https://doi.org/10.1007/sxxxxx-yyy-zzzz-1

RESEARCH ARTICLE

Toward City Foundation Models: From Task-Specific Spatial Approaches to Urban General Intelligence

Gao Cong

College of Computing and Data Science, Nanyang Technological University, Singapore

Received month dd, yyyy; accepted month dd, yyyy

E-mail: gaocong@ntu.edu.sg.

Abstract

Modern cities produce vast streams of multimodal data closely related to urban structure, dynamics, and human activities. These data have powered many specialized geospatial models for urban tasks such as spatial search, POI recommendation, and land-use inference, yet such approaches rely heavily on domain expertise for model design and label-intensive supervision, lacking cross-task generalization. Recent advances in self-supervised learning and foundation models, particularly large language models, present new opportunities for developing general-purpose geospatial intelligence. This Perspective synthesizes the methodological evolution from supervised spatial learning to foundation-model paradigms and examines how these advances can enable unified, data-driven urban intelligence. We highlight emerging efforts toward spatially grounded and multimodal City Foundation Models, which integrate spatial, temporal, and semantic modalities to support perception, prediction, retrieval, reasoning, and decision-making within a single framework. Finally, we outline key challenges in multimodal integration, spatial reasoning, cross-city generalization, agentic planning, and ethical governance, and discuss how addressing these challenges can pave the way for next-generation intelligent urban systems.

1 Introduction

The modern city is a dense network of sensors, infrastructures, and human activities, providing rich multimodal data such as points of interest (POIs), human mobility traces, road networks, building footprints, satellite imagery, street-view images, and textual descriptions. These data collectively capture how people move, interact, and experience urban spaces, driving *Urban Computing*, a field that leverages data-driven methods to address practical challenges such as spatial keyword queries [1–3], POI recommendations [4], socio-economic predictions [5], and traffic forecasting. However, over the past two decades, most of these methods have been tailored to individual tasks, relying heavily on task- and city-specific design, dedicated labeled datasets for supervision, and domain expertise for result interpretation. The learned models and representations cannot be readily adapted to other tasks or cities, leading to additional training and design efforts.

In the broader field of artificial intelligence, however, task-specific model design paradigms have been fundamentally reshaped by the rise of foundation models. Particularly, the remarkable success of large language models (LLMs) most vividly demonstrates this paradigm shift. Through large-scale self-supervised training, LLMs learn to organize and reuse knowledge across diverse tasks and domains, profoundly influencing how models are designed, trained, and evaluated. Their success naturally raises a new question for the urban computing community: *Can we build foundation models for cities that can understand, represent, reason about, and interact with urban environments across tasks and modalities?*

Recent studies offer encouraging evidence that such models are within reach. For example, self-supervised approaches on road networks demonstrate the feasibility of learning transferable geospatial embeddings [5]. Evaluations of general LLMs on spatial tasks reveal surprising semantic competence and fundamental geometric limitations, highlighting the need for explicit spatial grounding [6]. Early explorations in multimodal alignment, structured spatial grounding, and tool-augmented reasoning suggest that the capability of foundation models can extend naturally to urban contexts.

This Perspective synthesizes recent developments from our research team and proposes a research agenda toward *City Foundation Models (City FMs)*, a new class of models that integrate multimodal perception, transferable spatial representation, spatial reasoning, and urban decision functionalities. We trace the methodological evolution of geospatial AI, examine opportunities and limitations of foundation models for geospatial tasks, review emerging efforts toward spatially grounded City FMs, and outline open challenges that must be addressed to realize this vision.

2 Paradigm Shift in Geospatial AI: From Task-Specific Supervised Models to Self-Supervised Representation Learning

The trajectory of geospatial AI has long been shaped by methodological choices. Traditionally, the dominant paradigm has relied on

supervised, task-specific models, each tailored to a single urban problem. While highly successful for their intended tasks, such models produce specialized representations that transfer poorly across tasks or cities. Recently, the community has witnessed a transition toward *self-supervised representation learning*, which leverages large-scale unlabeled geospatial data to produce general-purpose spatial embeddings. This section summarizes this paradigm shift.

2.1 Traditional Task-Specific Supervised Models

Supervised models have made substantial contributions across multiple geospatial applications. To illustrate the traditional paradigm, we highlight three representative tasks. First, in geospatial information retrieval (a.k.a. spatial keyword query), early work combines textual relevance with spatial proximity to retrieve relevant spatial objects for a given query [1]. More recently, deep models like DrW [2] and GeoBloom [3] learn semantic relevance between queries and spatial objects from query–object pairs. These methods are task-specific and thus have limited generalizability. Second, in POI recommendation, existing models learn representations of users and POIs from historical user-POI interaction data [4]. Yet, the learned embeddings are specialized for recommendation and do not transfer to other geospatial tasks. Third, geospatial relationship extraction focuses on predicting spatial relations such as containment between two spatial objects, enabling the construction of geospatial knowledge graphs [7]. However, these models remain task-specific and cannot generalize to other tasks.

Limitations of the Traditional Paradigm The three examples illustrate that task-specific supervised approaches face common limitations: (1) Task-bounded embeddings: learned representations are optimized to preserve the particular signals of a single task, and they tend to discard information that is irrelevant to that task but crucial for others; (2) High annotation costs: labeled data are expensive, scarce, and often highly city-specific; and (3) Limited scalability: training is constrained by the coverage of labels and cannot effectively scale to massive urban datasets. These methodological bottlenecks motivate a move toward more unified and transferable approaches.

2.2 Emergence of Self-Supervised Spatial Representation Learning

The limitations of task-specific spatial models have motivated a shift toward self-supervised representation learning for geospatial entities. Unlike supervised approaches that require labeled data for each application, self-supervised methods exploit the intrinsic structure of raw spatial data, such as road networks, building footprints and mobility traces, and convert them into training signals. In practice, models can be pre-trained by predicting missing components, modeling local-to-global context, or contrasting co-occurring entities and patterns, so that embeddings reflect both connectivity (structure) and usage (function) in urban environments. Because supervision is derived from the data itself rather than task annotations, the learned representations are more label-efficient and transfer more readily across cities and downstream tasks. Below, we highlight two representative lines of work.

(1) Road network representation learning: Road networks are the backbone of urban mobility. Chen et al. [8] introduced a self-supervised framework to learn embeddings of road segments by modeling topological connectivity, traffic dynamics, and travel semantics from trajectories. Recently, Zhou et al. [9] further advanced this line of work using a contrastive learning framework that jointly models spatial proximity and geographic configuration similarity. The learned representations support multiple tasks, including travel time estimation, traffic speed prediction, trajectory similarity search, and anomaly detection, and often perform competitively against task-specific models.

(2) Region-level representation learning: Beyond roads, a city can be represented as regions that differ in land use, built form, and human activity. Li et al. [5] proposed a Transformer-based model to learn region embeddings from building footprints and road-defined spatial partitions. A two-level contrastive framework aligns representations across scales, encouraging consistency between building-group embeddings and region embeddings. These embeddings support a range of region-level tasks such as land-use inference, population density prediction, region clustering that matches zoning maps, and urban function discovery with improved interpretability.

Summary Compared with task-specific supervised models, self-supervised representations are easier to reuse across tasks and require far less manual labeling. More broadly, they provide a reusable representation layer that supports many urban analytics tasks, and they serve as a natural starting point for broader geospatial foundation-model efforts, including City Foundation Models.

3 From Geospatial Applications of Foundation Models to City Foundation Models (City FMs)

Following recent advances in self-supervised spatial representation learning, researchers have begun to explore how large foundation models can be extended to geospatial domains. Mai et al. [6] conducted one of the first systematic evaluations in this direction, revealing both promising capabilities and fundamental limitations when general foundation models are applied without spatial grounding. These observations highlight the need for spatially grounded and urban-specific foundation models, which we term City Foundation Models (City FMs).

3.1 Directly Applying General Foundation Models to Geospatial Tasks

General-purpose large language models (LLMs) and vision-language models (VLMs) contain substantial geospatial knowledge learned from web-scale corpora, and can be applied to some geospatial tasks in a zero-shot or few-shot manner. LLMs are particularly effective when the task is dominated by semantic or contextual understanding, such as toponym recognition, POI categorization, and location-description classification [6]. Complementing LLMs, VLMs provide an additional interface by directly operating on map-related images, where the model must read map texts and symbols and interpret marked locations. This suggests that general foundation models can provide strong baselines for geospatial applications that rely primarily on language understanding or visually explicit cues, without requiring task-specific model design or city-specific labeling.

Limitations General foundation models still face severe difficulties when tasks require explicit spatial grounding and reliable spatial reasoning. Dihan et al. [10] show that even strong LLMs often fail

at simple geometric computations and structured spatial reasoning, including straight-line distance estimation, cardinal-direction queries, and counting. Map-centric visual reasoning remains challenging for VLMs as well. Their performance degrades when map views become denser and more fine-grained, and common failure modes include POI miscounting and confusion among closely distributed spatial objects. These findings motivate spatially grounded, urban-specific City Foundation Models (City FMs).

3.2 Attempts Toward Spatially Grounded Foundation Models

Specialized approaches have emerged to combine the strengths of foundation models with the structured, spatially grounded nature of urban environments. These efforts include: (1) In CityFM [11], a self-supervised foundation model trained entirely on OpenStreetMap data to learn unified multimodal (spatial, visual, textual) representations for geospatial entities, a data-centric pretraining strategy to achieve competitive or superior performance across diverse downstream urban tasks; (2) In UrbanClip [12], an urban taxonomy and a set of urban-function prompt templates are designed to integrate domain knowledge with vision–language alignment for understanding the urban scene; (3) Techniques [13–15] finetune existing foundation models with domain data for spatial tasks. In LAMP [13], a language model is created by fine-tuning a pre-trained model on city-specific data to enable accurate answering of spatial object retrieval while minimizing hallucinations. QTMob [14] extends LLM reasoning to mobility tasks through introducing semantic location tokenization, while NextLocMoE [15] adopts a mixture-of-experts design to encode multi-functional location semantics and personalized mobility patterns; and (4) UrbanLLM [16], an agentic system, integrates a finetuned LLM with a zoo of spatio-temporal predictors and a planning pipeline capable of tool invocation, task decomposition, and result synthesis for urban planning and decision support. Together, these works show that LLMs can develop spatial competence when grounded in structured geospatial data.

Summary Together, evaluations of general foundation models and the initial specialized approaches point to a common direction. Foundation models provide strong semantic capabilities, yet reliable geospatial use typically requires spatial grounding, multimodal urban data, and tool support for precise computation. Existing attempts, including structural pretraining, multimodal alignment, prompt engineering, fine-tuned spatial LLMs, and tool-augmented agents, lay the groundwork for City Foundation Models capable of supporting perception, prediction, reasoning, and decision-making across diverse urban tasks.

4 Open Research Problems for City Foundation Models

Despite rapid progress in geospatial representation learning and the growing influence of language models, the development of a robust City Foundation Model (City FM) remains in its early stages. Building such a model raises several fundamental technical and societal challenges that differ from those addressed by existing LLMs and VLMs. We outline the key open research problems below.

- Defining capability scope and identifying new applications.

A key problem is to define the capability scope of a City FM—-i.e., the families of urban applications it should support—-and to identify novel applications it enables. This is not a return to task-specific modeling, but to specify transferable competencies that guide data pipelines, training, and benchmarks. The challenge is to make these capabilities comparable across cities while allowing city-specific data and constraints. A practical direction is to specify cross-city capabilities, such as cross-modal perception and retrieval, unified forecasting across signals and regions, and interactive urban planning, and to evaluate them with diverse application suites rather than one fixed task.

- Multimodal urban data integration and scalable model training.

A second bottleneck is how to train at city scale when urban data span maps, imagery, sensor streams, trajectories, point clouds, and spatial text, each with different semantics, resolutions, coverage patterns, and coordinate systems. Core questions include whether to adapt existing LLMs and VLMs with explicit geospatial grounding or to train models from scratch using geospatial corpora such as OpenStreetMap, remote-sensing archives, and large mobility datasets. Another challenge is architectural: city data contain network topology and spatial hierarchy, and it remains unclear how to encode connectivity, geometry, and multi-scale structure in Transformer-based models while preserving metric faithfulness. These training choices largely determine the upper limit of the model's downstream capabilities.

- Spatially grounded reasoning and verifiable performance gains.

Even when foundation models provide strong semantic understanding, many geospatial tasks require explicit spatial reasoning to be correct and reliable. Open challenges include reasoning about distances and travel times, handling connectivity and containment, and executing spatial operators such as routing, buffering, and neighborhood search. Current models also suffer from spatial factuality issues, including hallucinated POI entries, incorrect locations or coordinates, and inconsistent relations. A central question is how to integrate spatial constraints into model predictions so that reasoning steps become checkable, and performance improvements come from verifiable spatial operations rather than from uncontrolled text generation.

- Agentic capabilities and integration with existing spatial systems.

City-scale decision-making typically follows multi-step workflows and depends on external geospatial systems, such as routing engines, spatial databases, GIS operators, and mobility predictors. City FMs therefore require agentic capabilities for task decomposition, tool invocation, and constraint checking, and they must produce auditable intermediate states that can be inspected by domain experts. Early agentic urban LLMs illustrate promise [16], but robust deployment demands reliable tool selection across heterogeneous geospatial systems, seamless interoperability with existing spatial infrastructures (including semantic data governance), principled handling of uncertainty and missing city data, and explicit exception-handling mechanisms for cases in which spatial constraints cannot be satisfied.

- Benchmark datasets and evaluation protocols for City FMs.

Reliable evaluation is essential for measuring real progress toward general-purpose City FMs. At present, however, results are difficult to compare because the geospatial domain lacks large-scale, standardized, and cross-city benchmarks. A comprehensive evaluation should cover the full range of City FM capabilities: spatial semantics, metric and

topological reasoning, multimodal grounding across text, image, POI, trajectory, and point cloud, as well as urban prediction and agentic planning tasks that involve tool use. Benchmarks also need to be reproducible across cities while capturing city-specific differences, and should assess not only final outputs but also the correctness of intermediate spatial operations and tool interactions.

- Ethical and societal considerations.

Finally, City FMs will influence high-stakes decisions in transportation, housing, mobility, and public services. Open problems include fairness across regions and demographic groups, privacy protection for mobility traces and geo-tagged data, transparency and interpretability for both model outputs and tool actions, and responsible governance for data use and deployment. In this domain, safeguards must be built into data collection, training objectives, evaluation protocols, and deployment practices, with clear accountability for failures and misuse.

5 Conclusions

Geospatial AI is entering a new stage of integration and generalization. The long-standing paradigm of supervised, task-specific models is being replaced by self-supervised representation learning, which extracts transferable spatial knowledge directly from large urban datasets. Building on this foundation, large language and vision–language models bring powerful semantic reasoning and multimodal understanding, suggesting that general-purpose urban intelligence is within reach. Early progress in spatial grounding, multimodal alignment, prompt engineering, finetuning, and tool-augmented reasoning already shows that these capabilities can be combined into unified urban frameworks.

City Foundation Models represent the next step in this trajectory. They aim to unify perception, prediction, retrieval, reasoning, and decision-making across urban tasks within a single framework. Achieving this goal requires advances in multimodal fusion, spatial reasoning, cross-city generalization, agentic interaction, evaluation, and governance. Addressing these challenges will establish the methodological and ethical foundations for a new generation of geospatial AI systems that can reason about, plan for, and ultimately improve the functioning of modern cities.

Acknowledgement

We thank Yi Li for insightful discussions and proofreading the manuscript. Due to page limits, we were unable to include additional references.

References

[1] Cong G, Jensen C S, Wu D. Efficient retrieval of the top-k most relevant spatial web objects. Proc. VLDB Endow., 2009, 2(1): 337–348

[2] Liu S, Cong G, Feng K, Gu W, Zhang F. Effectiveness perspectives and a deep relevance model for spatial keyword queries. Proc. ACM Manag. Data, 2023, 1(1): 11:1–11:25

[3] Li Y, Cong G. Geobloom: Revisiting lightweight models for geographic information retrieval. Proc. VLDB Endow., 2025, 18(5): 1348–1361

[4] Yuan Q, Cong G, Ma Z, Sun A, Magnenat-Thalmann N. Time-aware point-of-interest recommendation. In: Proc. ACM SIGIR. 2013, 363–372

[5] Li Y, Huang W, Cong G, Wang H, Wang Z. Urban region representation learning with openstreetmap building footprints. In: Proc. ACM SIGKDD. 2023, 1363–1373

[6] Mai G, Huang W, Sun J, Song S, Mishra D, Liu N, Gao S, Liu T, Cong G, Hu Y, Cundy C, Li Z, Zhu R, Lao N. On the opportunities and challenges of foundation models for geospatial ai. ACM Trans. Spatial Algorithms Syst., 2024, 10(2): 11

[7] Balsebre P, Yao D, Cong G, Huang W, Hai Z. Mining geospatial relationships from text. Proc. ACM Manag. Data, 2023, 1(1): 93:1–93:26

[8] Chen Y, Li X, Cong G, Bao Z, Long C, Liu Y, Chandran A K, Ellison R. Robust road network representation learning: When traffic patterns meet traveling semantics. In: Proc. ACM CIKM. 2021, 211–220

[9] Zhou H, Huang W, Chen Y, He T, Cong G, Ong Y S. Road network representation learning with the third law of geography. In: Proc. NeurIPS. 2024, 11789–11813

[10] Dihan M L, Hassan M T, PARVEZ M T, Hasan M H, Alam M A, Cheema M A, Ali M E, Parvez M R. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. In: Proc. ICML. 2025

[11] Balsebre P, Huang W, Cong G, Li Y. City foundation models for learning general purpose representations from openstreetmap. In: Proc. ACM CIKM. 2024, 87–97

[12] Huang W, Wang J, Cong G. Zero-shot urban function inference with street-view images through prompting a pretrained vision–language model. Int. J. Geogr. Inf. Sci., 2024, 38(7): 1414–1442

[13] Balsebre P, Huang W, Cong G. Lamp: A language model on the map. CoRR, 2024, abs/2403.09059

[14] Chen Y, Tao Y, Jiang Y, Liu S, Yu H, Cong G. Enhancing large language models for mobility analytics with semantic location tokenization. In: Proc. ACM SIGKDD. 2025, 262–273

[15] Liu S, Cao N, Chen Y, Jiang Y, Cong G. Mixture-of-experts for personalized and semantic-aware next location prediction. CoRR, 2025, abs/2505.24597

[16] Jiang Y, Chao Q, Chen Y, Li X, Liu S, Cong G. Urbanllm: Autonomous urban activity planning and management with large language models. In: Proc. EMNLP Findings. 2024, 1810–1825