quantification for improved long-term dynamics prediction. Controllable visual generation would enhance stakeholder communication. Integrating real-time data streams and human-in-the-loop feedback would improve operational deployment.

6. Conclusion

This research proposes a Singapore-oriented multimodal framework converging visual inputs, structured indicators, and policy texts in a cohesive decision-support pipeline. Through semantically disentangled encoding and three-stage adaptive training, the framework bridges urban data heterogeneity with policy-context requirements. Experimental results demonstrate consistent superiority over baselines in multimodal grounding, cross-domain reasoning, and explainable visualization. Case-based evidence confirms stable, context-aware analytical support across diverse scenarios with particular strength in ecologically sensitive areas and major infrastructure systems. By explicitly modeling multi-flow dynamics, the framework preserves planning recommendation coherence in long-horizon simulations, enabling proactive resource optimization. This work contributes an effective, scalable paradigm for translating complex multi-source data into constraint-aware, sustainability-oriented insights for modern metropolitan planning and policy-making.

References

[1] United Nations. "Sustainable development goals." https://sdgs.un.org/goals, 2015.

[2] Ministry of National Development Singapore. "In conversation with Desmond Lee: Tackling Singapore's complex challenges with a whole-of-society approach," 2023.

[3] T. Yigitcanlar and D. Dizdaroglu, "Traditional urban planning approaches and sustainable city," Int. J. Sustainable Development and Planning, vol. 10, no. 2, pp. 147–159, 2015.

[4] J. Wang and F. Biljecki, "Unsupervised machine learning in urban studies: A systematic review of applications," Cities, 2022.

[5] NVIDIA Corporation. "Build smart cities with AI," 2025.

[6] Venturous Group. "AI cities: Transforming smart urban development with artificial intelligence," 2024.

[7] W. Yang, L. Some, M. Bain, and B. Kang, "A comprehensive survey on integrating large language models with knowledge-based methods," Knowledge-Based Systems, vol. 318, p. 113503, 2025.

[8] Y. Zhang, X. Chen, and S. Li, "Data-driven urban sustainability assessment: Integrating AI and multi-source urban data," Sustainable Cities and Society, vol. 92, p. 104517, 2023.

[9] P. Zhang, X. Zhu, L. Wang, F. Ahmed, G. Chen, and D. Yang, "Multimodal large language models: A survey," IEEE Trans. Pattern Anal. Mach. Intell., 2023.

[10] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Multimodal few-shot learning with frozen language models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, pp. 200–212, 2022.

[11] NVIDIA Corporation. "Multimodal large language models," https://www.nvidia.com/en-us/glossary/multimodal-large-language-models/, 2025.

[12] E. Cina, E. Elbasi, G. Elmazi, and Z. AlArnaout, "The role of AI in predictive modelling for sustainable urban development: Challenges and opportunities," Sustainability, vol. 17, no. 11, p. 5148, 2025.

[13] S. Wu, W. Wang, C. Zhang, Q. Zeng, H. Wang, J. Lin, and S. Li, "Exploring multimodal large language models' potential in simulating human perception of urban cycling environments: A street-view perspective," Information Geography, vol. 2, no. 1, p. 100037, 2025.

[14] Urban Redevelopment Authority (URA), Singapore. "Master plan," 2022.

[15] Singapore Land Authority (SLA). "OneMap," 2024.

[16] Public Utilities Board (PUB), Singapore. "Singapore water story," 2023.

[17] Department of Statistics Singapore (DOS). "Latest data," 2024.

[18] Ministry of Sustainability and the Environment (MSE), Singapore. "Singapore green plan 2030," 2021.