GenAI Models Capture Urban Science but Oversimplify Complexity

Yecheng Zhang [1] **Rong Zhao** [1] **Zimu Huang** [1] **Xinyu Wang** [1] **Yue Ma** [1] **Ying Long** [1 2]

Abstract

Generative artificial intelligence (GenAI) models are increasingly used for scientific data generation, yet their alignment with empirical knowledge in urban science remains unclear. Here, we introduce **AI4US** (Artificial Intelligence for Urban Science), a framework to systematically generate, evaluate, and calibrate urban data from leading GenAI models, testing their fidelity across both symbolic and perceptual domains. For the symbolic domain, we benchmark generated data against foundational urban theories concerning scale, space, and morphology. For the perceptual domain, we validate the models' visual judgments against human benchmarks and, critically, leverage their generative control to conduct in causal experiments on urban perception. Our findings show that while GenAI models reproduce core theoretical patterns, the generated data exhibit crucial limitations: poor diversity, systematic parametric deviations, and improvement from prompt engineering. To address this, we introduce a post-hoc calibration procedure using optimal transport, which produces synthetic symbolic datasets with demonstrably higher fidelity.

1. Introduction

Urban science has long sought to decipher the complex behaviors of cities from early demographic surveys to computational modeling (Geddes, 1915; Van Noorden and Perkel, 2023; Birhane et al., 2023; Binz et al., 2025), which manifest across multiple scales (macro, meso, micro) and dimensions (Long and Zhang, 2024; Zhang et al., 2025a). The endeavor to formulate urban theory from empirical observation has traditionally relied upon two principal data modalities: symbolic data, which abstracts urban phenomena into text and numerics, and perceptual data, which captures the direct sensory experience of the urban environment through modalities like visual imagery (Kitchin, 2016; Sudmant et al., 2024; Batty, 2024). Since the recent explosion of urban data (Long and Zhang, 2024; Zhang et al., 2025a) from remote sensing, sensor networks, and digital platforms promised a revolution, struggling to move beyond fragmented, context-dependent insights toward broadly generalizable urban theory.

However, theoretical innovation has stagnated (Kitchin, 2016; Sudmant et al., 2024; Batty, 2024). This impasse stems from a persistent methodological gap: the lack of scalable and systematic experimental designs that allow for robust, cross-city replication and validation of scientific claims (Sudmant et al., 2024; Wiig, 2018; Li et al., 2024a). This theoretical stagnation is rooted in the inherent difficulty of conducting controlled experiments in urban settings. Historically, urban experimentation has evolved through several paradigms, each with significant limitations. Early approaches treated the city as a living laboratory, relying on direct observation and policy trials, but these methods lacked scalability and control (Apostolos, 2010; Mayer-Schönberger and Cukier, 2013). Physical and analogue models offered controllable surrogates but were often oversimplified, failing to capture the non-linear dynamics of human behavior at a meaningful scale (Batty, 2012; Egger and Yu, 2022). The subsequent rise of digital models, such as cellular automata and agent-based simulations, introduced rule-based experimentation but hinged on hard-coded assumptions and static calibration, limiting their ability to represent the nuanced, adaptive behaviors of individuals (Wegener, 2021; Long and Shen, 2015; Zhang et al., 2025b). These shortcomings are magnified by the profound transformations of contemporary cities, from autonomous mobility to platform economies, creating an urgent need for a new experimental paradigm capable of capturing the complex human-in-the-loop dynamics that define modern urban life (Long and Zhang, 2024; Gottweis et al., 2025).

Nowadays, advances in GenAI are spearheading a new paradigm of data-driven scientific discovery (Balsa-Barreiro et al., 2024; Chen et al., 2024; Bhandari et al., 2024). In fields like biology and materials science, its success lies in predicting specific physical states from vast solution spaces, exemplified by AlphaFold's protein structure prediction (Jumper et al., 2021) and MatterGen's discovery of

[1] School of Architecture, Tsinghua University, Beijing, China [2] Hang Lung Center for Real Estate, Key Laboratory of Ecological Planning & Green Building, Ministry of Education, Tsinghua University, Beijing, China. Correspondence to: Ying Long <ylong@tsinghua.edu.cn>.

novel materials (Zeni et al., 2024). The challenge in urban science is fundamentally different. AI's role in urban science is to capture the broad statistical patterns emerging from complex, adaptive urban systems governed by context-dependent social behaviors rather than universal physical laws (Sun et al., 2023). This distinction makes GenAI models particularly suitable. Their evolution from Large Language Models (LLMs), capable of generating and reasoning over symbolic data, to Multimodal Large Language Models (MLLMs) that also encompass perceptual data, unifies the two critical domains of urban information. Trained on diverse records of human behavior and expression, these models can thus generate high-fidelity synthetic data, enabling scalable urban experimentation where formal modeling has reached its limits.

In urban science, however, the potential of GenAI to function as a new form of virtual city laboratory that can produce experimental data (Li et al., 2024b; Zhou et al., 2024), analyze urban phenomena, and drive theoretical advancement remains largely unexplored (Balsa-Barreiro et al., 2024; Ye et al., 2025). Initial studies have shown promise, using GenAI models to generate synthetic mobility data, simulate residents' daily behaviors, or assess urban systems (Bhandari et al., 2024; Wang et al., 2024). However, these pioneering efforts often remain fragmented. A significant research gap persists in systematically validating the fidelity of these worlds generated by GenAI against established urban theories (Li et al., 2025). Furthermore, the inherent geographical and social biases within GenAI models pose a critical challenge to their reliability as scientific instruments (Manvi et al., 2024; Abbasi et al., 2025). Therefore, a structured and rigorous framework is necessary to move beyond isolated applications and systematically evaluate the potential and pitfalls of GenAI as a foundation for the next generation of urban science.

To address these gaps, we introduce the AI4US (Artificial Intelligence for Urban Science) framework. This framework establishes a 'Generate-Evaluate-Calibrate' workflow to systematically evaluate generative AI as a first-generation proxy by probing the implicit urban knowledge embedded within them. Our diagnostic approach evaluates the fidelity of AI-generated data across both symbolic and perceptual domains. For symbolic data, we test the models' ability to reproduce foundational urban theories like urban scaling laws, distance decay, and urban vitality (Ye et al., 2025; Wang et al., 2024; Li et al., 2025). For perceptual data, we assess their multimodal capacity to generate realistic urban scenes and replicate human-centric visual perception. Crucially, this generative control enables a form of in causal experimentation, allowing for systematic tests of how specific urban elements impact human perception. These theories were chosen to create a comprehensive testbed, as they represent fundamental urban mechanisms across different scales and dimensions (Manvi et al., 2024; Abbasi et al., 2025; Tan et al., 2025; Xu et al., 2021; Jacobs, 1961; Batty and Longley, 1994). Moving beyond diagnosis, this study introduces a novel post-hoc calibration procedure, demonstrating how a small sample of real data can be leveraged to rapidly assess these biases for a given urban dataset and subsequently correct the distribution of the larger generated dataset. This computationally efficient approach enhances the fidelity for downstream tasks, such as training predictive models or seeding simulations, without requiring costly modifications to the foundational GenAI models. Our findings reveal that the path toward the future will require advanced models (Abbasi et al., 2025).

2. Results

2.1. GenAI models capture the fundamental statistical forms of urban theories

The investigation reveals that GenAI models possess a strong ability to recognize and reproduce the foundational forms of established urban theories across multiple scales, from macro-level city systems to micro-level neighborhood dynamics. At the macro-scale, when tasked with generating data for urban scaling laws, GenAI models consistently produce outputs that adhere to the characteristic power-law relationship. The generated data exhibits a strong fit to the scaling law equation, achieving a high average coefficient of determination ($R^2 = 0.804$) (Fig. 1a). Most models successfully replicated the expected super-linear relationship ($\beta > 1$) for socioeconomic output and, in the case of GPT-4o, the sub-linear relationship for infrastructure.

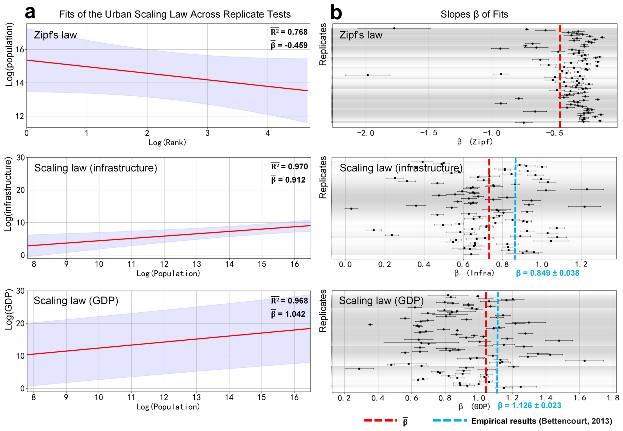


Figure 1. Urban scaling law reproduced on generated data with GenAI models: generated versus empirical coefficients. 'Replicate Tests' contains 100 independent duplicate data generation experiments for each GenAI model, and each experiment generates data of 100 cities, among which GPT-4o has best performance (See Supplementary Experiment 6). The diagram shows the results of GPT-4o.

Moving to the meso-scale, the models display a non-trivial grasp of the concentric-zone structure posited by distance-decay theory. When generating land density surfaces, the outputs fit the inverse-S formulation with high accuracy, showing a mean R^2 of 0.988 (Fig. 2a).

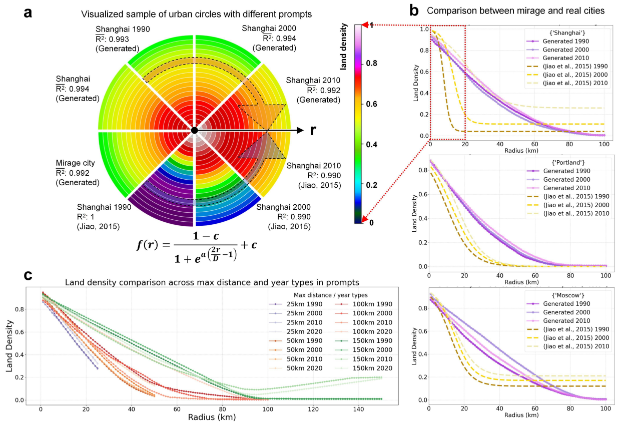


Figure 2. Urban decay reproduced on generated data with GenAI models: generated versus empirical statistical and spatial patterns. **a**, Concentric-ring land-density profiles for three representative cities on different continents across three decades (1990-2010), together with the corresponding 'Mirage cities' produced by GPT-4o when both city name and time period are included in the prompt. **b**, Sensitivity of the generated decay curves to prompts that specify different maximum radii and historical periods. **c**, Direct side-by-side comparison of Beijing's empirical profile with that of its mirage city generated by GenAI models.

Furthermore, this capability extends to descriptive theories that lack a single mathematical form, such as Jane Jacobs' theory of urban vitality. The test of this theory reveals a distinct capability of GenAI, where the model is prompted to act as a "resident" to simulate a "social consensus" on what makes a place livable. The model's outputs revealed nuanced and theoretically consistent insights, identifying 'Building Mix Index' and 'Tall Building' as having the highest alignment with the theory, while 'Aged Building' showed the greatest deviation (Fig. 3). This demonstrates that an GenAI model can process the complex, multi-faceted information embedded in qualitative urban theories to generate data that is structurally "form-similar" and directionally consistent with their core tenets. More importantly, it shows a capacity to translate these qualitative concepts into structured, testable hypotheses regarding the relative importance of different urban elements, establishing a baseline of theoretical competence.

2.2. A new paradigm for qualitative theory exploration in visual urban space

This research demonstrates that GenAI models can function as a new paradigm for exploring and validating theories in visual urban space. The fidelity of the generated images was first confirmed in a blind expert validation test. AI-generated images received a mean realism score of 7.83 (SD = 1.6), which was not statistically different from the 7.68 (SD = 1.8) scored by real photographs (t(198) = 0.65, p = 0.52). This high fidelity was complemented by substantial diversity, achieved via a structured-prompting strategy that increased the average CLIP semantic distance to 0.3310. This represents a 35.8% improvement over a baseline single-prompt approach ($D_{CLIP} = 0.2438$) and surpasses the diversity of a 1,000-image real-world benchmark ($D_{CLIP} = 0.2905$), enabling the creation of semantically rich datasets.

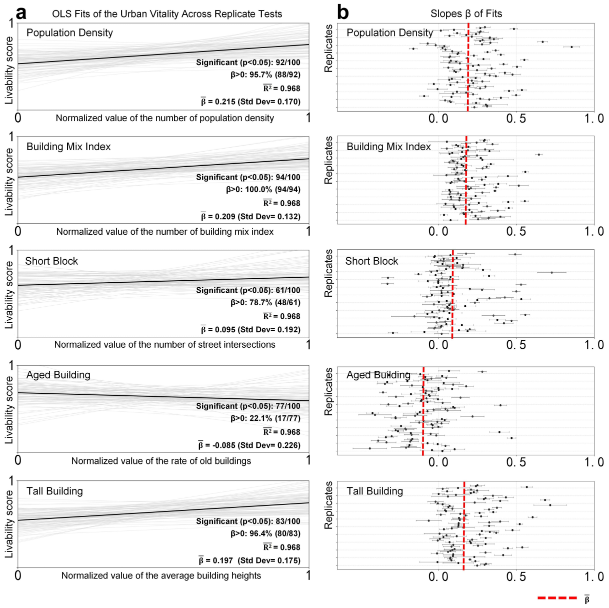


Figure 3. Urban vitality revealed on generated data with GenAI models. 'Replicate Tests' contains 100 independent duplicate data generation experiments for the GenAI models, and each experiment generates data of 100 blocks. The diagram shows the results of GPT-4o.

Having established the quality of the generated data, we validated the AI's ability to emulate human perception and identify its visual drivers. The AI's perceptual judgments showed substantial agreement with human choices from the Place Pulse dataset across all six evaluated categories, with Kappa values consistently above 0.2 and agreement rates approaching 70% (Fig. 4a). Using this validated AI perception, we then quantified the influence of visual elements on these attributes through a multiple linear regression model. The analysis revealed several significant relationships (Fig. 4b). For instance, the presence of a 'person' was a strong positive predictor for 'lively' ($\beta = 0.223, p < 0.001$), while features such as 'wall' negatively impacted the perception of 'beautiful' ($\beta = -0.247, p < 0.001$). To further validate these patterns, we compared our results to the benchmark findings of Place Pulse (Naik et al., 2014), which shows that 'tree' and 'grass' are strong positive predictors for 'beautiful', while 'sky' is a significant negative predictor for 'safe'.

Finally, the "generate-modify-evaluate" paradigm was used to test the AI's utility for causal inference by quantifying the impact of thematic interventions (Fig. 4c). The addition of Natural Elements produced the most pronounced effect, significantly increasing the perceived 'beautiful' score by an average of 2.5 points (Δ Score) and having the strongest negative impact on the 'depressing' score. Conversely, adding Traffic Elements was the most effective intervention for increasing the 'lively' score, demonstrating the framework's ability to test the perceptual outcomes of specific design strategies.

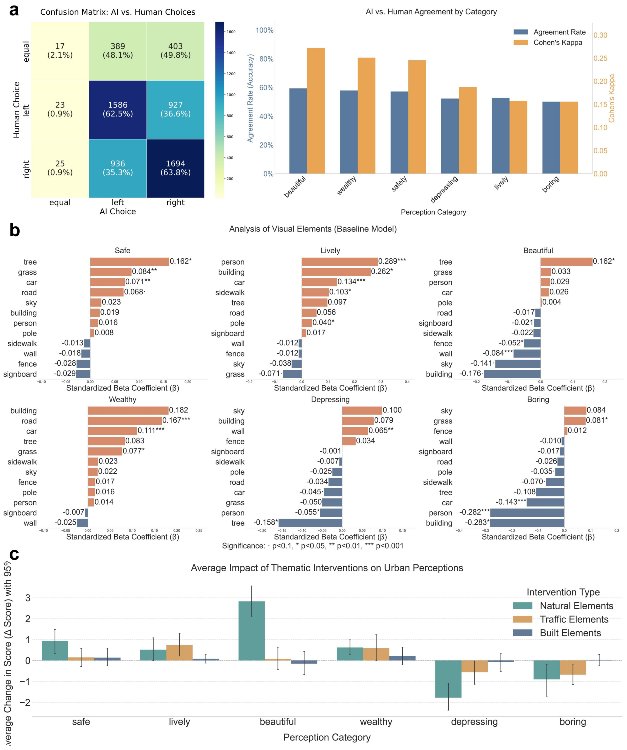


Figure 4. Potential of GenAI models in qualitative theory exploration in visual urban space. **a**, AI perception aligns with human judgment. A confusion matrix of pairwise choices (left) and per-category agreement rates with Cohen's Kappa scores (right). **b**, Identifying influential visual elements. Standardized beta coefficients (β) from regression models reveal the key visual features that are positively or negatively correlated with each of the six perceptual scores. **c**, Causal inference via thematic interventions. The average change in perception scores (Δ Score) quantifies the causal impact of programmatically adding thematic element categories to scenes: Natural Elements (trees, grass, sky, water, bushes), Traffic Elements (cars, trucks, bridges, roads), and Built Elements (buildings, fences, walls). Error bars represent 95% confidence intervals.

2.3. Mirage cities systematically oversimplify real-world complexity

While GenAI models can replicate theoretical forms based on statistical data, a deeper analysis reveals a significant gap between their generated "Mirage cities" and the complexity of real urban systems. This discrepancy manifests as a series of systematic, non-random deviations. A primary finding is the lack of diversity in the generated data compared to empirical data. The analysis of data distributions reveals that generated data is significantly more homogeneous. Without specific geographic prompts, the generated data's numerical scale diverges substantially from real data, resulting in a low average Overlap Ratio (OR) of just 28.6% per bin (Fig. 5a). More critically, the Jensen-Shannon Divergence (JSD) between real-world data samples (R-R) is, on average, 122.54% higher than the divergence within generated samples (G-G) (Fig. 5b). This indicates that real data exhibits substantially greater internal variation. This inherent "over-smoothing" tendency is a foundational reason why generated data, while theoretically clean, lacks the real-world noise and heterogeneity essential for robust scientific analysis.

This lack of diversity contributes to specific parametric and structural distortions when fitting theories. At the macro-scale, a systematic underestimation of the scaling exponent (β) is observed in the urban scaling law test (Fig. 1b). The models consistently produce lower exponents for both socioeconomic output and infrastructure compared to established empirical values (e.g., $\beta \approx 1.15$ for GDP and $\beta \approx 0.85$ for infrastructure), suggesting they capture the "shape" of the law but fail to grasp its true "steepness" or magnitude. At the meso-scale, this manifests as the generation of "idealized" spatial decay curves that are smoother than reality (Fig. 2a, Fig. 3b). The generated data gravitates toward the theoretical extrema, with values approaching 1 at the center and 0 on the periphery. Consequently, this idealized output fails to capture the "messiness" of real cities, which contain features like central parks or suburban industrial belts that cause deviations from a perfect curve. The distributional difference in these generated decay curves is 149% smaller than that of empirical data, further underscoring their idealized nature. This pattern of oversimplification extends to the visual domain. The AI's perceptual judgment, while directionally correct, showed weaknesses in abstract cases, achieving lower consistency on subjective concepts like 'boring' ($\kappa = 0.155$) and 'lively' ($\kappa = 0.157$). Most notably, the AI exhibited a strong bias against ambiguity, agreeing with human 'equal' judgments in only 2.1% of cases and instead forcing a definitive choice (Fig. 4a).

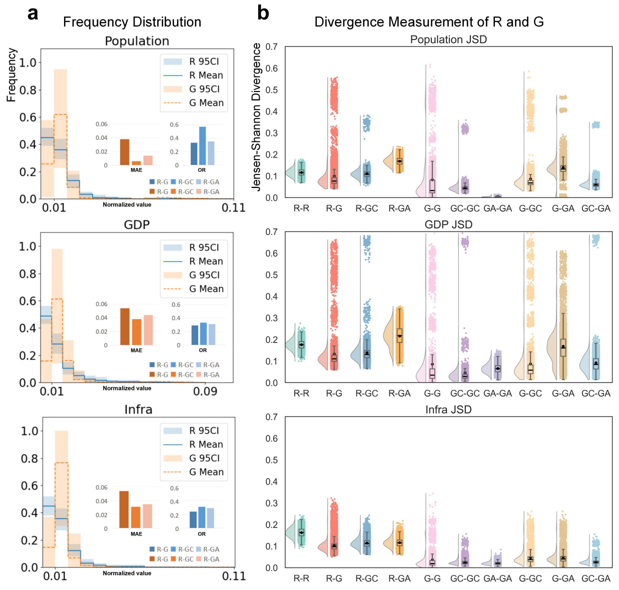


Figure 5. Data distribution divergence analysis. **a**, Equal-width binning with relative frequency normalization was applied to convert raw sample counts into scale-invariant frequency distributions. Mean Absolute Error (MAE) of bin-level frequencies quantified the central tendency divergence between real data (R) and generated data (G), calculated as the sum of absolute differences across all 15 bins. Overlap Ratio (OR) measured quantile interval congruence between corresponding bins of R and G distributions. **b**, Jensen-Shannon Divergence (JSD) comparisons among four datasets: real data (R), baseline generated data (G), China-prompted generated data (GC), and USA-prompted generated data (GA). The inclusion of geographical constraints, such as 'China', increased GPT-4o's constraint-triggering rate by 24% (24/100 trials) (e.g., 'I can't provide real-time data such as the current population').

2.4. The potential for improving AI-generated data

A post-hoc calibration procedure based on Optimal Transport (OT) is shown to be the most effective intervention in our comparative study for enhancing data fidelity (Supplementary Experiment 8). Our proposed OT method demonstrates consistent improvements over the raw generative output, reducing the Jensen-Shannon Divergence (JSD) for the GDP distribution from approximately 0.6 to 0.35 and increasing the Overlapping Ratio (OR) for infrastructure by a factor of five (Supplementary Experiment 8a). More critically, the calibration rectifies the systematic parametric biases found in the raw generative output. For instance, it corrects the Zipf's Law exponent from a biased value of approximately -0.5 to a value that aligns with the theoretical -1.0, and similarly adjusts the scaling exponents for infrastructure and GDP to match their empirical benchmarks (Supplementary Experiment 8b). The method's effectiveness is further validated by an ablation study, which shows that applying OT to uninformed noise (Noise + OT) yields only limited improvements, confirming the synergistic importance of the GenAI's sophisticated generative prior. Furthermore, the OT consistently outperforms the traditional statistical synthesizer (KDE), positioning it as a robust method for generating high-fidelity urban data.

2.5. The impact of different prompts and models

Given the observed deviations, this study investigated whether these gaps could be bridged through different forms of intervention. The results indicate that the fidelity of generated data possesses potential for improvement through both user interaction and model selection. Strategic prompting offers a viable pathway to enhance data fidelity. Prompting with specific geographic contexts demonstrates a clear effect on data distribution; when prompted with "China," the generated data showed a greater alignment with real data from Chinese cities, increasing the Overlap Ratio by 38.56% and decreasing the Mean Absolute Error by 52.14% (Fig. 5a, 5b). Similarly, at the meso-scale, prompting with parameters like city size or historical period directly influences the resulting decay curves (Fig. 2c). These results demonstrate that the model's output is sensitive to contextual cues, highlighting a clear potential for iterative refinement.

Furthermore, performance varies noticeably across the different GenAI models evaluated. In the urban scaling law test, GPT-4o demonstrated the closest alignment with empirical results, while Claude 3.5 was most accurate in reproducing Zipf's law. For distance decay, GPT-4o again showed more nuanced spatial reasoning, whereas models like ChatGLM-4 and DeepSeek-V3 tended to produce less diverse outputs. In the urban vitality assessment, GPT-4o consistently outperformed all other models. These variations suggest that factors intrinsic to the models, such as architecture, training data, and fine-tuning strategies, significantly influence the characteristics and fidelity of the generated urban data.

3. Discussion

This study reveals a fundamental duality in the capabilities of current GenAI models within urban science. They excel at reproducing the abstract, mathematical forms of established statistical theories but falter in capturing the complexity of real-world urban systems. This pattern of oversimplification extends directly to the visual domain. The systematic evaluation across different scales demonstrates that while GenAI models can generate data that is structurally "form-similar" to urban theories, this apparent competence masks significant deficiencies. In the statistical realm, this manifests as a lack of diversity and systematically distorted parameters (e.g., underestimated scaling exponents). In the visual realm, it appears as a bias against perceptual ambiguity (agreeing with human 'equal' judgments in only 2.1% of cases) and a failure to capture the full variance of global urban environments.

The initial finding that GenAI models consistently generate ”mirage cities” with low diversity and systematically distorted parameters (e.g., underestimated scaling exponents) points to a core limitation. This phenomenon, where models produce archetypal representations of cities, a tendency sometimes described as ”group flattening” (Bettencourt et al., 2007), aligns with observations of geographical and social bias in other domains (Manvi et al., 2024; Abbasi et al., 2025). This interpretation suggests that these distortions are symptomatic of the GenAI's nature as statistical pattern matchers rather than causal world models. Lacking a deep, generative understanding of the mechanisms that produce urban phenomena, they instead reproduce the most dominant statistical patterns from their training data, smoothing over the granular heterogeneity and outliers that are often crucial drivers of real world dynamics. This tendency, also noted in other applications of GenAI models for generating synthetic populations (Li et al., 2025), poses a significant risk for empirical research that relies on capturing the full spectrum of urban complexity.

The investigation into data fidelity reveals that GenAI models are malleable tools sensitive to user interaction and model specific characteristics. Our finding that prompt engineering can partially mitigate deviations by steering the model toward a specific context underscores the importance of the human AI interface in scientific applications (Wang et al., 2024; Li et al., 2025; Hou et al., 2025). The observed performance differences across the five evaluated GenAI models suggests that future improvements in data fidelity will depend on both more sophisticated interaction protocols and the rapid advancements in the models themselves (Balsa-Barreiro et al., 2024). Beyond these interactive refinements, our successful implementation of a post-hoc calibration procedure using Optimal Transport offers a crucial insight: the raw output of an GenAI model should be seen as a structurally rich yet uncalibrated starting point. This hybrid approach effectively overcomes the limitations of relying solely on the model's internal world representation, demonstrating a practical path toward creating high-fidelity synthetic datasets.

A distinct capability of GenAI models is revealed when they are tasked with navigating less formalized, qualitative theories like urban vitality. Here, the GenAI is prompted to simulate a ”social consensus” by synthesizing the vast amount of information about cities embedded in its training data, from academic texts to public discourse. The result is the generation of a plausible, structured set of relationships that can provide urban scientists with novel, testable hypotheses. This aligns with the growing recognition of GenAI models as tools for urban sensing and knowledge synthesis. In this role, the GenAI acts as a powerful catalyst for human led scientific inquiry, offering a new method to ”operationalize” qualitative knowledge into a quantitative, empirical research program.

The GenAI models also unveil a novel scientific opportunity. Their tendency to generate data by guidance of prompts with controlling variables in experiments, makes them powerful instruments for theoretical exploration (Ye et al., 2025; Wang et al., 2024; Li et al., 2025; Manvi et al., 2024; Abbasi et al., 2025). Urban scientists can now generate vast, fine grained datasets under specific theoretical assumptions, creating virtual laboratories to explore a theory's logical consequences or test its sensitivity to certain parameters. This moves beyond simply validating models against existing data and toward a new mode of inquiry focused on understanding the internal consistency and emergent properties of theories themselves. This approach directly addresses the challenge of how AI can generate new knowledge for urban studies, by providing a scalable method to rigorously interrogate theoretical constructs in a controlled environment. However, while our structured prompting strategy enhanced the diversity of the generated outputs, a fundamental limitation persists: the current methodology is incapable of synthesizing a truly de novo virtual city and is instead constrained to generating data corresponding to random, extant locations on Earth.

While GenAI models can be prompted to generate data for hypothetical or counterfactual scenarios, interpreting these outputs as scientifically valid predictions is challenging. For example, correcting known biases in generated data to test unknown scenarios where data cannot be obtained in the real world. The primary obstacle is the ”black box” nature of the models. We cannot audit their internal reasoning process. Furthermore, the systematic biases we measure in one context, such as under baseline conditions, are not guaranteed to be stable when the model is prompted with a novel, out-of-distribution scenario. Therefore, these counterfactual generations are most valuable when used to produce structured and testable hypotheses that guide subsequent empirical research.

While this study provides a foundational framework, its scope defines a clear agenda for future research. Continuous benchmarking is essential, as the quantitative performance of GenAI models will evolve with new architectures (Balsa-Barreiro et al., 2024). More importantly, our findings underscore a critical distinction. The systematic biases and lack of causal depth suggest that current GenAI models, while powerful, are not true world models. They are adept at learning the ”what” from data but struggle with the ”why”. The path toward transformative breakthroughs in urban science, particularly in areas like causal inference and policy simulation (Xia et al., 2025), will likely require a new generation of hybrid models. These findings position GenAI models as powerful engines for synthesizing urban scientific data and exploring the socio-cultural patterns for scientific inquiry. Future work should focus on integrating the rich, contextual knowledge of GenAI models with the causal integrity of traditional urban simulation frameworks, a synergy that could enable new avenues for a robust and predictive urban science.

4. Method

The intelligence of GenAI models essentially derives from their ability to compress world knowledge acquired during training (Delétang et al., 2024; **?**). To systematically evaluate the capabilities of GenAI models in urban science, this study employs a framework that tests their generative fidelity against two core domains of urban theory: (1) statistical laws based on symbolic data (different dimensions with scale, space, and morphology), and (2) cognitive laws based on perceptual data (such as street-view images in urban scenes).

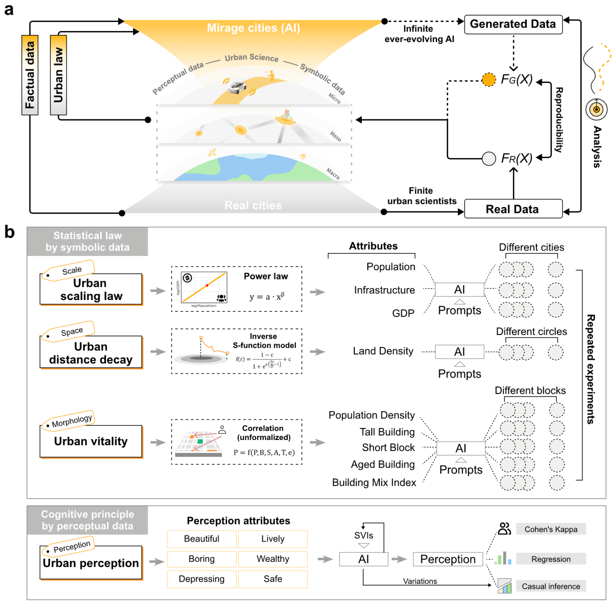


Figure 6. AI4US, the proposed framework utilizing GenAI models to advance urban science.

4.1. Urban theories based on generating symbolic data

For the first domain, we selected three foundational theories designed to test the GenAI's abilities across the distinct yet interconnected scales at which urban phenomena are understood: the selection of these theories is deliberate, designed to test the GenAI's generative fidelity across the distinct yet interconnected scales at which urban phenomena are understood: the macro-level system of cities, the meso-level internal structure of a single city, and the micro-level dynamics of a neighborhood (Supplementary Note 2). Substantial evidence suggests that, similar to regularities observed in natural complex systems, urban phenomena exhibit patterns and order across these multiple dimensions of size, space, and morphology (Batty and Longley, 1994; Bettencourt, 2013; Barthelemy, 2019; Arcaute, 2020). The chosen theories are canonical representations of these scaled phenomena (Table S1). Urban scaling laws capture nonlinear regularities at the macro-scale, describing how infrastructure and socioeconomic activity grow with city size. Distance decay theory reflects meso-scale spatial interaction friction and geographic reach, explaining the internal organization of cities. Finally, Jane Jacobs' theory of urban vitality articulates how fine-grained, micro-scale morphological configurations contribute to emergent urban dynamics. Together, these theories exemplify foundational mechanisms in urban systems. They provide a comprehensive and representative testbed for evaluating the core capabilities of GenAI models: from reproducing mathematically explicit laws to simulating the socio-cultural knowledge embedded in qualitative theory.

4.1.1. Urban Scaling Law

As a foundational concept in complex systems, urban scaling laws describe the power-law relationship between a city's size and its various metrics, serving as functional models for understanding cities as complex systems. These laws are broadly categorized into inter-urban processes (interactions between cities) (Ribeiro and Rybski, 2021), such as Zipf's Law, and intra-urban processes (dynamics within a city) that explore how human interactions and spatial geometry contribute to scaling. Zipf's Law posits an inverse relationship between a city's population and its rank.

$$\log_{10}(\text{Population}) = \log_{10} a + \beta \log_{10}(\text{Rank}) \qquad (1)$$

The intra-urban scaling law relates socioeconomic or infrastructural metrics y to city population.

$$\log_{10}(y) = \log_{10} a + \beta \log_{10}(\text{Population}) \qquad (2)$$

$\log_{10}(a)$ represents the intercept, while the slope β is the critical scaling exponent. Empirical evidence (Bettencourt et al., 2007) suggests two distinct patterns for intra-urban scaling: socioeconomic activities (e.g., GDP) exhibit superlinear scaling ($\beta > 1$), implying increasing returns to scale, while infrastructure metrics show sub-linear scaling ($\beta < 1$), indicating greater efficiency in larger cities. To test these laws, GenAI models were prompted to generate a dataset of 100 cities with attributes for population, infrastructure, and GDP.

4.1.2. Urban Distance Decay

This theory addresses the meso-scale structure of cities, positing that the intensity of interaction between two locations diminishes with increasing distance, a concept closely related to Tobler's First Law of Geography (Pun-Cheng, 2016; Tobler, 1970). This typically results in a pronounced "center-periphery" pattern in urban spatial structures, where attributes like population density and land value decrease from the city core. While several quantitative models exist, our study focused on the inverse S-function model for characterizing the spatial decay of urban land density $f(r)$ as a function of distance (r) from the city center (Rodrigue, 1998; Batty and Kim, 1992; Jiao, 2015).

$$f(r) = \frac{1-c}{1+e^{a\left(\frac{2r}{D}-1\right)}} + c \tag{3}$$

The parameters define the curve's shape: c represents the minimum land density in the urban periphery (the lower asymptote); a is the growth rate, controlling the steepness of the density decay; and D is a scaling parameter related to the city's spatial extent, where the inflection point of the curve occurs at the distance $r = D/2$. To assess the models' grasp of this concept, GenAI models were prompted to generate land density values for 100 concentric urban zones.

4.1.3. Jane Jacobs' Urban Vitality

At the micro-scale, we tested the GenAI's ability to operationalize Jane Jacobs' classic theory of urban vitality (Jacobs, 1961). The theory posits that vitality depends on four key elements: concentration, functional diversity, contact opportunity (short blocks), and aged buildings, where population density and building mix are considered core driving forces (Glaeser, 2010). Despite its wide acceptance, empirical tests of the theory remain limited and inconsistent (Huang et al., 2023), and debates persist around certain elements, such as the role of tall buildings.

$$P = f(P, B, S, A, T, e) \tag{4}$$

Here, vitality is a function of Population Density (P), Building Mix Index (B), Short Block (S), Aged Building (A), and Tall Building (T). We employed a two-step process: first, GenAI models were prompted to generate data for 100 urban blocks across these five key attributes; second, the models were prompted to act as a "resident" and assign a "livability score" to each block as a proxy for vitality, allowing us to test whether the models' inferred relationships were consistent with Jacobs' core tenets.

4.2. Urban perception through generating perceptual data

For the second domain, we evaluated the GenAI's multimodal capabilities in urban visual perception. We focused on human-scale street view imagery (SVI), a mature benchmark representing the daily urban life. The evaluation used a "Generate-Evaluate-Calibrate" paradigm grounded in the six perceptual indicators from the MIT Place Pulse project (Naik et al., 2014; 2016; Dubey et al., 2016; De Nadai et al., 2016; Naik et al., 2017; Zhang et al., 2018) (Table S2). Our methodology involved a three-phase process to validate the AI as a visual data generator, a perceptual judge, and a tool for causal inference.

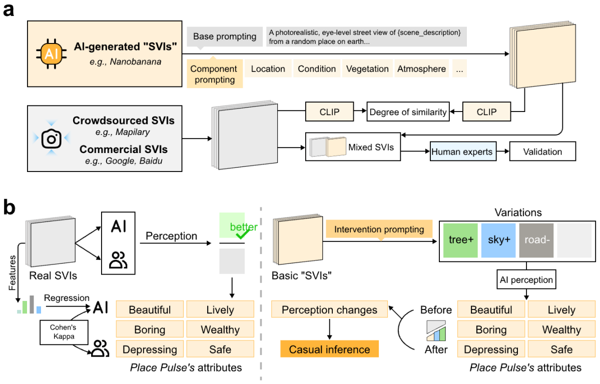


Figure 7. A computational framework for advancing urban perception through perceptual data. Edit text and images through Nano Banana.

First, to validate the AI's data generation capability, the generated human-scale streetscapes were assessed for realism and diversity. Realism was evaluated by human experts, who rated 100 images from AI generation and 100 images from Place Pulse on a 1-10 scale (Supplementary Experiment 9). The model's fidelity was then quantified by comparing the score distributions against real photographs using an independent samples t-test. Dataset diversity was ensured via a structured-prompting strategy. The effectiveness of this strategy was then quantitatively validated by using CLIP embeddings to compare the average semantic distance D_{CLIP} across three distinct image sets: a real-world street view benchmark from Place Pulse, a baseline set generated from simple prompts, and our proposed structured-prompt set.

$$D_{CLIP} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sqrt{1 - \frac{e_i \cdot e_j}{\|e_i\| \|e_j\|}} \tag{5}$$

Second, to validate the AI's perceptual judgment, its pairwise choices on original Place Pulse images were compared against the source dataset's human choices using Kappa.

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{6}$$

Furthermore, to assess if the AI could identify influential visual factors in alignment with human patterns (Zhang et al., 2018), we fit a multiple linear regression model correlating the AI's perception scores (Y_p) with the pixel proportions of 150 visual elements (X_k) from semantic segmentation.

$$Y_p = \beta_{p,0} + \sum_{k=1}^{K} \beta_{p,k} X_k + \varepsilon_p \tag{7}$$

Third, to validate the AI's utility for causal inference, we performed thematic interventions on baseline images to create "before-and-after" pairs. These interventions were designed around three interpretable categories: Natural, Traffic, and Built Elements. The causal effect of each category was then quantified by calculating the average change in perception scores ($\Delta S_{p,c}$) between the variant and baseline images for each perception type.

$$\Delta S_{p,c} = \frac{1}{N_c} \sum_{i=1}^{N_c} (S_{p,i}^{\text{variant}} - S_{p,i}^{\text{baseline}}) \tag{8}$$

Where $\Delta S_{p,c}$ is the average change in score for perception p under category C, $S_{p,i}^{\text{variant}}$ and $S_{p,i}^{\text{baseline}}$ are the scores for a single image pair i, and N_c is the total number of image pairs in that category.

4.3. Comparison of data distribution

First, this study directly compares the data distributions of generated and real-world samples. Since generated values without geographic prompts often differ from real samples by orders of magnitude, this study applies equal-width binning and relative frequency normalization to eliminate differences in data scale, converting the raw sample counts within each bin into frequencies. The Mean Absolute Error (MAE) of bin-level frequencies is then used to measure the average difference in relative frequency between real and generated data across bins, serving as an indicator of how well the generated data approximates the overall distribution trend of the real data. This is calculated by summing the absolute differences of mean frequencies across all bins, thereby capturing the degree of deviation in central tendency between the two distributions. For each bin, the average frequency is computed separately for the real data (R) and the generated data (G). N is the number of bins, $\mu_{R,i}$ is the average relative frequency of R in the ith bin, $\mu_{G,i}$ is the average relative frequency of G in the i th bin.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\mu_{R,i} - \mu_{G,i}| \tag{9}$$

Measure the overlapping degree of quantile intervals between real data and generated data in each bin, and evaluate whether the fluctuation range of generated data covers the distribution interval of real data. The overlapping ratio calculates the ratio of the interval intersection of the two to the union, and divides the sum of the intersection areas of all bins by the sum of the union areas.

$$\text{Overlap Ratio} = \frac{\sum_{i=1}^{N} \text{Intersection}(Q_{R,i}, Q_{G,i})}{\sum_{i=1}^{N} \text{Union}(Q_{R,i}, Q_{G,i})} \tag{10}$$

Secondly, based on the urban data generated by GenAI models and the real data of Chinese cities, this study measures the pairing difference between the two data and three discrete distributions by Jensen-Shannon Divergence (JSD). Include 'distance between real sample and real sample data distribution' (R-R), 'distance between real sample and generated sample data distribution' (R-G), 'Distance between generated sample and generated sample data distribution'(G-G). Taking the calculation of average divergence of R-G as an example, there are n true distributions R_i and m generated distributions G_j.

$$\text{Distance}_{R-G} = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{JSD}(R_i, G_j) \tag{11}$$

$$\text{JSD}(R, G) = \frac{1}{2}\text{KL}(R \| M) + \frac{1}{2}\text{KL}(G \| M) \tag{12}$$

$$\text{where } M = \frac{1}{2}(R + G)$$

$$\text{KL}(R \| M) = \sum_{x} R(x) \log \frac{R(x)}{M(x)} \tag{13}$$

4.4. Post-hoc calibration procedure for data distribution

To address the systematic bias and lack of diversity observed in raw generated data, we introduce a post-hoc calibration procedure (detailed explanations refer to Supplementary Note 3). This challenge of aligning a generated ('source') distribution with an empirical ('target') distribution is a fundamental problem in machine learning, analogous to domain adaptation (Courty et al., 2016; Fatras et al., 2021) or batch effect correction (Bunne et al., 2024). While prior work in generated data synthesis has focused on prompt engineering or simple rule-based adjustments (Li et al., 2024b; 2025; Manvi et al., 2024), these methods may not correct for systemic distributional shifts. We therefore employ an approach grounded in Optimal Transport (OT) theory (Courty et al., 2016; Bonet et al., 2025) for its ability to transform an entire dataset while preserving its core structure. The procedure learns an optimal transport map T that transports samples from the preprocessed generated distribution (P_G) to the preprocessed real data distribution (P_{real}). This map is then applied to the full generated dataset (x_i) to produce a calibrated set ($D_{\text{calib, scaled}}$):

$$D_{\text{calib, scaled}} = \{T(x_i) | x_i \in D_{G,\text{scaled}}\} \tag{14}$$

Here, x_i is the set of generated data points, and T is the learned mapping function. The distance between distributions is quantified using the Wasserstein distance (W_1), which represents the minimum "cost" required to transform one distribution into another:

$$W_1(P_{real}, P_G) = \inf_{\gamma \in \Pi(P_{real}, P_G)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| d\gamma(x, y) \tag{15}$$

Where $\Pi(P_{real}, P_G)$ is the set of all possible joint distributions of the source P_G and target (P_{real}) distributions. The efficacy of this OT approach was validated against several baselines, including raw GenAI output, random noise, and data synthesized via KDE. This comprehensive correction is achieved because the OT procedure learns a mapping that transforms the entire joint distribution of the multi-dimensional data. By adjusting the data structure holistically, rather than treating each variable independently, it inherently corrects the inter-variable correlations that define the scaling exponents.

Competing Interests

All authors declare no financial or non-financial competing interests.

Data Availability

We have provided information on all publicly available data used in our analysis in the Methods. Please refer to Table S3 for all prompts in this study.

Code Availability

The code to replicate the results in the paper will be available via Figshare.

References

P. Geddes. *Cities in Evolution: An Introduction to the Town Planning Movement and to the Study of Civics*. Williams, London, 1915.

R. Van Noorden and J. M. Perkel. AI and science: what 1,600 researchers think. *Nature*, 621(7980):672–675, 2023. doi: 10.1038/d41586-023-02980-0.

A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023. doi: 10.1038/s42254-023-00581-1.

M. Binz et al. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122:e2401227121, 2025. doi: 10.1073/pnas.2401227121.

Y. Long and E. Zhang. City laboratory: Embracing new data, new elements, and new pathways to invent new cities. *Environment and Planning B: Urban Analytics and City Science*, 51:1068–1072, 2024.

Y. Zhang, H. Zhao, and Y. Long. CMAB: A Multi-Attribute Building Dataset of China. *Scientific Data*, 12:430, 2025a. doi: 10.1038/s41597-025-04730-5.

R. Kitchin. The ethics of smart cities and urban science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374: 20160115, 2016.

A. Sudmant, F. Creutzig, and Z. Mi. Replicate and generalize to make urban research coherent. *International Journal of Urban Sciences*, 2024.

M. Batty. 50 years and going strong: Into the next half century. *Environment and Planning B: Urban Analytics and City Science*, 2024.

A. Wiig. Is urban theory changing big data? https://medium.com/global-infrastructures/is-urban-theory-changing-big-data-dc2120e5e263, 2018. Accessed May 18, 2025.

W. Li, Y. Zhang, M. Li, and Y. Long. Rethinking the country-level percentage of population residing in urban area with a global harmonized urban definition. *iScience*, 27:110125, 2024a.

L. Apostolos. Urban growth simulation through cellular automata: A way to explore the fractal nature of cities. *SCIENZE REGIONALI*, 9:5–23, 2010.

V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013.

M. Batty. A generic framework for computational spatial modelling. In A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, editors, *Agent-Based Models of Geographical Systems*, pages 19–50. Springer, 2012.

R. Egger and J. Yu. Epistemological challenges. In R. Egger, editor, *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, pages 17–34. Springer, 2022.

M. Wegener. Land-use transport interaction models. In M. M. Fischer and P. Nijkamp, editors, *Handbook of Regional Science*, pages 229–246. Springer, 2021.

Y. Long and Z. Shen. Population spatialization and synthesis with open data. In Y. Long and Z. Shen, editors, *Geospatial Analysis to Support Urban Planning in Beijing*, pages 115–131. Springer, 2015.

Y. Zhang, T. Tu, and Y. Long. Inferring ghost cities on the globe in newly developed urban areas based on urban vitality with multi-source data. *Habitat International*, 158:103350, 2025b.

J. Gottweis et al. Towards an ai co-scientist, 2025.

J. Balsa-Barreiro, M. Cebrián, M. Menéndez, and K. Axhausen. Leveraging generative ai models in urban science. In T. Paus, J. R. Brook, K. Keyes, and Z. Pausova, editors, *Principles and Advances in Population Neuroscience*, pages 239–275. Springer, 2024.

J. Chen et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks, 2024.

P. Bhandari, A. Anastasopoulos, and D. Pfoser. Urban mobility assessment using llms. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 67–79. Association for Computing Machinery, 2024.

J. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

C. Zeni et al. Mattergen: a generative model for inorganic materials design, 2024.

G. Sun, E. Y. Choe, and C. Webster. Natural experiments in healthy cities research: how can urban planning and design knowledge reinforce the causal inference? *Town Planning Review*, 94:87–108, 2023.

X. Li et al. Be more real: Travel diary generation using llm agents and individual profiles, 2024b.

Z. Zhou, Y. Lin, D. Jin, and Y. Li. Large language model for participatory urban planning, 2024.

X. Ye et al. Artificial intelligence in urban science: why does it matter? *Annals of GIS*, 31:181–189, 2025.

J. Wang et al. Large language models as urban residents: an llm agent framework for personal mobility generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 124547–124574, 2024.

H. Li, J. Ding, J. Gong, and Y. Li. Large language model as user daily behavior data generator: balancing population diversity and individual personality, 2025.

R. Manvi, S. Khanna, M. Burke, D. Lobell, and S. Ermon. Large language models are geographically biased, 2024.

O. R. Abbasi, F. Welscher, G. Weinberger, and J. Scholz. The world as large language models see it: Exploring the reliability of llms in representing geographical features, 2025.

X. Tan et al. The spatiotemporal scaling laws of urban population dynamics. *Nature Communications*, 16:2881, 2025.

F. Xu, Y. Li, D. Jin, J. Lu, and C. Song. Emergence of urban growth patterns from human mobility behavior. *Nature Computational Science*, 1:791–800, 2021.

J. Jacobs. *The Death and Life of Great American Cities*. Random House, 1961.

M. Batty and P. Longley. *Fractal Cities: A Geometry of Form and Function*. Academic Press, London; San Diego, 1994.

N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 779–785, 2014.

L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104:7301–7306, 2007.

C. Hou et al. Urban sensing in the era of large language models. *The Innovation*, 6:100749, 2025.

Y. Xia et al. Reimagining urban science: Scaling causal inference with large language models, 2025.

G. Delétang et al. Language modeling is compression, 2024.

L. M. A. Bettencourt. The origins of scaling in cities. *Science*, 2013.

M. Barthelemy. The statistical physics of cities. *Nature Reviews Physics*, 1:406–415, 2019.

E. Arcaute. Hierarchies defined through human mobility. *Nature*, 587:372–373, 2020.

F. L. Ribeiro and D. Rybski. Mathematical models to explain the origin of urban scaling laws: a synthetic review, 2021.

L. S. C. Pun-Cheng. Distance decay. In *International Encyclopedia of Geography*, pages 1–5. John Wiley & Sons, Ltd, 2016.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240, 1970.

J.-P. Rodrigue. *The Geography of Transport Systems: Levels of Integration*. Hofstra University, 1998.

M. Batty and K. S. Kim. Form follows function: Reformulating urban population density functions. *Urban Studies*, 1992.

L. Jiao. Urban land density function: A new method to characterize urban expansion. *Landscape and Urban Planning*, 139:26–39, 2015.

E. L. Glaeser. Preservation follies. *City Journal*, 2010.

J. Huang, Y. Cui, L. Li, C. Webster, et al. Re-examining Jane Jacobs' doctrine using new urban data in Hong Kong. *Environment and Planning B: Urban Analytics and City Science*, 50(1):76–93, 2023.

N. Naik, R. Raskar, and C. A. Hidalgo. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *American Economic Review*, 106(5):128–132, 2016.

A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision*, pages 196–212. Springer International Publishing, 2016.

M. De Nadai, R. L. Vieriu, G. Zen, S. Dragicevic, N. Naik, M. Caraviello, et al. Are safer looking neighborhoods more lively? a multimodal investigation into urban life. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1127–1135, 2016.

N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160, 2018.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

K. Fatras, T. Séjourné, R. Flamary, and N. Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International conference on machine learning*, pages 3186–3197. PMLR, 2021.

C. Bunne, G. Schiebinger, A. Krause, A. Regev, and M. Cuturi. Optimal transport for single-cell and spatial omics. *Nature Reviews Methods Primers*, 4(1):33, 2024.

C. Bonet, C. Vauthier, and A. Korba. Flowing datasets with wasserstein over wasserstein gradient flows, 2025.

K. Lynch. *The image of the city*. MIT press, 1964.

Supplementary Information

Supplementary Note 1. Models selection

For model selection, we evaluated top MEGA-Bench (Chen et al., 2024) models and conducted experiments to validate their real-world stability and applicability. While some models excelled in mathematical reasoning and code generation, they proved unsuitable for urban science data generation due to shortcomings in output completeness, stability, and controllability. For example:

- **Doubao-Pro-128k:** The model produced incomplete outputs and failed to strictly follow the required prompt structure during experiments.
- **Hunyuan-Large-LongContext:** The model produced incomplete outputs and exhibited high repetition across multiple runs, preventing diverse data generation. Additionally, it imposed usage frequency restrictions, impacting the reproducibility of batch experiments.
- **Qwen-Plus-Latest:** The model produced incomplete outputs and failed to generate the required content in full, leading to missing information and reducing its suitability for rigorous data validation tasks.

ChatGPT, Claude, Gemini, DeepSeek, and ChatGLM demonstrated superior data generation completeness, theoretical alignment, and experimental reproducibility. Future research may further explore ways to optimize the application of other models to better meet the demands of urban science studies.

Supplementary Note 2. Detailed explanations for selected theories

Urban scaling law

Scaling laws are a fundamental concept in complex systems, describing how certain quantities change with system size, typically following a power-law relationship. In the context of urban systems, urban scaling laws serve as functional models for understanding cities as complex systems. According to Ribeiro and Rybski (Ribeiro and Rybski, 2021), these scaling laws can be categorized into two broad types: inter-urban processes (models describing interactions between cities) and intra-urban processes (models describing dynamics within a single city). Inter-urban processes focus on the exchange of information between cities as a mechanism for explaining the emergence of scaling laws. Key theoretical explanations for inter-urban scaling include Zipf's Law, hierarchical organization, and interactions between people in different cities. In contrast, intra-urban processes examine scaling laws based solely on factors within a city. These models explore how human interactions, urban geometry, and spatial distribution contribute to urban scaling. Within this category, gravity models represent a generalized formalization of intra-urban scaling dynamics.

Empirical studies on inter-urban processes have shown that city rank (Y) and city population size (N) exhibit a scaling relationship (Ribeiro and Rybski, 2021). For intra-urban processes, empirical evidence suggests two distinct scaling laws: (1) variables related to socioeconomic activity (e.g., GDP, patents, HIV cases) exhibit super-linear scaling ($\beta > 1$), meaning that as a city's population increases, these socioeconomic indicators grow at a faster rate; and (2) variables related to infrastructure (e.g., the number of charging stations or gas stations) exhibit sub-linear scaling ($\beta < 1$), indicating that as the population increases, these infrastructure-related variables grow at a slower rate.

Urban distance decay

The term distance decay is used to describe the diminishing intensity of interaction between two locations as the distance between them increases. It refers to the phenomenon where the strength of connections—especially in economic activities—tends to weaken with greater spatial separation (Pun-Cheng, 2016). At its core, the distance decay law posits that interactions between geographic entities are inversely related to distance; under otherwise equal conditions, such interactions typically follow an inverse-square relationship with distance. As a concept that reflects the decline of influence over space, distance decay is closely related to Tobler's First Law of Geography (Tobler, 1970): 'Everything is related to everything else, but near things are more related than distant things'. In urban geography, cities and towns often have clearly defined centers, and most urban attributes—such as population density, land value, and the quality or quantity of services—tend to decrease with increasing distance from the center (Pun-Cheng, 2016). Due to spatial agglomeration, urban spatial structures frequently exhibit a pronounced 'center-periphery' pattern, where the intensity of urban elements decays with increasing distance from the city core. The most typical manifestation of this is the spatial decay of population density. Common quantitative models include the negative exponential model (Rodrigue, 1998) and the inverse power model (Batty and Kim, 1992). For urban land use density, the inverse S-function model has also been proposed to characterize its spatial decay (Jiao, 2015).

Jane Jacobs' urban vitality

Urban vitality theory is a classic urban theory put forward by Jacobs (Jacobs, 1961). She pointed out that urban vitality often depends on four key elements: concentration (population density), functional diversity (architectural function mixture), contact opportunity (short block), and need for aged buildings (proportion of old buildings). These factors together constitute the foundation of urban vitality, among which population density and architectural function mixing are the core driving forces, while short blocks and old buildings enhance urban vitality by increasing contact opportunities and reducing operating costs. However, in recent years, there have been disputes about Jacobs' views on high-rise buildings, and some studies believe that it has a positive impact on urban vitality (Glaeser, 2010). Although Jacobs' theory is widely accepted, empirical tests remain limited and inconsistent. The reasons for uncertainty are the robustness of the research design and the interpretability of the theory to cross-geographical areas.

Urban perception through images

The AI4US framework extends its evaluation beyond statistical regularities to the domain of urban visual perception. Urban science is concerned not only with emergent patterns but also with the human-centric experience of the built environment, which is profoundly shaped by its visual characteristics. Understanding the link between physical appearance and human psychological responses, such as feelings of safety or aesthetic pleasure, is fundamental to urban design and planning.

The study of urban perception has evolved significantly over the past half-century. It originated from foundational, qualitative approaches, such as Kevin Lynch's work on the urban image (Lynch, 1964), which relied on manual surveys and expert analysis to decode how people form mental maps of their cities. The subsequent integration of photography and GIS allowed researchers to begin spatializing these subjective perceptions. A paradigm shift occurred with the rise of computational methods and large-scale crowdsourcing, exemplified by the MIT Place Pulse project (Naik et al., 2014; 2016; Dubey et al., 2016; De Nadai et al., 2016; Naik et al., 2017). This initiative collected millions of human judgments on street view imagery (SVI), creating a global dataset that links images to perceptual scores. This enabled the next stage: using machine learning to automate perception analysis at an unprecedented scale, and leveraging Explainable AI (XAI) to quantify which visual elements correlate with specific perceptions (Zhang et al., 2018). The current era of generative multimodal models marks a new frontier, allowing for end-to-end experiments where scenes can be generated, programmatically modified, and re-evaluated to test causal hypotheses about perception.

Supplementary Note 3. Post-hoc calibration procedure for data distribution

To overcome the systematic bias and lack of diversity observed when using GenAI to synthesize data, it is necessary to calibrate the larger dataset generated by GenAI by utilizing a small portion of real empirical data. This challenge of aligning a generated ('source') distribution with an empirical ('target') distribution is a fundamental problem in machine learning (Courty et al., 2016). It is conceptually analogous to domain adaptation in computer vision, where synthetic data is transformed to match real-world image characteristics (Fatras et al., 2021), or to batch effect correction in bioinformatics, where data from different experiments are normalized to a common distribution (Bunne et al., 2024). In these contexts, the goal is to learn a mapping that corrects for distributional shifts while preserving the core data structure.

Within the specific context of GenAI-based data synthesis, prior research has largely focused on enhancing output fidelity through sophisticated prompt engineering or agent-based architectures (Li et al., 2025; Manvi et al., 2024). Post-processing, when applied, has typically involved direct, rule-based adjustments. These include enforcing strict output formats and value ranges to ensure data validity at the point of generation (Li et al., 2025), or subsequently grounding the GenAI's conceptual outputs (e.g., a "trip to the office") in physical space by mapping them to specific geographic coordinates based on real-world constraints like a city's road network (Li et al., 2024b). While these methods improve the plausibility of individual data points, they may not correct for systematic biases in the overall data distribution.

To address this need, we selected an approach grounded in Optimal Transport (OT) (Courty et al., 2016; Bonet et al., 2025) theory for its distinct advantages over simpler alternatives. Direct probability density estimation methods (e.g., Kernel Density Estimation) often struggle with the curse of dimensionality in multi-attribute urban data. Simpler filtering methods, such as classifier-based selection, can discard a significant portion of the generated samples and may fail to preserve the latent structure of the data. In contrast, OT offers a more powerful solution by transforming the entire generated dataset rather than merely selecting from it, providing a robust method for distributional alignment that is well-suited to this task.

Supplementary Experiment 1. Comparison of Model Ability in the Data Distribution

As evidenced in Figure S1, GPT-4o and DeepSeek-V3 demonstrated superior alignment with real-world data among the five evaluated models. Under the JSD metric, these two models exhibited aggregate mean divergence values of 0.115 (R-G) and 0.116 (R-GC), representing reductions of 41.9%, 23.2%, and 1% compared to Claude-3.5, Gemini-2.0, and ChatGLM-4, respectively. Real data (R) comprises statistical records of population, GDP, and road network length across 685 Chinese municipal administrative regions. Each iteration involved random sampling of

Table S1. A Multi-Scale Framework for Evaluating GenAI models in Urban Science.

Scale	Theory	Significance in Urban Science
Macro (City system)	Urban scaling law	Explains universal, size-based regularities of cities as complex systems, central to understanding urban growth and efficiency.
Meso (Internal city structure)	Urban distance decay	Represents the foundational role of spatial friction and accessibility in shaping internal urban structure and organization.
Micro (Block)	Urban vitality	Provides a human-centric perspective linking the micro-design of the built environment to social and economic outcomes; a cornerstone of urban design.

Table S2. Benchmark for Urban Perception Indicators based on Visual Elements. Summary of regression coefficients (β) for all unique top-10 visual objects influencing six urban perceptions. The significance of each coefficient is indicated by stars. Adapted from the analysis in Zhang et al. (2018) (Zhang et al., 2018). The top 10 objects that positively/negatively contributed to each of the 6 perception types are shown. Beta Coefficient (*p¡0.1,**p¡0.05,***p¡0.01).

Visual Object	Safe	Lively	Beautiful	Wealthy	Depressing	Boring
Bridge	-	-0.0144***	-0.0120**	-	0.0178***	0.0076
Building	-0.1163***	-0.0625***	-0.0689***	-	0.0483***	-
Car	0.0600***	0.0887***	-	0.0295***	-0.0243***	-0.0329***
Earth	-	-	-	-0.0174***	-	-
Fence	-0.0327***	-	-0.0201***	-	0.0198***	0.0144***
Field	-	-0.0543***	-	-0.0248***	-	0.0130**
Floor	-	-	-0.0172***	-0.0219***	-	-
Grass	0.0711***	-	0.0465***	0.0272***	-0.0204***	-
Hill	-	-	-	-	-	0.0105*
House	-	-	-	-	-0.0123**	-
Path	0.0154***	-	0.0126***	0.0115**	-	-
Pole	-	-	-	-	-	-0.008
Road	0.0395***	0.0228***	-	0.0122*	-	-
Signboard	-	-	-	-	-	0.0173***
Sidewalk	0.0821***	0.0596***	-	-	-0.0350***	-0.0320***
Sky	-0.2068***	-0.2389***	-0.1179***	-0.1142***	0.1063***	0.0969***
Tree	-0.0392***	-0.1055***	0.0980***	0.0244***	-0.0610***	-
Truck	-	-	-0.0125***	-	-	-
Wall	-0.0764***	-0.0828***	-0.0458***	-0.0264***	0.0389***	0.0382***
Water	-	-	0.0070*	-	-	-

100 cities, with this process repeated 100 times to ensure statistical robustness. Generated data (G) were generated by corresponding models, producing 100 city-level datasets per iteration over 100 cycles. The distinction between R-G origin and R-GC lies in the inclusion of the geographical cue 'China' in the latter's generation prompts. Pairwise Jensen-Shannon Divergence (JSD) analyses were conducted to quantify discrepancies among three datasets: real data (R), baseline generated data (G), and generated data with geographical prompt 'China' (GC).

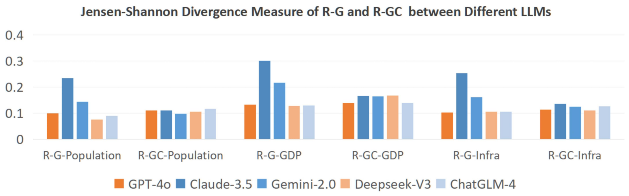


Figure S1. Comparative model capabilities based on R-G data difference.

Supplementary Experiment 2. Human Height Data

In this supplementary experiment, we evaluated the ability of GenAI models to generate human height data. Given that height distributions in the real world follow well-documented statistical patterns, such as normal distributions with known mean and standard deviation values for different populations, this task serves as a baseline assessment of whether generated data aligns with real-world human height distributions. We conducted ten independent trials for each GenAI model to generate male and female height datasets. The generated data were then compared against the global average height distributions for males (mean = 178.4 cm, SD = 5.6 cm) and females (mean = 164.7 cm, SD = 4.7 cm), as reported by Silverman (2022). Generated height distributions generally conformed to a normal pattern, though accuracy varied notably across models. Claude-3.5 consistently delivered the most realistic estimates, closely matching both the mean and standard deviation of true height data across genders. GPT-4o followed closely, offering stable performance with slightly conservative estimates. In contrast, DeepSeek-V3 exhibited the largest deviations, producing overly dispersed distributions that failed to capture real-world variance. Gender-based differences were also observed. For male height, GPT-4o achieved the best fit, with the highest R^2 (0.95) and near-identical distribution characteristics, while Claude-3.5 and ChatGLM-4 also performed well. For female height, Claude-3.5 was the most accurate overall, with the smallest deviation from the true mean and standard deviation, followed by ChatGLM-4, which demonstrated the best curve fit ($R^2 = 0.95$).

Supplementary Experiment 3. Individuals' Wealth Data

We evaluated GenAI's ability to generate individual-level wealth data and observed that the resulting distributions generally conformed to a Pareto pattern, capturing the right-skewed nature of wealth concentration. However, the estimated Pareto coefficients were consistently lower than the empirical benchmark of 1.5, as reported by Atkinson and Piketty, suggesting that GenAI models tend to understate the extent of real-world wealth inequality. Despite this, the log-log linear fits between rank and wealth yielded relatively high R^2 values, indicating that GenAI models can reproduce the structural form of the Pareto distribution. However, high R^2 does not imply accurate calibration of inequality levels, as the slope of the fitted line (i.e., the Pareto coefficient) remained biased toward more equal distributions. Among the tested models, Gemini-2.0 generated the most realistic distribution, producing both Pareto coefficients and tail be-

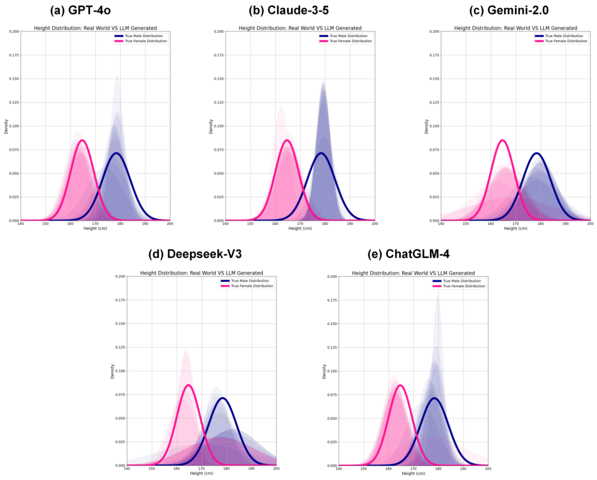


Figure S2. Main GenAI models tests: Human height distribution.

haviors closest to empirical values, thereby better reflecting the heavy-tailed nature of actual wealth distributions. Real-world wealth decile data used for comparison were obtained from the World Inequality Database.

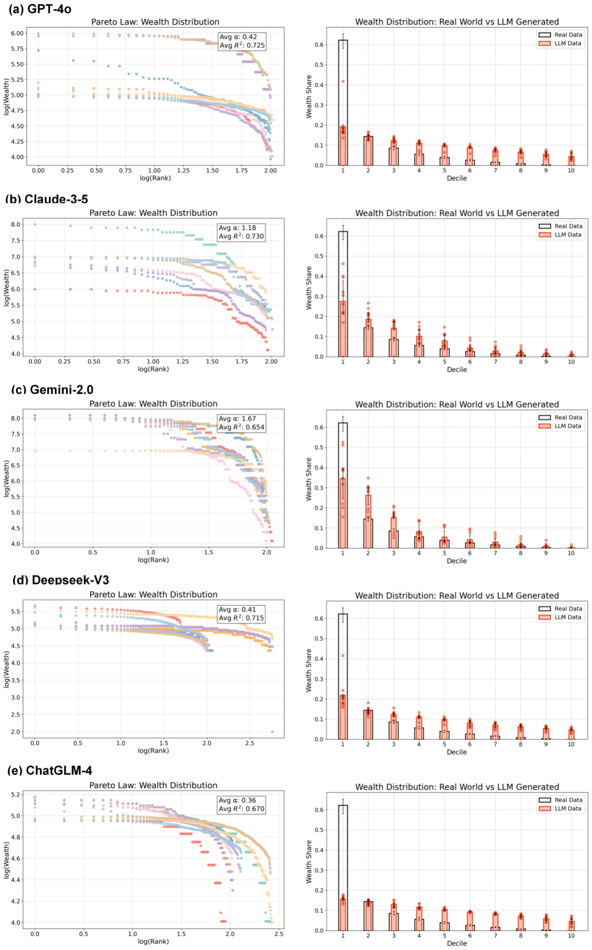
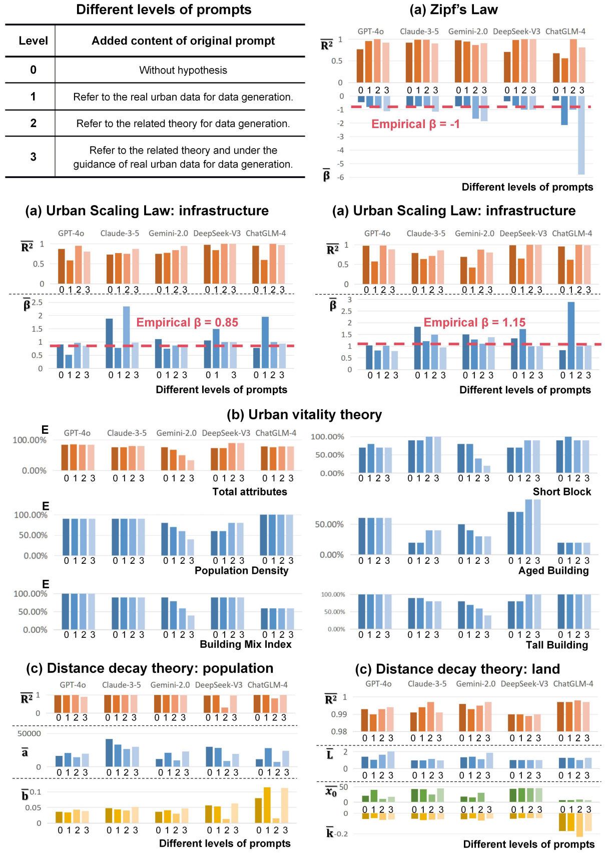


Figure S3. Main GenAI models tests: Individuals' wealth data distribution.



Different levels of prompts

Level	Added content of original prompt
0	Without hypothesis
1	Refer to the real urban data for data generation.
2	Refer to the related theory for data generation.
3	Refer to the related theory and under the guidance of real urban data for data generation.

Figure S4. Sensitivity test of different levels of prompts.

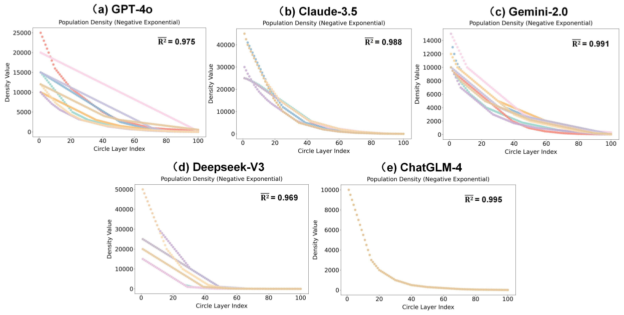


Figure S5. Main GenAI models tests of Distance decay theory: In terms of population density.

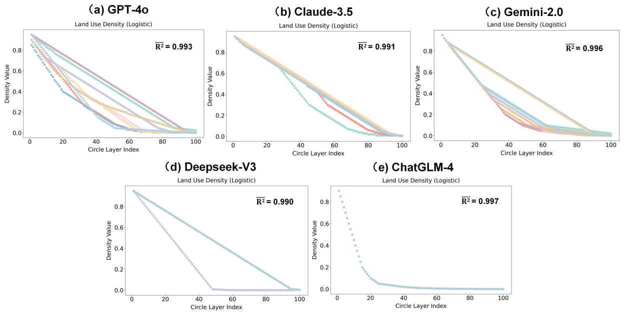


Figure S6. Main GenAI models tests of Distance decay theory: In terms of urban land density.

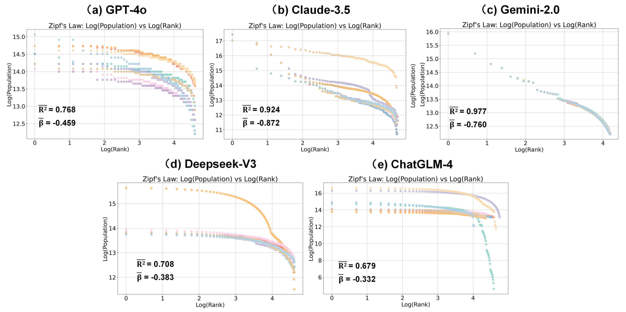


Figure S7. Main GenAI models tests of Zipf's Law between cities.

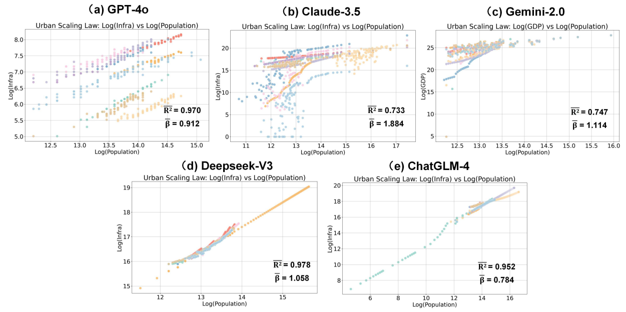


Figure S8. Main GenAI models tests of Urban Scaling Law in cities: infrastructure variables.

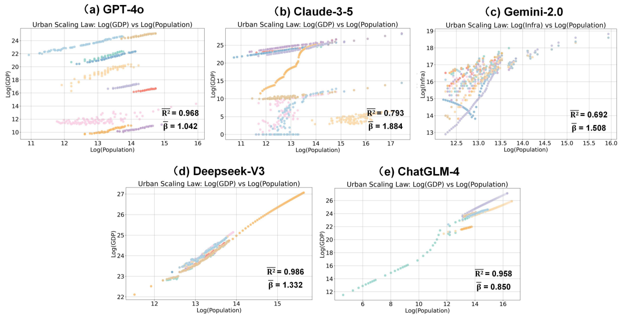


Figure S9. Main GenAI models tests of Urban Scaling Law in cities: socio-economic variables.

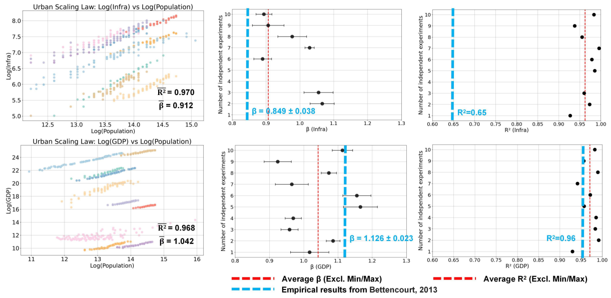


Figure S10. GPT-4o with the best performance of urban scaling law.

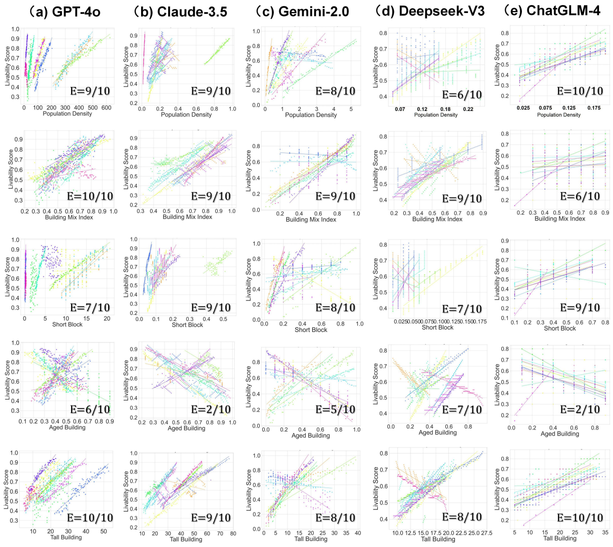


Figure S11. Main GenAI models tests of Jane Jacobs' urban vitality theory. E = effective significant positive slope, P ¡ 0.05 and slope ¿ 0.

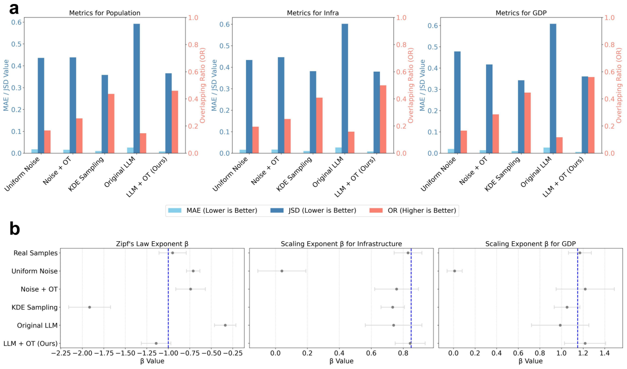
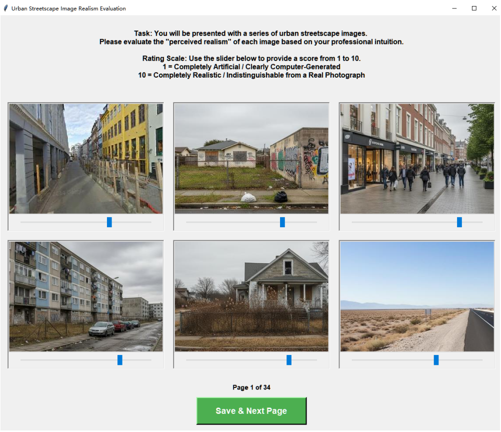


Figure S12. Enhancing the effectiveness of GenAI as a data synthesizer based on optimal transport. The top panels display the mean relative frequency distributions, while the bottom panels provide a quantitative evaluation using three key performance metrics. All results are aggregated from 100 independent experimental runs. In each run, a large synthetic dataset (1,000 data points) is generated by each method, calibrated against a new random sample of 100 real cities.


Urban Streetscape Image Realism Evaluation

Task: You will be presented with a series of urban streetscape images.
Please evaluate the "perceived realism" of each image based on your professional intuition.

Rating Scale: Use the slider below to provide a score from 1 to 10.
1 = Completely Artificial / Clearly Computer-Generated
10 = Completely Realistic / Indistinguishable from a Real Photograph

Page 1 of 34

Figure S13. Urban streetscape image realism evaluation.

Table S3. Prompts and specific descriptions in this study.

Prompt	**Description**
Urban scaling law	Generate the dataset containing 100 cities in a country. Format the output as: CityName1, Population1, Infrastructure volume1, GDP1 CityName2, Population2, Infrastructure volume2, GDP2 ... Output exactly 100 lines without any additional text or explanations. - Infrastructure volume: Total road miles - GDP: all Gross Domestic Product of the city in one year
Urban distance decay theory	Generate attributes for 100 concentric city circles arranged in order in a city. From the city center (Circle 1) to the outermost ring (Circle 100). Format the output as: Circle 1, Population Density1, Land Density 1 Circle 2, Population Density2, Land Density 2 ... Output exactly 100 lines without any additional text or explanations. - Population Density. - Land Density: Defined as the ratio of impervious surface area to the available land area.
Urban vitality theory – generation	Generate attributes for 100 blocks in a city. Format the output as: Block Name1, Population Density1, Building Mix Index1, Short Block1, Aged Building1, Tall Building1 ... Output exactly 100 lines without any additional text or explanations. - Population Density: People per square meter. - Building Mix Index: Entropy index based on building use. Value 0 to 1. - Short Block: Total number of street intersections divided by the area of the block. - Aged Building: The rate of old buildings on each block. Value 0 to 1. - Tall Building: The average building heights on each block.
Urban vitality theory – evaluation	You are a resident of a city. Assign a livability score between 0 and 1 for each block, 0 represents the least livable and 1 represents the most livable. Evaluate each block based on the following attributes: - Block name. - Population Density: People per square meter. - Building Mix Index: Entropy index based on building use. - Short Block: Total number of street intersections of the block. - Aged Building: The rate of old buildings on each block. - Tall Building: The average building heights on each block.
Streetscape Generation (Phase 1 Base Template)	”A photorealistic, eye-level street view of {scene_description} from a random place on earth. The screen should be completely filled with the generated image, without white borders. Try to approach the real scene captured by the car (A specialized vehicle equipped with professional camera equipment for capturing street view images) as closely as possible, without any retouching or beautification. The camera perspective is a forward-facing view, looking straight down the road. It should be a head-up view without any perspective distortion.”
Causal Intervention (Phase 3 Base Template)	Realistically edit the image to {intervention_description}. Keep everything else in the image the same.