nature cities

Review article

https://doi.org/10.1038/s44284-026-00492-2

Generative AI in urban science and practice

Received: 3 July 2025

Accepted: 22 June 2026

Published online: 5 August 2026

Check for updates

Yunke Zhang [1,16], Fengli Xu [1,16], Qi R. Wang [2], Esteban Moro [3,4], Yang Yue [5], Huandong Wang [1], Bin Chen [6,7], Yu Liu [8], Dongping Fang [9], Peng Gong [6,10,11], Luís M. A. Bettencourt [12,13,14], Michael Batty [15] & Yong Li [1]

Cities are changing faster than the tools used to understand them. A new generation of AI can now synthesize images, model aspects of human behavior and reason across heterogeneous data—capabilities that are beginning to reshape how researchers and practitioners observe, model and support decisions about urban life. Here, in this Review, we explore what these tools can genuinely offer, where the evidence is strong and where it remains thin. The answer depends less on technical novelty than on whether generative AI can be evaluated honestly, validated in context and deployed in ways that keep human judgment and accountability at the center.

Urban analytics has emerged as an interdisciplinary field concerned with understanding, modeling and shaping the form and functioning of cities. The discipline comprises two interconnected domains: urban science, which focuses on the systematic analysis of urban phenomena through hypothesis testing, data analysis and interpretation modeling[1,2]; and urban practice, which applies this knowledge to guide planning, design and governance[3]. Researchers have historically relied on rule- and mechanism-based generative frameworks—from shape grammars[4] and pattern languages[5] to agent-based simulations[6]—to simulate urban processes using explicit assumptions and rules[7]. However, because these approaches depend on fixed assumptions and handcrafted rules, their capacity to adapt to new data, represent heterogeneous urban contexts or keep pace with fast-changing contemporary environments remains limited[8].

Recent advances in artificial intelligence (AI), combined with the rapid expansion of large-scale urban data, are reshaping how urban complexity is represented, inferred and acted upon. As illustrated in Fig. 1, much of this shift is driven by two categories of generative AI. The first comprises distribution-fitting generative models, such as generative adversarial networks (GANs)[9] and diffusion models[10], which learn statistical structure from high-dimensional data, ranging from satellite imagery to mobility traces, to generate plausible synthetic outputs that augment sparse sensing coverage and scenario generation[11]. The second comprises foundation models, including large language models (LLMs), pretrained on broad corpora and refined through instruction tuning[12,13], enabling instruction following, tool use and multimodal reasoning. Together, these categories expand 'generation' beyond data synthesis towards workflows that support analysis, communication and decision support in urban science and practice[14,15].

At the same time, deploying generative AI in urban environments introduces challenges that technical novelty alone cannot resolve. Distribution-fitting models can inherit and amplify biases in training data and often provide limited mechanistic interpretability[8], and foundation models can be sensitive to prompting, weakly grounded in domain evidence, and difficult to validate[16]. In high-stakes settings such as disaster response or large-scale infrastructure planning, these limitations become governance risks, including hallucinated claims, opaque uncertainty and institutional burdens of accountability[17]. Responsible integration therefore requires task-specific evaluation, transparent data provenance and institutional safeguards that preserve human oversight.

[1]Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. [2]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA. [3]Network Science Institute, Department of Physics, Northeastern University, Boston, MA, USA. [4]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Thrust of Urban Governance and Design, Society Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. [6]Institute for Climate and Carbon Neutrality, Urban Systems Institute, The University of Hong Kong, Hong Kong SAR, China. [7]Future Urbanity and Sustainable Environment Lab, Department of Architecture, Faculty of Architecture, The University of Hong Kong, Hong Kong SAR, China. [8]Institute of Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China. [9]School of Civil Engineering, Tsinghua University, Beijing, China. [10]Department of Geography, The University of Hong Kong, Hong Kong SAR, China. [11]Department of Earth and Planetary Sciences, The University of Hong Kong, Hong Kong SAR, China. [12]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [13]Department of Sociology, University of Chicago, Chicago, IL, USA. [14]Santa Fe Institute, Santa Fe, NM, USA. [15]Centre for Advanced Spatial Analysis, University College London, London, UK. [16]These authors contributed equally: Yunke Zhang, Fengli Xu. e-mail: liyong07@tsinghua.edu.cn

 https://doi.org/10.1038/s44284-026-00492-2

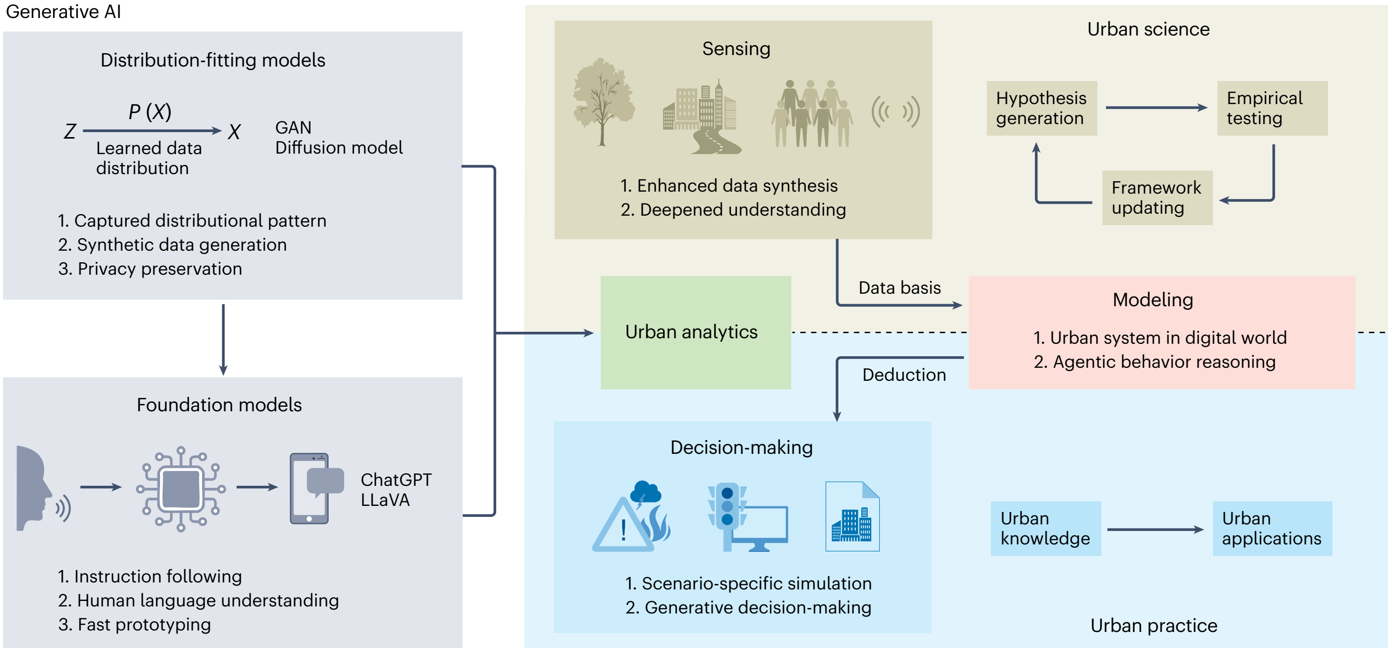

Fig. 1 | A conceptual illustration of how generative AI enables new capabilities in urban science and practice. Two categories of generative AI—distribution-fitting generative models and foundation models—support tasks across urban sensing, modeling and decision-making, while raising evaluation and governance requirements that shape responsible use.

These challenges are unfolding across multiple communities in parallel: remote sensing and geographic information systems (GIS) advancing geospatial AI (geoAI) for observation and spatial modeling[18,19]; urban science developing digital-twin pipelines for prediction and scenario generation[1,20]; design and planning exploring generative interfaces for participatory prototyping[21,22]; and governance research examining accountability and legitimacy under high-stakes decisions[23,24]. These approaches are often discussed in isolation, even though they form a single end-to-end workflow from observation to explanation to action.

We therefore organize this Review around three interconnected domains of the urban workflow: urban sensing, urban modeling and urban decision-making (Fig. 1). Within each domain, we critically synthesize how the two categories of generative AI are integrated, what problems they address and what constitutes credible evidence of reliability in urban settings, foregrounding evaluation under distribution shift, reproducibility and governance implications. We first clarify how traditional structural models differ from modern probabilistic generation, then review each domain in turn, before synthesizing cross-cutting challenges and pathways forward. Broader sociopolitical urban scholarship that does not engage generative computational methods is outside the scope of this Review.

Traditional urban models with generative structure

To clarify the novelty of modern generative AI in urban analytics, it is essential to distinguish it from the long-standing use of 'generative' structure in urban modeling. We use structurally generative models to denote rule- and mechanism-based frameworks that produce urban patterns through explicit assumptions, equations or decision rules, rather than by learning high-dimensional data distributions[4,7]. Modern distribution-fitting models generate samples by fitting distributions, and foundation models produce conditional outputs by following instructions and leveraging broad pretrained representations. Four families of structurally generative models have shaped urban research.

The first concerns spatial form and layout. Shape grammars[4] and pattern languages[5] use hierarchical constraints to generate complex urban configurations, and cellular automata[6] model spatial dynamics through local grid interactions, enabling interpretable simulations of urban growth. Their generative logic is manually encoded and tightly coupled to predefined design rules.

A second family comprises social and behavioral models. Gravity models[25] generate spatial flows through interaction potentials and distance frictions, and discrete choice models[26] formalize preferences over alternatives such as residential location or travel mode. Multi-agent systems[7,27] extend these ideas by simulating heterogeneous agents whose interactions yield emergent phenomena such as segregation and neighborhood change, supporting counterfactual reasoning about policy levers.

Transport and environmental models form the remaining two families. Traffic flow and assignment models[28,29] simulate congestion dynamics and network-wide path distributions under explicit behavioral assumptions. Urban canopy models[30] represent thermal exchanges between built form and atmosphere to support climate resilience analysis under interpretable physical parameterizations.

Collectively, these models offer mechanistic coherence and interpretability. However, their reliance on fixed rules and limited observational inputs constrains their capacity to capture high-dimensional, implicit patterns in contemporary urban environments[8]. Crucially, this distinction also reframes what validation means: structural models are assessed through mechanistic coherence; distribution-fitting models require distributional fidelity and robustness under extrapolation; and foundation models further introduce alignment, provenance and governance constraints that shape responsible use.

Conceptual framework for generative AI

The recent proliferation of urban big data has created a distinctive opportunity for modern generative AI. We organize this landscape into two categories that capture the dominant technical mechanisms and, importantly, imply different evaluation regimes: distribution-fitting generative models, exemplified by GANs[9] and diffusion models[10], and foundation models, including LLMs[13]. Distribution-fitting models learn statistical structure from observed urban data and have been primarily applied to data augmentation, reconstruction and scenario

 https://doi.org/10.1038/s44284-026-00492-2

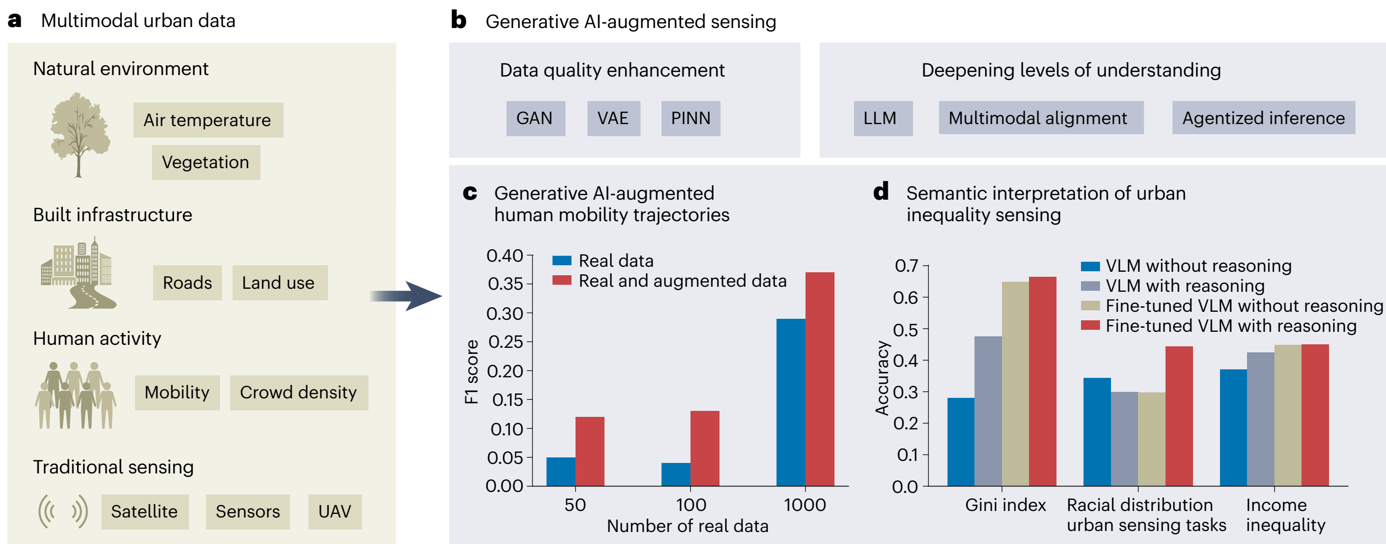

Fig. 2 | Paradigms of generative AI-augmented urban sensing. **a**,**b**, Generative AI methods extend traditional sensing pipelines by enabling distribution-fitting data synthesis for reconstruction and gap-filling (**a**) and foundation model-based semantic interpretation (**b**). PINN, physics-informed neural networks; VAE, variational autoencoder. **c**, Based on data from Yuan et al.[41]: synthetic augmentation can improve task-specific performance in daily activity prediction under sparse observations. **d**, Based on data from Zhang et al.[42]: VLMs fine-tuned on urban data can improve proxy estimation of inequality-related indicators from street-view imagery when paired with structured reasoning.

generation; they are assessed through distributional fidelity and robustness under extrapolation. Foundation models, pretrained on broad corpora and refined through instruction tuning[12], learn transferable statistical representations across language, images and code; they further introduce alignment, provenance and governance constraints that shape responsible use. Together, these two categories span the predominant modes by which contemporary systems 'generate' in urban contexts: either by sampling from learned data distributions, or by producing conditional outputs that integrate heterogeneous evidence through prompted reasoning.

A central implication concerns the nature of the knowledge produced: what foundation models acquire through pretraining is best understood as broadly transferable statistical representations, rather than verified, theory-grounded knowledge. Consequently, their outputs should be interpreted as conditional hypotheses or scenarios, whose scientific value depends on grounding, validation and transparent provenance.

To connect these model categories to urban science and practice, we adopt a workflow view of urban inquiry in which observation, explanation and action are tightly coupled yet institutionally distinct (Fig. 1). Urban science must integrate heterogeneous measurements into coherent representations and testable explanations, whereas urban practice must translate evidence into decisions under constraints of legitimacy, equity and accountability. Generative AI increasingly mediates both sides of this cycle: distribution-fitting models can enrich observations and simulate plausible counterfactuals, and foundation models are increasingly explored as interfaces for coordinating heterogeneous evidence and supporting participation around complex trade-offs[21].

Throughout this Review, we use three interconnected domains—urban sensing, urban modeling and urban decision-making—as a workflow scaffold from observation to action (Fig. 1). This scaffold is not merely organizational; it clarifies what counts as evidence in each domain and how different communities contribute to urban claims.

- Urban sensing concerns data acquisition and enrichment, including filling spatial and temporal gaps. Credible evidence hinges on quantified coverage and bias, uncertainty characterization and robustness under distribution shift, particularly across cities and data-sparse contexts.
- Urban modeling transforms observations into representations, simulations and forecasts. Evaluation requires grounding and coherence checks, calibration against held-out observations or trusted simulators, and stress tests probing extrapolation beyond observed regimes.
- Urban decision-making translates evidence into governance processes and practitioner-facing tools. Evaluation should extend beyond predictive metrics to transparency and provenance of generated claims, procedural fairness and institutional accountability—framing generative systems as decision support rather than automated decision-makers[31].

Generative AI-augmented urban sensing

Urban sensing underpins empirical urban inquiry by collecting and interpreting multimodal measurements of the natural environment, built infrastructure and human activity. These data enable hypothesis testing and model calibration, yet they remain incomplete, unevenly distributed and often difficult to translate from raw signals into urban meaning[32]. Traditionally, sensing has relied on physical instrumentation (for example, global navigation satellite systems (GNSS), light detection and ranging (LiDAR) and satellite remote sensing) to capture spatial and behavioral dynamics. Although indispensable, these technologies face recurrent limitations, including spatial sparsity, temporal gaps and a semantic gap between pixels and social interpretation. Generative AI can address these limitations through two complementary capabilities (Fig. 2): distribution-fitting data synthesis and foundation model-based interpretation.

- Synthesizing urban data with distribution-fitting models. Distribution-fitting generative models can mitigate data sparsity by learning the statistical structure of urban observations to reconstruct missing values and synthesize fine-grained detail. In physical sensing, GANs and diffusion models have demonstrated the ability to perform super-resolution from low-resolution satellite imagery and to fill missing observations, such as cloud-obscured pixels[11,33,34]. Beyond purely statistical reconstruction, physics-informed generative frameworks can embed physical constraints

Table 1 | Comparison of classical and generative AI approaches in urban modeling

Modeling domain	Classical methods	Generative AI tasks and examples
Natural environment	Numerical simulations (CFD, WRF), interpolation (Kriging)	Flood and climate forecasting: GANs for rapid flood inundation mapping[51]; super-resolution for urban heat island monitoring[34]; AI-driven extreme event prediction[52]
		Emission estimation: generative inference of street-level carbon footprints[53]
Built environment	Procedural generation (CityEngine), GIS manual drafting	Route network generation: diffusion-based synthesis of topology-preserving street networks (for example, RN-Diff)[54]
		Urban layout synthesis: conditional generation of building footprints and zoning plans from text or satellite imagery[21,55]
		Traffic simulation: generative stress testing of traffic scenarios[56]
Human activity	Gravity models, survey-calibrated heuristics	Mobility dynamics simulation: synthesizing aggregate mobility traces via distribution fitting (for example, MoveSim)[57,58]
		Generative agents: LLM-driven simulation of detailed daily routines, navigation and social interactions[59,60]
Hybrid systems	Coupled static models	Holistic social simulation: large-scale simulators (for example, AgentSociety and OpenCity) co-simulating urban mobility, economic activities and social network dynamics[60,63]; dynamic forecasting of city-wide phenomena[106]

Generative AI complements mechanistic and rule-based models by learning probabilistic emulators and agent-based simulators that support scenario generation and stress testing. CFD, computational fluid dynamics; WRF, weather research and forecasting model.

to improve plausibility when simulating urban microclimates from sparse sensors[35], and real-time models deployed on unmanned aerial vehicles (UAVs) can dynamically augment data streams for time-sensitive infrastructure monitoring[36]. However, strictly distinguishing signal recovery from hallucination remains challenging; generative models may prioritize perceptual plausibility over physical veracity, necessitating rigorous verification against ground truth before such data are used for downstream analytics. In human activity sensing, generative models have also been used to synthesize privacy-preserving mobility trajectories that preserve aggregate statistical properties while obfuscating individual identities[37], creating a pathway to share granular mobility proxies where raw data access is restricted by privacy constraints.

- Interpreting urban context with foundation models. Foundation models offer a complementary route to address the semantic gap between raw sensing signals and urban interpretation. Many urban sensing pipelines now combine heterogeneous inputs—from imagery and street view to mobility traces and administrative records—whose integration and interpretation can be prohibitive at scale. Foundation models can learn latent representations that compress heterogeneous signals into shared spaces for retrieval, comparison and downstream inference. For example, complex geospatial and business activity data can be embedded into low-dimensional representations to support analysis of population dynamics and urban behavior[18]. Building on such representations, vision–language models (VLMs) leverage broad pretraining to move beyond object detection towards functional semantics and contextual interpretation. Recent benchmarks suggest that VLMs can support tasks such as identifying functional attributes, detecting anomalous events and narrating changes across space and time[38–40]. These capabilities shift urban sensing from passive measurement toward task-driven semantic inference, while raising new requirements for evidence grounding and failure-mode auditing.

The utility of generative augmentation is supported by emerging quantitative evidence, although any reported gains are inherently task and dataset specific and should not be extrapolated into claims about long-horizon prediction or policy outcomes without downstream evaluation. Augmenting sparse survey data with GAN-generated trajectories has been reported to improve daily activity prediction relative to baselines trained on real data alone[41] (Fig. 2c), and incorporating chain-of-thought reasoning in VLM pipelines can improve proxy estimation of socioeconomic indicators from street-view imagery[42] (Fig. 2d). Operationally, validating generative augmentation requires making the evidentiary burden explicit: quantifying coverage and representation bias in the underlying observations[43]; benchmarking against independent ground truth or external datasets where available[44]; stress-testing robustness under missingness, noise and rare regimes; and assessing transfer under distribution shift[19], with particular caution in data-sparse contexts.

Despite their promise, generative sensing systems pose ethical and representational risks. Distribution-fitting models can fail to reproduce rare but critical events (for example, extreme weather or atypical mobility behaviors), and foundation models trained predominantly on global north data risk hallucinating Westernized urban features when applied to underrepresented regions[45]. Such errors can be consequential: reconstructing an informal settlement with the visual texture of a planned suburb may erase markers of inequality relevant to downstream analysis[46]. Responsible deployment therefore requires rigorous auditing for distribution shift and representational harm, alongside active debiasing strategies, to avoid reinforcing the 'digital invisibility' of marginalized communities[47].

Generative AI-augmented sensing thus extends traditional geomatics beyond measurement alone by combining physical observations with synthetic augmentation and semantic interpretation. When evaluated with task-appropriate protocols and deployed with governance safeguards, these methods can expand observational coverage and enrich urban understanding, while clarifying where unresolved representation and reliability challenges remain.

Generative urban modeling with digital twins and language agents

Generative AI is reshaping urban modeling by complementing static representations with probabilistic emulation and interactive simulation[1]. Whereas classical urban models typically rely on theory and calibrated equations or rule-based heuristics to represent system states, modern generative AI introduces two complementary modeling capabilities: fast, high-dimensional emulation of physical dynamics using distribution-fitting models, and semantic simulation of human behavior using foundation model-based agents. These advancements support next-generation digital twins that can simulate not only what is observed, but also plausible counterfactual and test scenarios under explicit assumptions[14,15,48]. We organize the literature into four core domains that recur across digital-twin pipelines: the natural environment, the built environment, human activity and hybrid systems (Table 1).

In the natural and built environment, distribution-fitting models often act as high-dimensional surrogates for expensive physical or procedural pipelines. Unlike mechanistic numerical solvers, which offer high precision but are computationally expensive, probabilistic emulators learn input-to-field mappings at substantially reduced runtime, trading mechanistic transparency for speed and flexibility. Their scientific value is therefore best established through evidence tied to the task—agreement with trusted simulators or held-out observations under defined regimes, sensitivity to distribution shift, and robustness under test scenarios[49]—avoiding generic claims of 'accuracy' when models are used for forward-looking scenario exploration where policy outcomes lack ground truth[50].

- Natural environment. Generative models facilitate environmental emulation and downscaling by synthesizing high-resolution fields from sparse or coarse inputs. GAN-based surrogates have been used to emulate flood inundation and downscale climate-related fields relevant to urban heat and hazard assessment[34,51,52]. Notably, recent work fuses traffic signals with vehicle physics to estimate road-transport emissions at street level in near real time[53]—illustrating how generative inference can bridge physical and behavioral data streams beyond purely statistical reconstruction.
- Built environment. Graph-based models and diffusion processes can be conditioned to synthesize road networks and building layouts satisfying connectivity or efficiency constraints[21,54,55]. Importantly, generative models can also produce rare or adversarial regimes for scenario stress testing, complementing classical simulators that undersample extremes[56]—a capability that is difficult to achieve through procedural generation alone.

Modeling human activity combines aggregate distribution fitting with agent-based simulation at finer semantic resolution. Distribution-fitting approaches reproduce city-scale mobility statistics[57,58], and foundation-model-based generative agents introduce a distinct capability: simulating routines and interactions via natural-language goals, memory, reflection and planning[59,60]. This shift does not imply that agents understand urban processes in any deep sense; rather, it enables conditional behavioral narratives that function as hypotheses requiring empirical assessment.

- From rules to semantic simulation. Compared with heuristic agent-based models that encode behavior through fixed utility functions, LLM-based agents can condition actions on natural-language constraints and contextual cues, generating diverse schedules and emergent social outcomes such as polarization[59,60]. However, cross-national evaluations across 107 countries document systematic clustering of LLM responses toward Western cultural dimensions, with non-Western value systems consistently underrepresented[61]—suggesting that the challenge for urban social simulation is representational rather than merely linguistic. Without careful calibration, agents risk exhibiting 'simulacra' behavior, mimicking superficial social norms in training data rather than authentic decision-making[62].
- Scalability of agents. Recent systems have scaled agent-based simulations to the order of 10^4 LLM agents using distributed orchestration and efficient memory management[60,63,64]. Policy-grade city-scale deployment, however, remains constrained by inference cost, calibration overhead and governance requirements[65]. Current trajectories therefore favor hybrid architectures that combine foundation-model agents with structured simulators, hierarchical role decomposition and distillation or caching to reduce per-step inference costs, while retaining auditable evaluation protocols[23].

Across domains, generative modeling claims are best evaluated against explicit empirical tests rather than generalized notions of 'accuracy': environmental surrogates against hydrodynamic simulations or held-out measurements, and behavioral agents against external observational benchmarks or controlled experiments. Such evaluations support the use of generative models for scenario exploration and stress testing but do not by themselves establish validated long-horizon policy prediction where ground truth is unavailable[51,60].

Probabilistic generation foregrounds uncertainty in ways that differ from deterministic simulation. Urban modeling faces stochastic uncertainty from human behavior and measurement noise, epistemic uncertainty from underspecified models and distribution shift, and scenario uncertainty from structural change and policy feedbacks[66]. Failures often arise when plausible samples are misread as forecasts, or when uncertainty is not explicitly communicated. Mitigating these risks requires grounding at two levels: at the technical level, constraining generation with physical priors and explicit consistency checks to reduce implausible outputs[35,67]; and at the deployment level, institutional guardrails such as fail-safe defaults, audit trails and human-in-the-loop review to ensure digital twins function as decision support rather than automated decision-makers[68].

These components are increasingly assembled into digital-twin-style pipelines that couple natural, built and human subsystem models with streaming observations[20,69], supporting interactive scenario comparison and structured deliberation over trade-offs. Substantial open problems remain, however, in reliability under distribution shift, representational coverage and institutional accountability.

Informing decision-making in urban space with generative AI

Recent advances in generative AI are beginning to reshape urban decision support, complementing traditional optimization- and simulation-based tools with scenario generation, interactive interfaces and language-mediated reasoning[55,70]. Unlike many conventional decision support systems that operate on fixed objective functions and closed-form constraints, foundation-model-based systems can synthesize heterogeneous evidence, translate between technical and narrative representations, and scaffold deliberation over competing goals[8]. At the same time, because urban decisions involve legal authority, distributional consequences and contested values, the governance literature consistently positions generative systems as decision support rather than automated decision-makers, with responsibility for justification and outcomes remaining with human institutions[23].

Decision support has historically relied on mechanistic simulation, mathematical optimization and reinforcement learning. These approaches offer interpretability and constraint guarantees, yet they can be brittle under distribution shift or limited by model specification. Generative models trade strict guarantees for flexibility, enabling exploration of open-ended scenarios and richer representations of heterogeneous behavior[48].

The decision-making literature can be organized across three temporal scales, comprising short-term response, medium-term management and long-term planning (Fig. 3), to reflect how generative systems interface with governance under differing time constraints, evidence standards and legitimacy requirements.

- Short-term emergency management. Rapid-onset events such as extreme weather or crowd surges demand swift, informed responses under uncertainty. Generative AI can contribute to anticipatory governance by pairing environmental nowcasting with behavioral scenario exploration. For instance, digital-twin pipelines coupled with weather prediction systems can model evolving flood conditions[71], and LLM-based agents can explore plausible evacuation narratives under varying constraints[72]. Physics-informed generative methods can also simulate pedestrian dynamics in emergency settings, reporting improved benchmark

https://doi.org/10.1038/s44284-026-00492-2

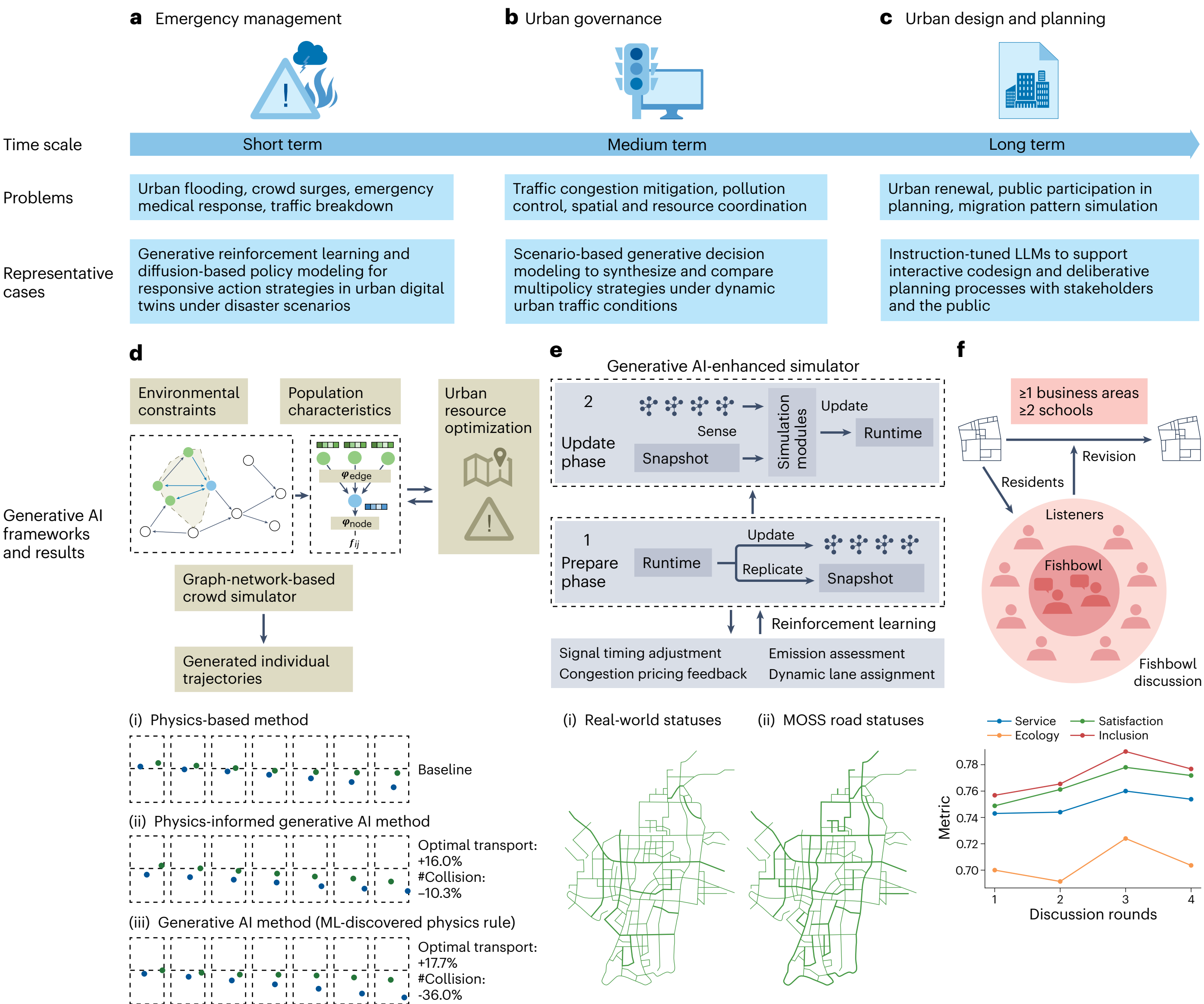

Fig. 3 | Representative uses of generative AI for informing urban decision-making. a–c, Generative methods can augment decision processes across temporal scales: short-term emergency management (**a**), medium-term urban governance (**b**) and long-term urban design and planning (**c**). **d**, Physics-informed generative models simulate pedestrian evacuation dynamics, reporting improved agreement with observed trajectories in controlled benchmarks relative to social force baselines[73]. ML, machine learning. **e**, Generative traffic modeling (mobility simulation system (MOSS)) supports rapid scenario evaluation, reporting reduced mismatch to held-out simulation or observation proxies relative to rule-based baselines under specific test settings[75]. **f**, LLM-based agents have been used to prototype participatory planning deliberation, with pilot studies reporting measurable changes in participant-rated satisfaction under defined protocols[16].

agreement over classical social-force baselines[73] (Fig. 3d). Such tools are best understood as generators of testable scenarios for staffing, routing and resource allocation, whose outputs should be verified against operational constraints rather than treated as forecasts.

- Medium-term urban governance. Operational governance (weeks to months) involves coordinating interdependent systems such as traffic, utilities and public services. Traditional optimization can struggle with combinatorial complexity and shifting regimes[74]. Generative models can support policy design by rapidly synthesizing plausible system trajectories and stress-test scenarios that complement mechanistic simulation, while retaining classical simulators as verification and constraint-checking domains for accountability. In traffic governance, generative modeling can accelerate the evaluation of interventions by enabling broader coverage of rare regimes[75,76] (Fig. 3e), and reinforcement learning pipelines trained against generative simulation can support fast prototyping of adaptive management strategies[77].
- Long-term urban design and planning. Strategic planning operates over decades, where evidence is incomplete and goals are often value-laden. Here, generative AI functions less as an optimizer and more as a mediator for participation and communication. Beyond generating candidate layouts from contextual inputs[21,78], instruction-tuned LLMs can lower barriers to engagement by translating technical proposals into accessible narratives and facilitating iterative feedback loops. Recent frameworks deploy LLM agents to prototype community deliberation and multiround refinement of land-use plans[16], with pilot studies reporting measurable shifts in participant-rated satisfaction (Fig. 3f)—evidence of deliberation quality rather than validation of long-horizon urban outcomes. These systems still require human facilitation and institutional safeguards[20].

 https://doi.org/10.1038/s44284-026-00492-2

Table 2 | Capabilities, challenges and pathways forward across the urban workflow

Domain	Representative capabilities	Challenges and impacts	Open problems and pathways forward
Sensing	Super-resolution and gap-filling; privacy-preserving trajectory synthesis; semantic interpretation with VLMs	Validation and reliability: plausible reconstructions can hide systematic error under distribution shift	Unsolved: standardized protocols for bias, uncertainty and cross-city transfer; auditing representational harm, not only average error
		Representation and equity: uneven coverage and global north pretraining can erase informal or underrepresented urban forms[45,46]	Pathways: data-based evaluation with coverage and bias quantification, external benchmarking, robustness stress tests under missingness/noise, and cross-city generalization tests; reporting standards for uncertainty and failure modes[19,47]
		Impact: biased or hallucinated observations propagate downstream into models and decisions	
Modeling	Surrogate emulation for floods, heat and emissions; scenario generation and stress testing; LLM-enabled agent simulation	Validation and reliability: forward-looking scenarios often lack direct ground truth, and samples can be misread as forecasts[84]	Unsolved: policy-grade evaluation for coupled socio-physical feedbacks and long horizons; uncertainty communication that prevents forecast misinterpretation
		Uncertainty: epistemic uncertainty under structural change and feedbacks is often opaque[66]	Pathways: grounding and calibration via agreement with trusted simulators or held-out observations under defined regimes, uncertainty calibration and out-of-regime stress tests; hybrid architectures that retain verification domains and auditable assumptions[67,68]
		Impact: uncalibrated extrapolation can create spurious precision in policy discussion	
Decision-making	Scenario ideation; interface prototyping; summarization of stakeholder input; participatory deliberation scaffolding	Governance and accountability: unclear responsibility when generated claims shape high-stakes choices	Unsolved: auditable standards linking model outputs to decision justification under legal authority and distributional consequences
		Process risks: automation bias and institutional black-boxing reduce contestability[24,81]	Pathways: application-based evaluation in real workflows, provenance and documentation requirements, human-in-the-loop review for high stakes, participatory safeguards and outcome audits where feasible[79]
		Impact: legitimacy and due process can erode even when task metrics improve	

Each domain has distinct evidentiary burdens. The table summarizes how challenges translate into downstream impacts, and what remains open for research and practice.

Because generative systems can produce plausible outputs without sufficient evidentiary grounding, responsible use requires interpreting uncertainty, detecting hallucinations and maintaining audit trails and contestability mechanisms in high-stakes settings[31,79]. Two institutional risks are well documented in empirical studies of government algorithm deployment[80]: automation bias, where overreliance on algorithmic advice narrows deliberation and reduces the likelihood that officials detect model failures[81]; and 'algorithmic bureaucracy', where complex systems become institutionally black-boxed, obscuring how evidence becomes recommendations and complicating legal requirements for transparency[24]. Addressing these risks requires governance by design: human-in-the-loop review for high-stakes decisions, auditable provenance and evidence tracing that makes uncertainties and trade-offs explicit.

Challenges and debates

Generative AI is increasingly embedded across urban sensing, modeling and decision support, but its adoption sharpens three cross-cutting debates that align with the workflow of this Review: validation and reliability, representation and equity, and governance and accountability (Table 2). These debates are not abstract concerns. They determine what claims can be supported as evidence in each domain, how errors propagate from observation to explanation to action, and which safeguards are required for responsible use. The shift from deterministic pipelines to probabilistic generation expands the space of plausible outputs, making it easier to produce convincing artifacts while harder to establish when, where and why they should be trusted.

Across domains, the validation debate is best framed not merely as a search for plausibility, but as a question of whether generative claims can sustain scientific and ethical criticism. We distinguish two complementary families of standards[23]. Data-based evaluation asks whether generative outputs remain faithful to observations under defined regimes, including quantified coverage and bias, benchmarking against independent ground truth, robustness under noise and missingness, and transfer under distribution shift[82]. Application-based evaluation asks whether outputs remain reliable within real decision processes—whether claims are traceable to evidence, whether uncertainty is communicated in decision-relevant forms, and whether governance constraints prevent overreliance in high-stakes settings[23,31].

- Validation and reliability: from plausible outputs to testable claims. Generative systems expand the space of plausible reconstructions and futures, but plausibility is not validity[83]. In modeling, forward-looking scenarios often lack ground truth, so reliability must be established through agreement with trusted simulators or held-out observations, calibrated uncertainty and out-of-regime stress tests[84]. In decision support, reliability further includes whether generated claims can be inspected, compared and contested in institutional workflows[79].
- Representation and equity: whose city is learned and whose is erased. Because urban observations are unevenly distributed and socially patterned, generative systems can inherit representational asymmetries and amplify them through synthesis and semantic interpretation[83]. This debate is especially acute in data-sparse contexts, where failures may map onto vulnerable communities and informal urban forms[45,46]. The methodological frontier is operationalizing representational harm through auditing that reports not only aggregate performance, but also the spatial and social structure of omissions and errors, and how these propagate through the workflow[47].

- Governance and accountability: from model capability to institutional legitimacy. Even when task performance improves, generative systems can increase governance burdens by obscuring how evidence becomes recommendations, encouraging automation bias and complicating transparency and contestability requirements[24,81]. Harm can arise from process failures, not only from output errors[83]. Pathways forward require governance-by-design, including provenance and documentation, human review for high-stakes use, and participatory mechanisms that surface contested values[79,85].

Future research directions

Generative AI has the potential to reshape urban science and practice, but its credible use depends less on novelty than on whether generated outputs can be validated, whether representation gaps are reduced rather than amplified, and whether governance capacity and ethical standards keep pace with deployment. We outline five evaluation-forward directions, each targeting a specific evidence gap and clarifying how progress would interface with the three domains of the urban workflow[19,23,68].

Multiscale and multimodal urban models

Urban systems are observed through heterogeneous signals across scales, including satellite imagery, street view, administrative records, sensor streams and mobility traces[18], yet many generative pipelines still operate on narrow inputs that limit external validity and extrapolation under distribution shift. A growing body of recent work has explored multiscale, multimodal integration as a means to improve validation and transfer: combining complementary evidence sources has been shown to reduce spurious correlations[86] and to support stress tests over missingness and rare regimes[19].

The implications of this integration have been examined separately for each domain of the urban workflow. In the sensing domain, multimodal fusion has been applied to strengthen coverage auditing, provenance tracking, uncertainty characterization and bias assessment[19,68]. In the modeling domain, integrated observations have been used to support calibration, data assimilation and consistency checks—including physics- or mechanism-informed constraints where appropriate—to improve robustness under regime change[35,67]. In the decision domain, the shift from single-modality to multimodal inputs has raised questions about how to produce interpretable, decision-relevant summaries rather than opaque outputs—an area of active methodological debate[31,68]. Large urban models (urban LMs), including CityGPT[87] and UrbanLLaVa[88], represent one emerging attempt at cross-modal unification; how to couple such representations to transparent validation and governance protocols remains an open challenge.

Human experience and evaluation

Cities are lived environments, and many urban objectives are normative (safety, dignity and access) rather than purely physical. Embedding human experience into generative urban pipelines calls for evaluation and governance commitments that go beyond richer data alone. Experiential ground truth can be partially operationalized through participatory measurement (surveys, community reporting and qualitative annotation) and careful triangulation across administrative records, mobility proxies and imagery, with explicit documentation of whose experiences are captured and whose are missing[23,79]. Models can be audited for stereotyping and narrative collapse: instruction-following interfaces can unintentionally harden simplified stories about neighborhoods unless counter-evidence and contestation are built into the workflow[24,45]. Participatory validation is increasingly treated as a methodological component: model outputs should be critique-able by affected stakeholders through structured feedback loops, not merely explained after deployment[21].

The literature points to a documented competency gap in urban AI education between model-centric training and the skills needed for responsible decision support: uncertainty interpretation, evidence tracing, comparative scenario analysis and institutional documentation for audit and contestability[31]. Field-building work on toolkits, reporting standards and curriculum development has been characterized as enabling infrastructure that aligns technical innovation with reproducible evaluation and accountable use[68].

Neurosymbolic and grounding approaches

Scaling matters because policy-relevant urban questions are rarely neighborhood-sized: resilience planning involves cascading failures across mobility, energy, health and supply chains. Recent work demonstrates that LLM-enabled agent societies and hybrid simulators are beginning to scale into the 10^4-agent regime, enabling richer scenario exploration than traditional rule-based agent-based models in some settings[60,63]. Scaling alone, however, does not by itself guarantee policy-grade reliability: as simulations become larger and longer-horizon, calibration overhead and failure detection become harder, not easier.

Recent work on physics-informed and constraint-aware learning has developed complementary grounding strategies, offering mechanisms to reduce physically implausible generations and improve out-of-regime behavior by incorporating conservation laws, operators and structured priors[35,67,89]. For urban resilience, a key unresolved challenge is how to design grounding that is heterogeneous: accommodating physical constraints for environmental fields, behavioral constraints for human activity, and institutional constraints for decision workflows simultaneously. Neurosymbolic approaches and hybrid architectures have been explored as one direction toward this goal, combining generative flexibility with explicit rules, simulators and verifiers[35,89]. Several methodological questions remain unresolved, including how to evaluate such grounded systems under structural change, how to communicate uncertainty in decision-relevant forms, and how to prevent grounded constraints from inadvertently encoding contested values as if they were neutral rules[23,31].

Bridging global south data gaps

Much future urban population growth is projected to occur in Africa and in parts of South and Southeast Asia, with additional pressures across Latin America and the Caribbean[90]. Cities across these regions combine rapid structural change with informality, fragmented registries, limited sensing infrastructure and constrained institutional capacity for data governance[91]. These conditions do not merely reduce the quantity of available data; they also change the ways in which generative AI can fail. Distribution-fitting models trained on global north imagery have been documented to smooth over or erase the morphology of informal settlements[92], and foundation-model interfaces predominantly trained on global north corpora encode formal urban assumptions poorly aligned with informal morphology and local planning realities—with the risk of misrepresentation concentrated in precisely those settings where vulnerability is greatest[45,46].

The literature documents both progress and persistent limits. Remote sensing and open mapping have enabled systematic settlement characterization in data-scarce contexts, yet cross-city generalization has consistently emerged as a bottleneck: models trained on one city routinely degrade when transferred to morphologically distinct cities, even within the same country[92–94]. Global building-footprint datasets have expanded baseline coverage, yet empirical audits reveal systematic biases against poorer and more rural populations[95], and cross-dataset comparisons find that major products—including Google Open Buildings, Microsoft Building Footprints and OpenStreetMap—are not interchangeable, with none capturing building type, use or occupancy beyond geometry[96]. OpenStreetMap coverage is similarly uneven, with cities in South Asia averaging 9% building completeness compared with 71% in Europe, and more than half of all edits in sub-Saharan

Africa originating from organized humanitarian campaigns rather than local contributors[97].

Open tools and benchmark frameworks have lowered barriers to reproducible evaluation, but technical adaptation alone has not proved sufficient where training capacity, documentation infrastructure and contestability mechanisms remain underdeveloped[19,98,99].

The evidence reviewed above suggests that, in many global south planning contexts, the binding constraints are institutional legitimacy, fiscal capacity and service delivery rather than scenario generation. Simpler causal-inference designs or domain-adapted tools for regulatory audit may offer more viable near-term pathways than high-capacity generative systems. Recent work in Santiago, Chile, illustrates this directly: LLMs fine-tuned on local legal-urban discourse were used to audit 31 communal zoning ordinances, uncovering normative asymmetries that perpetuate spatial segregation—a use case grounded in institutional accountability rather than synthetic scenario generation[100]. The governance literature consistently points to participatory problem definition, local validation and sustained maintenance as conditions for responsible deployment[23,31].

Ethical considerations and inclusive modeling

Ethics in generative AI for urban research is increasingly framed around auditable requirements rather than broad principles. The literature can be organized around five recurring themes, each involving both technical methods and reporting standards.

- Bias and fairness. Generative systems can inherit representational asymmetries and amplify them through reconstruction and synthesis, with disproportionate effects in marginalized and data-sparse communities[45,46]. Mitigation requires place-aware fairness assessment that reports the spatial and social structure of errors, not only average performance[47,101]. What remains unresolved is standardizing fairness metrics that reflect urban impacts rather than abstract parity.
- Privacy. Distribution-fitting synthesis can enable privacy-preserving sharing of mobility or activity traces, but privacy gains are not automatic and require explicit threat models and evaluation[37]. A major open problem is establishing urban-appropriate privacy benchmarks that reflect realistic re-identification risks and institutional data governance constraints.
- Explainability. In decision contexts, the relevant question is not only interpretability of internal representations, but traceability of claims to evidence and assumptions. Documentation tools and provenance tracking can make outputs contestable and reviewable, which is critical for legitimacy in public-sector decisions[79,102]. A persistent gap is aligning explainability with the needs of non-technical stakeholders without collapsing uncertainty.
- Controllability and safety. Safety failures often arise when plausible generations are misread as forecasts or when models behave unpredictably under distribution shift. Control mechanisms therefore include stress testing under shift, calibrated uncertainty reporting and fail-safe defaults for high-stakes use[31]. Standards for such controls are beginning to emerge in risk-management frameworks, but urban deployments still lack consistent thresholds for escalation, human review and rollback[80].
- Participatory governance. Urban decisions embed contested values and the literature increasingly treats participatory mechanisms as important safeguards for exposing differences in how benefits and burdens are distributed. Generative interfaces can lower barriers to engagement, but participation is most consequential when it supports contestability and accountability rather than token consultation[23,85]. Across all five themes, reproducibility and transparent reporting are prerequisites for ethics and governance, because accountability cannot be enforced without auditable inputs, versions and protocols[68,102].

These five themes interact, and each has developed a distinct empirical base. Place-aware auditing frameworks have been developed to report the geographic structure of model errors rather than aggregate performance[101]. Federated learning has been applied to urban problems such as traffic flow prediction to enable collaborative training without centralizing sensitive mobility data[103], although privacy protections depend heavily on implementation choices and cannot be assumed automatically. Empirical analyses of municipal algorithmic deployments in New York City, Seattle and San Diego find that transparency requirements are routinely prioritized over structural accountability, and that community participation is frequently reduced to token consultation without pathways to institutional contestation[85]; mandates for individual end-stage review similarly provide limited protection against automation bias under institutional pressure[80]. Across these strands, the literature consistently points to combining technical auditing with institutional mechanisms for contestability, documentation and iterative local validation as necessary conditions for accountable deployment[104,105].

Conclusion

This Review synthesizes how two categories of generative AI—distribution-fitting models and foundation models—are being integrated into a three-domain urban workflow spanning sensing, modeling and decision-making. Rather than cataloging techniques, we foreground evaluation: in each domain, we highlight what constitutes credible evidence, where validation protocols remain underspecified, and how uncertainty and distribution shift complicate claims about future scenarios. Across the literature, three unresolved tensions recur: validation and reliability, representation and equity, and governance and accountability.

Generative systems can meaningfully augment urban analysis and deliberation when they are deployed as decision support under explicit guardrails: grounded constraints, transparent documentation of data and inference settings, and human-in-the-loop oversight that preserves institutional accountability. Progress in this area depends not only on better models, but also on shared standards for evaluation, reporting and participatory governance that make model assumptions and trade-offs legible. Under these conditions, generative AI can contribute to more robust urban science and more responsible urban practice, while clarifying the limits of what can be claimed, validated and justified.

References

1. Bettencourt, L. M. A. *Introduction to Urban Science: Evidence and Theory of Cities as Complex Systems* (MIT Press, 2021).
2. Batty, M. *The Computable City: Histories, Technologies, Stories, Predictions* (MIT Press, 2024).
3. Batty, M. Big data, smart cities and city planning. *Dialogues Hum. Geogr.* **3**, 274–279 (2013).
4. Stiny, G. & Gips, J. Shape grammars and the generative specification of painting and sculpture. In *IFIP Congress* vol. 2, 125–135 (Citeseer, 1971).
5. Alexander, C., Ishikawa, S. & Silverstein, M. *A Pattern Language: Towns, Buildings, Construction* (Oxford Univ. Press, 1977).
6. White, R. & Engelen, G. Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. *Environ. Plann. A* **25**, 1175–1199 (1993).
7. Schelling, T. C. Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971).
8. Zhang, Y. et al. Metacity: data-driven sustainable development of complex cities. *Innovation* **6**, 100775 (2025).
9. Goodfellow, I. J. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, 2672–2680 (NeurIPS, 2014).
10. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).

11. Wang, Q. et al. Generative ai for urban planning: synthesizing satellite imagery via diffusion models. *Comput. Environ. Urban Syst.* **122**, 102339 (2025).
12. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
13. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
14. Gao, C. et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanit. Soc. Sci. Commun.* **11**, 1259 (2024).
15. Wang, J. et al. Large language models as urban residents: an LLM agent framework for personal mobility generation. In *Proc. 38th Annual Conference on Neural Information Processing Systems*, 124547–124574 (NeurIPS, 2024).
16. Zhou, Z., Lin, Y., Jin, D. & Li, Y. Large language model for participatory urban planning. Preprint at https://arxiv.org/abs/2402.17161 (2024).
17. Huang, L. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **43**, 1–55 (2025).
18. Agarwal, M. et al. General geospatial inference with a population dynamics foundation model. Preprint at https://arxiv.org/abs/2411.07207 (2024).
19. Tuia, D. et al. Artificial intelligence to advance earth observation: a review of models, recent trends, and pathways forward. *IEEE Geosci. Remote Sens. Mag.* **13**, 119–141 (2025).
20. Batty, M. Digital twins in city planning. *Nat. Comput. Sci.* **4**, 192–199 (2024).
21. Zheng, Y. et al. Spatial planning of urban communities via deep reinforcement learning. *Nat. Comput. Sci.* **3**, 748–762 (2023).
22. Fu, X., Li, C., Quan, S. J., Yigitcanlar, T. & Wasserman, D. Large language models in urban planning. *Nat. Cities* **2**, 585–592 (2025).
23. Taeihagh, A. Governance of generative ai. *Policy Soc.* **44**, 1–22 (2025).
24. Kronblad, C., Essén, A. & Mähring, M. When justice is blind to algorithms: multilayered blackboxing of algorithmic decision-making in the public sector. *MIS Quart.* **48**, 1637–1662 (2024).
25. Wilson, A. G. A family of spatial interaction models, and associated developments. *Environ. Plann. A* **3**, 1–32 (1971).
26. Ben-Akiva, M. E. & Lerman, S. R. *Discrete Choice Analysis: Theory and Application to Travel Demand* Vol. 9 (MIT Press, 1985).
27. Bonabeau, E. Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl Acad. Sci. USA* **99**, 7280–7287 (2002).
28. Lighthill, M. J. & Whitham, G. B. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc. R. Soc. A* **229**, 317–345 (1955).
29. Sheffi, Y. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods* (Prentice-Hall, 1985).
30. Arnfield, A. J. Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *Int. J. Climatol.* **23**, 1–26 (2003).
31. Marchau, V. A., Walker, W. E., Bloemen, P. J. & Popper, S. W. *Decision Making under Deep Uncertainty: From Theory to Practice* (Springer, 2019).
32. Bettencourt, L. M. A. Recent achievements and conceptual challenges for urban digital twins. *Nat. Comput. Sci.* **4**, 150–153 (2024).
33. Khanna, S. et al. DiffusionSat: a generative foundation model for satellite imagery. In *The 12th International Conference on Learning Representations*, 5586–5604 (ICLR, 2024).
34. Yu, Z., Liu, C., Liu, L., Shi, Z. & Zou, Z. MetaEarth: a generative foundation model for global-scale remote sensing image generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**, 1764–1781 (2024).
35. Xu, E. et al. A survey of physics-informed AI for complex urban systems. *Inf. Fusion* **129**, 104012 (2026).
36. Jankovic, B., Jangirova, S., Ullah, W., Khan, L. U. & Guizani, M. UAV-assisted real-time disaster detection using optimized transformer model. In *2025 IEEE Symposium on Computers and Communications* 1–7 (IEEE, 2025).
37. Berke, A., Doorley, R., Larson, K. & Moro, E. Generating synthetic mobility data for a realistic population with rnns to improve utility and privacy. In *Proc. 37th ACM/SIGAPP Symposium on Applied Computing* 964–967 (2022).
38. Feng, J. et al. CityBench: evaluating the capabilities of large language models for urban tasks. *Proc. 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5413–5424 (ACM, 2025).
39. Xuan, W. et al. DynamicVL: benchmarking multimodal large language models for dynamic city understanding. In *The 39th Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, https://neurips.cc/virtual/2025/loc/san-diego/poster/121371 (Curran Associates, 2025).
40. Zhan, Y., Xiong, Z. & Yuan, Y. Skyeyegpt: unifying remote sensing vision–language tasks via instruction tuning with large language model. *ISPRS J. Photogramm. Remote Sens.* **221**, 64–77 (2025).
41. Yuan, Y., Wang, H., Ding, J., Jin, D. & Li, Y. Learning to simulate daily activities via modeling dynamic human needs. In *Proc. ACM Web Conference 2023*, 906–916 (Association for Computing Machinery, 2023).
42. Zhang, Y., Ma, R., Zhang, X. & Li, Y. Perceiving urban inequality from imagery using visual language models with chain-of-thought reasoning. In *Proc. ACM on Web Conference 2025*, 5342–5351 (Association for Computing Machinery, 2025).
43. Paolanti, M. et al. Ethical framework to assess and quantify the trustworthiness of artificial intelligence techniques: application case in remote sensing. *Remote Sens.* **16**, 4529 (2024).
44. Koldasbayeva, D. et al. Challenges in data-driven geospatial modeling for environmental research and practice. *Nat. Commun.* **15**, 10700 (2024).
45. Manvi, R., Khanna, S., Burke, M., Lobell, D. B. & Ermon, S. Large language models are geographically biased. In *Proc. 41st Int. Conf. Mach. Learn. 235*, 34654–34669 (PMLR, 2024).
46. Hall, M. et al. Towards geographic inclusion in the evaluation of text-to-image models. In *Proc. 2024 ACM Conference on Fairness, Accountability, and Transparency*, 585–601 (Association for Computing Machinery, 2024).
47. Ma, Y., Chen, S., Ermon, S. & Lobell, D. B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **301**, 113924 (2024).
48. Cheng, Y. et al. Exploring large language model based intelligent agents: definitions, methods, and prospects. Preprint at https://arxiv.org/abs/2401.03428 (2024).
49. Sun, Y. Q. et al. Can AI weather models predict out-of-distribution gray swan tropical cyclones? *Proc. Natl Acad. Sci. USA* **122**, e2420914122 (2025).
50. Papyshev, G. & Yarime, M. Exploring city digital twins as policy tools: a task-based approach to generating synthetic data on urban mobility. *Data Policy* **3**, e16 (2021).
51. Hofmann, J. & Schüttrumpf, H. Floodgan: using deep adversarial learning to predict pluvial flooding in real time. *Water* **13**, 2255 (2021).
52. Camps-Valls, G. et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nat. Commun.* **16**, 1919 (2025).
53. Zeng, J. et al. Opencarbon: a contrastive learning-based cross-modality neural approach for high-resolution carbon emission prediction using open data. In *Proc. 34th International Joint Conference on Artificial Intelligence* 9999–10007 (IJCAI, 2025).

54. Gao, J. et al. GenFusion: crafting future urban building layouts via diffusion model and genetic algorithm. *Inf. Geogr.* **1**, 100015 (2025).
55. Li, Y. et al. Generative models in decision making: a survey. Preprint at https://arxiv.org/abs/2502.17100 (2025).
56. Ding, W., Chen, B., Xu, M. & Zhao, D. Learning to collide: an adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2243–2250 (IEEE, 2020).
57. Yuan, Y., Ding, J., Wang, H. & Jin, D. Generating daily activities with need dynamics. *ACM Trans. Intell. Syst. Technol.* **15**, 1–28 (2024).
58. Yuan, Y., Ding, J., Jin, D. & Li, Y. Learning the complexity of urban mobility with deep generative network. *PNAS Nexus* **4**, pgaf081 (2025).
59. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology* (eds Follmer, S. et al.) 1–22 (ACM, 2023).
60. Piao, J. et al. AgentSociety: large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. Preprint at https://arxiv.org/abs/2502.08691 (2025).
61. Tao, Y., Viberg, O., Baker, R. S. & Kizilcec, R. F. Cultural bias and cultural alignment of large language models. *PNAS Nexus* **3**, pgae346 (2024).
62. Park, J. S. et al. Social simulacra: creating populated prototypes for social computing systems. In *Proc. 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18 (Association for Computing Machinery, 2022).
63. Yan, Y. et al. Opencity: a scalable platform to simulate urban activities with massive LLM agents. Preprint at https://arxiv.org/abs/2410.21286 (2024).
64. Yang, Z. et al. Oasis: open agent social interaction simulations with one million agents. Preprint at https://arxiv.org/abs/2411.11581 (2024).
65. Chopra, A., Kumar, S., Kuru, N. G., Raskar, R. & Quera-Bofarull, A. On the limits of agency in agent-based models. In *Proc. 24th International Conference on Autonomous Agents and Multiagent Systems*, 500–509 (International Foundation for Autonomous Agents and Multiagent Systems, 2025).
66. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021).
67. Kashinath, K. et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Trans. R. Soc. A* **379**, 20200093 (2021).
68. Kapoor, S. et al. Reforms: consensus-based recommendations for machine-learning-based science. *Sci. Adv.* **10**, eadk3452 (2024).
69. Xu, H. et al. Leveraging generative ai for urban digital twins: a scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *Urban Inform.* **3**, 29 (2024).
70. Zheng, Y. et al. A survey of machine learning for urban decision making: applications in planning, transportation, and healthcare. *ACM Comput. Surv.* **57**, 1–41 (2024).
71. Ravuri, S. et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672–677 (2021).
72. Otal, H. T., Stern, E. & Canbaz, M. A. LLM-assisted crisis management: building advanced LLM platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence*, 851–859 (IEEE, 2024).
73. Zhang, G., Yu, Z., Jin, D. & Li, Y. Physics-infused machine learning for crowd simulation. In *Proc. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2439–2449 (Association for Computing Machinery, 2022).
74. Chen, L. et al. Strategic COVID-19 vaccine distribution can simultaneously elevate social utility and equity. *Nat. Hum. Behav.* **6**, 1503–1514 (2022).
75. Zhang, J., Ao, W., Yan, J., Jin, D. & Li, Y. A GPU-accelerated large-scale simulator for transportation system optimization benchmarking. Preprint at https://arxiv.org/abs/2406.10661 (2024).
76. Zhang, H. et al. Cityflow: a multi-agent reinforcement learning environment for large scale city traffic scenario. In *The World Wide Web Conference*, 3620–3624 (Association for Computing Machinery, 2019).
77. Chen, H. et al. DiffLight: a partial rewards conditioned diffusion model for traffic signal control with missing data. In *Proc. 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 123353–123378 (NeurIPS, 2024).
78. Hu, X. et al. Dual-stage flows-based generative modeling for traceable urban planning. In *Proc. 2024 SIAM International Conference on Data Mining* (eds Shekhar, S. et al.) 370–378 (SIAM, 2024).
79. Busuioc, M. Accountable artificial intelligence: holding algorithms to account. *Public Adm. Rev.* **81**, 825–836 (2021).
80. Green, B. The flaws of policies requiring human oversight of government algorithms. *Comput. Law Secur. Rev.* **45**, 105681 (2022).
81. Alon-Barkat, S. & Busuioc, M. Human–AI interactions in public sector decision making: automation bias and selective adherence to algorithmic advice. *J. Public Adm. Res. Theory* **33**, 153–169 (2023).
82. Raji, I. D. et al. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proc. 2020 Conference on Fairness, Accountability, and Transparency*, 33–44 (Association for Computing Machinery, 2020).
83. Suresh, H. & Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proc. 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9 (Association for Computing Machinery, 2021).
84. Chattopadhyay, A. & Hassanzadeh, P. Long-term instabilities of deep learning-based digital twins of the climate system: the cause and a solution. Preprint at https://arxiv.org/abs/2304.07029 (2023).
85. Baykurt, B. Algorithmic accountability in US cities: transparency, impact, and political economy. *Big Data Soc.* **9**, 20539517221115426 (2022).
86. Rong, C., Ding, J., Liu, Y. & Li, Y. A large-scale dataset and benchmark for commuting origin–destination flow generation. In *The 13th International Conference on Learning Representations*, https://neurips.cc/virtual/2025/loc/san-diego/poster/121846 (NeurIPS, 2025).
87. Feng, J. et al. Citygpt: empowering urban spatial cognition of large language models. In *Proc. 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 591–602 (Association for Computing Machinery, 2025).
88. Feng, J., Wang, S., Liu, T., Xi, Y. & Li, Y. UrbanLLaVA: a multi-modal large language model for urban intelligence with spatial reasoning and understanding. In *Proc. IEEE/CVF International Conference on Computer Vision*, 6209–6219 (IEEE, 2025).
89. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
90. United Nations, Department of Economic and Social Affairs, Population Division. World urbanization prospects 2018: highlights. *un.org* https://population.un.org/wup/assets/WUP2018-Highlights.pdf (2019).
91. Bettencourt, L. M. & Marchio, N. Infrastructure deficits and informal settlements in sub-Saharan Africa. *Nature* **645**, 399–406 (2025).
92. Kuffer, M., Pfeffer, K. & Sliuzas, R. Slums from space—15 years of slum mapping using remote sensing. *Remote Sens.* **8**, 455 (2016).

93. da Silva, J. P., Rodrigues-Jr, J. F. & de Albuquerque, J. P. On the power of CNNs to detect slums in Brazil. *Comput. Environ. Urban Syst.* **121**, 102306 (2025).
94. Brelsford, C., Martin, T., Hand, J. & Bettencourt, L. M. A. Toward cities without slums: topology and the spatial evolution of neighborhoods. *Sci. Adv.* **4**, eaar4644 (2018).
95. Gevaert, C. M., Buunk, T. & Van Den Homberg, M. J. Auditing geospatial datasets for biases: using global building datasets for disaster risk management. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **17**, 12579–12590 (2024).
96. Chamberlain, H. R. et al. Building footprint data for countries in Africa: to what extent are existing data products comparable?. *Comput. Environ. Urban Syst.* **110**, 102104 (2024).
97. Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J. & Zipf, A. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in openstreetmap. *Nat. Commun.* **14**, 3985 (2023).
98. Lacoste, A. et al. GEO-Bench: toward foundation models for Earth monitoring. *Adv Neural Inf. Proc. Syst.* **36**, 51080–51093 (2023).
99. Hong, D. et al. Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sens. Environ.* **299**, 113856 (2023).
100. Vergara-Perucich, J. F. Ai-driven deconstruction of urban regulatory frameworks: unveiling social sustainability gaps in santiago's communal zoning. *Urban Sci.* **9**, 186 (2025).
101. Xie, Y. et al. Fairness by where: a statistically-robust and model-agnostic bi-level learning framework. *Proc. AAAI Conf. Artif. Intell.* **36**, 12208–12216 (2022).
102. Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
103. Liu, Y., James, J. Q., Kang, J., Niyato, D. & Zhang, S. Privacy-preserving traffic flow prediction: a federated learning approach. *IEEE Internet Things J.* **7**, 7751–7763 (2020).
104. Li, X. & Dang, A. The disruptive effect of ai on urban planning. *Nat. Cities* **2**, 568–570 (2025).
105. Sanchez, T. W., Brenman, M. & Ye, X. The ethical concerns of artificial intelligence in urban planning. *J. Am. Plann. Assoc.* **91**, 294–307 (2025).
106. Choi, S. L. et al. *eGridGPT: Trustworthy AI in the Control Room. NREL/TP-5D00-87440* (National Renewable Energy Laboratory, 2024).

Author contributions

Y. Li, Q.R.W., E.M., Y.Y., Y. Liu, D.F., P.G., L.M.A.B. and M.B. conceived the research framework and initial design of the study. Y.Z., F.X. and Y. Li wrote the original draft of the manuscript. F.X. and Y. Li formulated the conceptual framework regarding the emergence of generative AI. M.B. contributed to the historical review of generative models. Y.Y., B.C., Y. Liu and P.G. developed the urban sensing section. Q.R.W., E.M. and D.F. developed the urban modeling section. H.W. developed the decision-making section. Y.Y. and L.M.A.B. contributed to the challenges and future directions sections. Y.Z. integrated inputs from all co-authors and managed the revision process. All authors provided critical revisions and approved the final version of the paper.

Funding

Y. Li discloses support for the research of this work from The National Key Research and Development Program of China (grant number 2024YFC3307605) and The National Natural Science Foundation of China (grant number 62472241).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yong Li.

Peer review information *Nature Cities* thanks Riccardo Di Clemente, Orlando Woods and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.