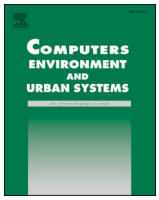
Computers, Environment and Urban Systems 126 (2026) 102416

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Fine-grained urban land use simulation: Integrating spatial dynamic modeling with a pre-trained vision-language model☆

Zipan Cai [a,b], Andrew Karvonen [c], Cong Cong [d], Weiming Huang [e,*]

[a] Department of Urban Planning, School of Architecture, Southeast University, Nanjing, China
[b] Department of Sustainable Development, Environmental Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden
[c] Department of Architecture and the Built Environment, Lund University, Lund, Sweden
[d] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA
[e] School of Geography, University of Leeds, Leeds, United Kingdom

ARTICLE INFO

Keywords:
Land use change
Vision-language models
Foundation models
Spatial dynamic modeling
Street view images

ABSTRACT

Accurate prediction of urban land use changes at fine spatial scales is essential for developing healthy and sustainable cities, yet traditional simulation models struggle to capture local dynamics due to limited availability of fine-grained data and insufficient complexity in modeling urban systems. To address these limitations, we propose a novel approach that leverages advances in pre-trained vision-language foundation models combined with spatial dynamic modeling to forecast detailed urban land use patterns. Specifically, we collected a spatially dense collection of street view images (SVIs) throughout Shenzhen, China, and applied UrbanCLIP, a specialized vision-language prompting framework, to perform zero-shot inference of urban land use directly from images without labeled datasets and model retraining. The resulting fine-grained classifications delineate eight distinct urban land use types, producing a detailed urban functional map. These high-resolution patterns were then integrated into a spatial dynamic model enhanced by polynomial regression to simulate urban evolution toward 2035. This approach effectively captures neighborhood influences, socioeconomic drivers, and urban planning policies. Our simulation provides actionable insights for sustainable development in Shenzhen by identifying areas for balanced growth, targeted infrastructure investments, and ecological preservation. Compared to conventional methods, our methodology significantly improves predictive accuracy and spatial granularity. By incorporating foundation models, our approach addresses traditional data constraints, offering scalable and robust tools for informed urban governance and decision-making.

1. Introduction

Urbanization is transforming societies worldwide, with rapid city expansion driven by population growth and migration (Cohen, 2006). This growth poses challenges for sustainable planning, requiring a balance among economic, environmental, and social objectives (Hariram et al., 2023). Predictive urban modeling has thus become critical (Batty, 2024), which offers insights into potential future urban land use patterns. By simulating growth and transformations, these models help policymakers make informed (Murphy, 2012), data-driven decisions for more resilient cities (Karvonen et al., 2021).

Simulating future land use patterns is fundamental to predict the likely futures of our cities (Pijanowski et al., 2002). Traditional models, such as cellular automata (CA) and agent-based models (ABMs), have long been employed to forecast urban growth and land use changes. Although both CA and ABM frameworks are capable of simulating fine-scale dynamics, their application at large spatial scales often face practical constraints, such as computational load, data availability, and calibration complexity, which can result in spatially coarser implementations in practice (An et al., 2021). In addition, these models typically rely heavily on authoritative land use datasets to inform transition rules or agent behaviors (Clarke, 2021). Although such data sometimes can be acquired from local planning authorities, access is not always guaranteed. To this end, numerous studies have utilized different geospatial data sources and machine learning methods to derive land use information (Cai, 2025). However, such processes typically require

☆ This article is part of a Special issue entitled: 'Digital Twins and AI for Cities' published in Computers, Environment and Urban Systems.

* Corresponding author.
E-mail address: W.Huang@leeds.ac.uk (W. Huang).

https://doi.org/10.1016/j.compenvurbsys.2026.102416
Received 14 May 2025; Received in revised form 7 January 2026; Accepted 13 February 2026
Available online 26 February 2026

extensively labeled datasets for model training, which are large collections of annotated data that are costly and time-consuming to produce.

Fine-grained descriptions of urban land use are pivotal for the realization of nuanced and refined urban planning and simulations (Wen & Li, 2021; Wu et al., 2024). In this regard, the recent proliferation of street view images (SVIs) and their use in various urban analyses have shed light on obtaining highly granular understandings of the urban environment. Specifically, SVIs can offer delineation of the built environment from a grounded perspective, which is particularly useful for sensing fine-grained (e.g., building-level) land use distributions (Kang et al., 2018). However, the heavy reliance on extensive labeled datasets for machine learning model training remains a prominent challenge. In fact, the use of SVIs further exacerbates the problem, as it is hardly possible to obtain ground truth labels at the single-SVI level (a small urban area captured in an individual image), which is at a finer scale than land parcels.

Recent advancements in computer vision and large-scale pre-trained foundation models usher in a promising way to tackle the challenge (Janowicz et al., 2025; Mai et al., 2024). With pre-trained vision-language models (VLMs), it becomes possible to infer fine-grained land use from SVIs in a zero-shot manner, i.e., with no ground truth labels and no further model training/tuning (Li et al., 2024). The Contrastive Language–Image Pre-training (CLIP) model is a representative model that has demonstrated remarkable performance in various image recognition tasks (Radford et al., 2021). CLIP is pre-trained on enormous amounts of image-text pairs, enabling it to be applied to zero-shot image classification, namely classifying images with no labeled samples and model training (finetuning). Specifically, the pre-trained CLIP can produce effective embeddings (vector representations) for both images and textual descriptions. Zero-shot image classification is achieved by simply comparing the (cosine) similarities between image and text embeddings. For example, given an image and the textual descriptions like “dog”, “plane”, and “building”, the cosine similarities between the image embedding and the text embeddings are derived, and the text label with the largest similarity metric is deemed to be the category of the image. In this context, Huang et al. (2024) has proposed a prompting framework UrbanCLIP, to effectively infer fine-grained urban land use from individual SVIs in a zero-shot manner. In this way, we can obtain fine-grained and current land use status in cities.

Despite the potential of VLMs (UrbanCLIP in this case), their use in predictive urban modeling remains rare, as most studies focus on static image classification or scene understanding rather than dynamic simulations of future urban changes (Ho et al., 2024; Liang et al., 2024). Integrating these models with spatial dynamic approaches offers a missed but promising opportunity: the detailed functional information from VLMs can substantially improve both the accuracy and granularity of forecasts of the urban fabric, i.e., the physical and functional layout of buildings, streets, and open spaces (Araldi & Fusco, 2019).

For land use simulation, the complexities of urban growth are influenced not only by physical factors but also by socioeconomic drivers and policy constraints (Chang et al., 2020; Wilkerson et al., 2018). Traditional models do not adequately account for these multifaceted influences, particularly at fine spatial scales (Bahers et al., 2022; Herold et al., 2005). Incorporating detailed urban land use patterns derived from VLMs into spatial dynamic frameworks can enable more realistic simulations that reflect the intricate interplay among various factors, e.g., population density (He et al., 2020), transportation accessibility (Rode et al., 2017), socioeconomic composition (Cavicchia & Cucca, 2022), environmental constraints (Li et al., 2020), and policy-driven zoning regulations (Marey et al., 2024) that shape urban development. Achieving such fine-grained simulations inherently requires careful integration of machine learning techniques within the modeling frameworks to calibrate transition probabilities and capture diverse urban growth drivers (Wang et al., 2022).

Given these considerations, the primary aim of this research is to develop a novel methodology to leverage both pre-trained VLMs and spatial dynamic modeling (SDM) to predict fine-grained urban land use changes. Specifically, we:

1. Leverage UrbanCLIP to classify land use patterns from SVIs in a fine granularity (~50 m), and in a zero-shot manner, reducing the reliance on labeled datasets.
2. Incorporate these classifications into a SDM enhanced with machine learning to simulate future urban changes under neighborhood influences, growth drivers, and policy constraints.
3. Evaluate predictive performance and assess real-world planning implications, offering actionable insights for policymakers.

By addressing traditional modeling limitations through the integration of VLMs and SDM, this study contributes to predictive urban planning. The proposed approach enables planners to detect nuanced variations in urban functions and anticipate urban development with greater precision, ultimately supporting more sustainable and resilient urban futures. In the following sections, we present related works, outline the methodology, present simulation results, and discuss the key findings. We emphasize how integrating advanced machine learning techniques with SDM addresses persistent challenges in predictive urban planning, elevating the detail and accuracy of planning assessments and facilitating sustainable urban transformations.

2. Related works

2.1. Fine-grained urban land use prediction

Fine-grained urban land use prediction has gained great popularity in recent urban studies, especially with the evolution of machine learning–enhanced SDM (Wang et al., 2022). High spatial resolution is increasingly crucial for cities undergoing rapid urbanization, where minor and micro-level changes can reshape urban form and function (Shukla et al., 2021; Xia et al., 2019). For example, Xu and Zhao (2024) emphasized the need for fine-grained spatial analysis in China, revealing detailed urban growth patterns that traditional blue-green-gray landscape studies overlooked. Similarly, Yin et al. (2021) developed a fine-resolution population grid by integrating remote sensing and GIS data, enabling more precise urban analysis and better resource allocation.

Efforts to predict urban land use changes at the building level have shown potential but face challenges in practice. For instance, several studies have relied on remote sensing images to analyze building-level land use changes of built and unbuilt areas (Liu et al., 2017; Shi et al., 2019; Wu et al., 2022). While these approaches provide valuable insights, they often fail to capture granular and ground-level semantics, including functional nuances and human-scale interactions, limiting their effectiveness for fine-grained land use modeling.

Incorporating SVIs provides a ground-level perspective of urban environments, enhancing fine-grained modeling capabilities. Chen and Biljecki (2023) employed SVIs to assess streetscape greenery at high spatial resolution, contributing to studies on urban livability and environmental quality. These detailed visual sources enable models to capture micro-scale interactions and variations in urban spaces (e.g. signage, storefronts, and informal activities) that remote sensing methods cannot fully represent.

In this study, we tackle the long-standing spatial granularity challenge by using SVIs and the zero-shot inference capability from VLMs. The fine-grained land use patterns can then be fed into a spatial dynamic framework, to represent local influences, growth drivers, and policy constraints more accurately.

2.2. Machine learning enhancements in SDM

SDM has long been central to simulating urban growth and land use changes, with CA models being among the most widely used. Traditional CA approaches rely on predefined transition rules based on theoretical

assumptions, which can be too simplistic for the complex, nonlinear dynamics of urban environments (Cai, 2025). Consequently, there is growing interest in integrating machine learning techniques to calibrate rules and transition probabilities using observed data.

For example, Meresa et al. (2024) introduced a neural network-based CA model that employs machine learning to derive transition rules from empirical data. Their approach improved the model's ability to simulate urban development patterns by capturing intricate relationships between various urban growth factors. Similarly, Cai et al. (2023) utilized linear regression to estimate transition probabilities in a CA model, enhancing the simulation of urban growth in Swedish cities. These studies underscore the capability of machine learning algorithms to model nonlinear interaction among multiple variables influencing urban dynamics.

The integration of machine learning offers several distinct benefits. First, data-driven calibration allows transition rules and probabilities to be derived from observed data, capturing real-world dynamics more accurately (Cheng et al., 2023). Second, machine learning excels at handling the intricate relationship among variables such as population density, accessibility, socioeconomic factors, and environmental conditions, which are pivotal to understanding urban growth (Fan et al., 2023; Kutty et al., 2022). This capability enhances the flexibility and predictive power of the models, resulting in more accurate simulations of future urban changes.

A critical advancement facilitated by machine learning is the incorporation of socioeconomic factors and policy constraints into SDM. Silva and Wu (2012) emphasized the importance of integrating socioeconomic factors and policy constraints into spatial models. Our work aligns with this perspective by embedding machine learning–based calibration within SDM. By doing so, we explicitly account for multiple drivers of urban growth, e.g., population distribution, accessibility to urban functions and road networks, while reflecting policy decisions in the simulation outcomes.

2.3. Vision-language models for urban land use inference

The advent of pre-trained VLMs has reformed the landscape of various computer vision tasks, e.g., image classification (Radford et al., 2021), semantic segmentation (Zhou et al., 2023), and video clip retrieval (Luo et al., 2022). In the meantime, pioneering works have been implemented to use the zero-shot capability of CLIP for urban analysis, particularly by enabling the interpretation and classification of urban images (SVIs) without labeled samples. Nevertheless, applying CLIP to answer urban questions is not straightforward due to the significant gap between the general-purpose vision pre-training and the specialized urban questions. This gap in fine-grained urban land use inference was described in detail in Huang et al. (2024), where they observed that simply prompting CLIP with the raw urban function category names (e.g., residential and commercial) does not suffice. The first challenge is that CLIP is more capable of handling concrete concepts, and often fails to understand the abstract language phrases of urban function categories (Liao et al., 2023). The second obstacle is that CLIP can be easily misguided by common but interfering elements in real-world SVIs, such as vehicles, road surface, and street-side landscape such as trees. In this context, Huang et al. (2024) developed the UrbanCLIP framework to overcome the challenges, so that the inference capability for urban land use can be significantly improved.

In addition, there have been other attempts of using pre-trained VLMs for urban analysis. Wu et al. (2023) used CLIP for urban land use analysis with SVIs, and found that the incorporation of spatial context information (e.g., city names) can have a subtle influence on performance. Vivanco Cepeda et al. (2023) adapted CLIP for geo-localization of SVIs, in which a location encoder that continuously models the earth's surface was used to find the correct places where SVIs were taken.

Overall, the utilization and adaptation of pre-trained VLMs in urban analyses are still in a preliminary stage. The effectiveness of deriving fine-grained urban land use information without requiring any ground truth labels opens tremendous opportunities to simulate urban land use changes in an unprecedented, detailed manner. This remains uncharted territory.

3. Methodology

This study develops a novel workflow to predict fine-grained urban land use changes by integrating vision-language analysis and SDM. Fig. 1 demonstrates the main technical pipeline. First, we conducted data collection and preprocessing by gathering SVIs and relevant spatial data, followed by cleaning and preparing these datasets for analysis. Second, we performed vision-language analysis using UrbanCLIP to infer urban land use from the images, enabling us to classify urban areas without relying on extensive labeled datasets. Third, we conducted SDM to integrate the inferred land use to simulate future urban changes at a highly localized level. Finally, we validated and generated the final outputs to assess the model's performance and produce detailed predictions of urban transformations. This comprehensive methodology enhances predictive accuracy and spatial granularity, providing a robust framework for simulating future urban developments.

3.1. Study area and data

We selected Shenzhen, located in Guangdong Province, China, as the study area due to its rapid urbanization and diverse mix of urban functions. Since its designation as a Special Economic Zone in 1980, Shenzhen has transformed from a small fishing village into a bustling metropolis with a population exceeding 12 million. The city's dynamic urban landscape and rapid development make it an ideal study area to test the efficacy of fine-grained urban predictive modeling. Predicting Shenzhen's future urban fabrics is particularly compelling due to its role as a global innovation hub and a model for urbanization in developing economies. The city's development trajectory directly impacts regional economic growth, environmental sustainability, and urban policy strategies, and thus accurate predictions are crucial to inform planning and sustainable development.

To capture the detailed urban land use of Shenzhen, we collected SVIs across the entire city at a high spatial resolution. The data collection process involved several steps: (1) partitioning the entire territory of Shenzhen into hexagons of 10,000 m^2 each; (2) retrieving the centroids of the hexagons; (3) filtering out the centroids whose 100 m buffer zones do not intersect with the urban planning zones of the built up environment; (4) aligning each centroid to its closest point in the road network; (5) using the aligned closest points on the roads as sampling points; (6) requesting the SVIs at each sampling point with heading angles of 0° (due north), 90° (due east), 180° (due south), and 270° (due west). Overall, in 2022, we captured four SVIs at each sampling point with an interval of roughly 100 m, resulting in 226,881 SVIs from Baidu Map API in Shenzhen. With four images collected at each sampling point and a conservative estimate of 25 m as the average visual depth of SVIs in open street spaces (Miao et al., 2025), this results in an effective sampling interval of approximately 50 m, which allows for a fine-grained capture of the urban fabric.

Beyond SVIs, a range of complementary datasets was utilized (see Table A.1 in Supplementary Materials). These datasets include population distribution data for density analysis, detailed zoning regulations, future development policy maps, administrative boundaries, and road networks. Specifically, road networks sourced from OpenStreetMap provided essential spatial connectivity information, while population data from the Shenzhen Bureau of Planning and Natural Resources enabled detailed demographic analyses. Additionally, policy maps, zoning regulations, and administrative boundaries provided critical contextual constraints and policy-driven spatial guidelines necessary for accurate scenario modeling.

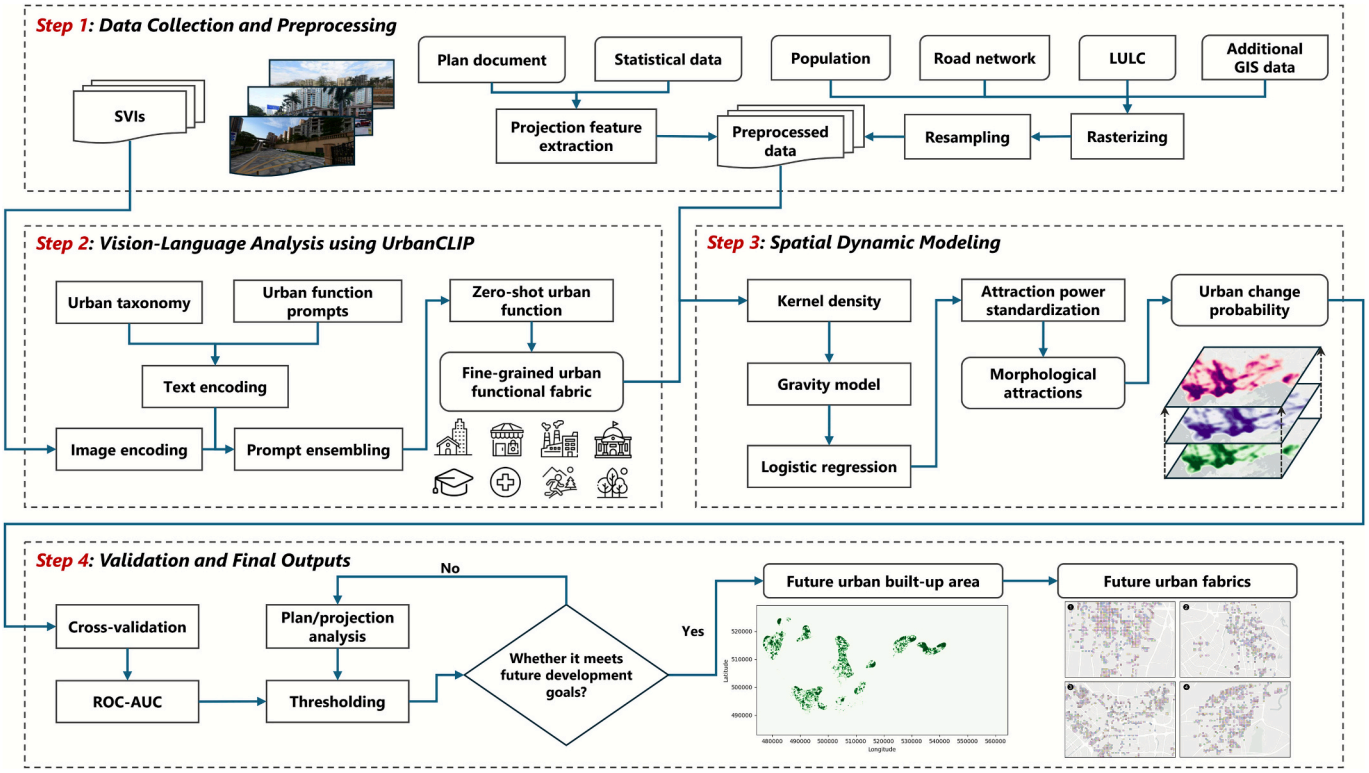


Fig. 1. Technical workflow of the proposed method: (1) data collection and preprocessing from SVIs, planning documents, and GIS layers (see Annex for details); (2) fine-grained urban function classification using UrbanCLIP; (3) SDM combining kernel density, a gravity model, and logistic regression; (4) validation and output generation through cross-validation and projection alignment.

Integrating UrbanCLIP-derived land use with authoritative planning data combines the strengths of both sources. UrbanCLIP captures detailed street-level urban functions overlooked by traditional methods, while planning authority data ensure comprehensive spatial coverage, including areas unobservable by SVIs. This integration facilitates a robust and practical simulation of Shenzhen's urban dynamics, aligned with strategic planning frameworks of the Shenzhen Master Plan for 2035 (see Table A.2 in Supplementary Materials). It ensures the simulation adheres to critical indicators such as ecological preservation, controlled urban expansion, and enhanced infrastructure, effectively supporting sustainable urban policy decisions.

3.2. Zero-shot urban land use inference with UrbanCLIP

Inferring the fine-grained land use status from a large number of SVIs (more than 200,000 images with a roughly 50-m interval along the road network) involves laborious and computationally intensive tasks to annotate many labeled samples (SVIs) and train a deep learning model. There are also major barriers in previous studies to downscale the simulation of urban land use change to such a fine scale (Ren et al., 2019).

To address this, we employed UrbanCLIP (Huang et al., 2024), a prompting framework of the pre-trained VLM-CLIP. UrbanCLIP produces zero-shot urban land use classification using SVIs by inferring the land use type reflected from each individual SVI without the need of model training and labeled samples. In the following sub-sections, we summarize the key components of the UrbanCLIP framework and how zero-shot inference is accomplished.

3.2.1. Urban taxonomy

Each urban land use category entails abstract and sometimes polysemous semantics, and can be mapped to various visual clues if directly prompting CLIP with land use category names, e.g., residential, commercial, and industrial. This poses significant difficulties for CLIP, leading to suboptimal performance. To address this, the UrbanCLIP framework incorporates an urban taxonomy to map each land use category to tens of specific urban objective types (UOTs). For example, *residential* is linked to 49 urban objective types, such as apartment, attached housing, bungalow, central-passage house, and condominium.

In the original paper of UrbanCLIP (Huang et al., 2024), ten urban land use types were defined, which were linked to 354 UOTs in the urban taxonomy. In this paper, we modified the urban taxonomy by merging *commercial* and *hotel* to a single land use type (*commercial*) and removing the *transportation* category, as this category is readily represented by the road network. In this context, we have 8 urban land use types (*residential, commercial, industrial, education, healthcare, civic and governmental, outdoors and natural, and sports and recreation*) and 328 UOTs in the modified urban taxonomy. The categories capture the predominant urban land use types and reflect the diversity of land uses in Shenzhen.

With this modified urban taxonomy, the abstract land use types were transformed into specific urban objects with less ambiguous visual clues. This facilitated the zero-shot inference process of the pre-trained VLM and improved the accuracy of the zero-shot inference. For a comprehensive presentation of urban taxonomy, see Huang et al. (2024).

3.2.2. Urban function prompt template

While the challenge of abstract semantics of urban land use types was addressed by the urban taxonomy, another challenge remained. Due to the "street" nature of SVIs, they commonly contain visual clues of several types of prevalent but distracting objects (e.g., road surface, road-side trees, and vehicles on the streets). Such objects could potentially misguide the zero-shot inference process of CLIP, and this problem is especially notable as such objects often appear in the foreground of SVIs.

In this context, Huang et al. (2024) developed six prompt templates

to mitigate these misguiding objects. The first prompt template is "{UOT}" with no contextual information, enforcing the model to focus on the UOT itself. The second is "A street photo of {UOT} in city." to provide the urban context for the inference. The third template is "A street photo of {UOT} with many trees." to minimize misguidance due to road-side greenery. Several other prompts were also used to train the model to pay less attention to cars, road surfaces, and parking lots in the images.

3.2.3. Zero-shot urban land use inference with UrbanCLIP

During inference, each UOT was inserted into the placeholder in the six prompt templates, generating six sentences. The six sentences were fed into the pre-trained CLIP encoder, to generate six text embeddings. The six embeddings were then ensembled through element-wise averaging. The 328 UOTs in the modified urban taxonomy resulted in 328 ensembled text embeddings.

On the vision side, each SVI is processed through a pre-trained image encoder to generate an image embedding. This embedding is then compared with the 328 UOT embeddings to identify the one with the highest cosine similarity. The corresponding urban land use type of the top-1 similar UOT is then used as the zero-shot inference result of UrbanCLIP. For example, if the UOT *apartment* has the highest cosine similarity with an SVI, then the corresponding urban land use type *residential* in the modified urban taxonomy is deemed as the result for this image (location). After this process, the reflected urban land use status from each SVI can be inferred without the need for labeled samples and model training. With the large number of SVIs in this study area, an urban land use map can then be produced at the scale of roughly every 50 m along the road network.

To assess UrbanCLIP's zero-shot land use inference, Huang et al. (2024) manually annotated a ground truth dataset of 1518 SVIs in Shenzhen. The labels were determined by analyzing visual indicators within the images and cross-referencing them with detailed planning maps and online map services. The results in these annotated SVIs indicate that UrbanCLIP's inference accuracy is nearly 70 % in Shenzhen, in a zero-shot setting. An analysis further showed that the errors of misclassification in Shenzhen are due to several major factors: (1) uncommon building appearances, e.g., commercial complex transformed from a cruise; (2) similar design forms shared by different urban functions, e.g., a factory is misinterpreted as a school due to high visual similarity; (3) being distracted by side information, e.g., a residential area is misclassified as a commercial area due to a prominent food-related advertisement on a wall.

The UrbanCLIP model's capacity is likely higher than the ~70 % accuracy suggested by test dataset, due to factors related to both the test's design and the methodology's application. The validation was conducted on a "high-pressure testbed" that deliberately included a significant portion of challenging cases to rigorously test the model's limits. Even on this difficult dataset, UrbanCLIP performed strongly on the most predominant urban land uses, such as residential (F1: 0.83), commercial (F1: 0.78), and outdoors/natural (F1: 0.72). Furthermore, the methodology is inherently resilient to individual classification errors. The dense sampling of SVIs ensures that a given site, like a university campus, is captured in multiple images. The collective evidence from these images makes the overall deduction reliable, even if a fraction are misclassified. Finally, the subsequent SDM involves a smoothing process (see Section 2 in Supplementary Materials), which leverages the data's density to naturally amplify the consensus from correct classifications while diminishing the impact of sporadic errors.

3.3. SDM with machine learning

Building on the fine-grained urban land use patterns derived from UrbanCLIP, we developed a predictive urban modeling framework to simulate future land use changes in Shenzhen. This framework integrates SDM techniques with machine learning to capture the complex interplay of factors influencing urban growth and land use transformations. The modeling framework operates at a spatial resolution of ~50 m, directly matching the granularity of UrbanCLIP-derived urban land use patterns from street view imagery.

The modeling process involved three interconnected components: kernel density estimation for quantifying spatial distributions, a gravity model for assessing interactions between urban areas, and polynomial regression for estimating the probability of urban transitions. These methods are carefully calibrated and integrated to ensure an accurate and nuanced representation of Shenzhen's rapid urbanization dynamics.

3.3.1. Kernel density estimation (KDE)

KDE was employed as the foundational step to model the spatial distribution of urban functions and population density across Shenzhen. This method provides continuous surfaces representing the intensity of urban functions, enabling the identification of potential growth hotspots and understanding the spatial patterns of urban activities.

The input data for this step included geocoded locations of various urban functions derived from SVIs, such as residential, commercial, and industrial points, alongside population data obtained from the latest census. The KDE implementation utilized the Gaussian kernel, a widely adopted function that ensures smooth transitions in spatial data. The density at any location x is estimated as:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where n is the number of observations, h is the bandwidth parameter controlling smoothness, and x_i represents the locations of the i-th data point (urban functions or population centers), and $K(u)$ is the Gaussian kernel defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{2}$$

The bandwidth parameter h plays a pivotal role in controlling the level of smoothness in the resulting density surface. Smaller bandwidths emphasize local variations, capturing granular patterns of urban intensity, while larger bandwidths generalize trends over broader spatial extents. To determine an optimal bandwidth for KDE, we conducted a systematic evaluation of candidate values (500, 1000, 2000, 5000 m) using both likelihood-based cross-validation scores and visual comparison of the resulting density curves (see Section 2 in Supplementary Materials). The cross-validation metric indicated that mid-range bandwidths (1000–2000 m) achieved the best fit, while visual inspection confirmed that they provided a balance between preserving meaningful spatial patterns and avoiding oversmoothing. Finally, KDE outputs were resampled to a 10-m grid, creating continuous raster layers that formed the foundation for subsequent modeling steps.

3.3.2. Gravity model for attraction power

To quantify the spatial attraction power across the urban areas, we adopted a gravity model framework commonly used in urban interaction and accessibility studies. This model estimates the relative influence of one location on another based on its functional intensity and accessibility. The gravity model conceptualizes attraction as a function of the mass of the destination area and the distance between locations. The attraction between cell j and cell i is expressed as:

$$A_{ij} = \frac{M_j}{D_{ij}^{\alpha}} \tag{3}$$

where M_j represents the mass of the cell j, derived from KDE outputs, while D_{ij} is the distance between i and j, adjusted for road network accessibility. The parameter α controls the sensitivity of attraction to spatial separation. A higher value of α indicates a steeper decline in attraction with increasing distance. To calculate mass values M_j, we

combined the KDE-derived densities of population and urban functions. Each KDE layer was normalized and weighted to reflect its relative contribution to urban interaction potential, assigning stronger influence to cells with greater population or functional intensity.

Distance D_{ij} was adjusted to reflect network-based accessibility rather than using raw Euclidean distance. Travel impedance was derived from Shenzhen's road network, considering road hierarchy and average speed to better approximate real-world travel times between cells. This adjustment ensures that the model accounts for the spatial friction imposed by the transportation system.

The decay parameter α was fixed at 0.5, a commonly used value in spatial interaction modeling that balances strong local influence with persistence of medium-range effects (Cai et al., 2020). This choice avoids overfitting to short-range noise while remaining computationally tractable for large-scale raster calculations. To ensure robustness, we conducted a sensitivity test with alternative values (α = 0.25, 0.75, and 1.0), which produced qualitatively similar attraction gradients and nearly identical predictive accuracies (see Section 3 in Supplementary Materials). This stability suggests that the model outcomes are not overly dependent on the exact choice of α. While the approach assumes that distance-mediated interactions will continue to structure urban growth, we acknowledge that policy, economic, or environmental shifts may reshape these dynamics in practice.

Attraction values were computed for each cell by summing distance-decayed potentials with surrounding cells within a defined spatial radius. The resulting attraction power surface was then standardized using z-score normalization:

$$MA_i = \frac{A_i - \mu_A}{\sigma_A} \tag{4}$$

where A_i is the (unstandardized) attraction power of cell i, and μ_A and σ_A are the mean and standard deviation of attraction values across all cells. The standardized indices were used as inputs for the polynomial regression model, linking spatial interaction dynamics to the probability of urban transitions.

3.3.3. Regularized polynomial regression for urban land use change probability

Polynomial regression was selected over more complex models (e.g., random forests, neural networks) to balance predictive power with interpretability and efficiency. This approach allows us to model non-linear relationships between spatial drivers and land use transitions while maintaining transparency of variable influences essential for urban planning decisions. It ensures computational scalability at high spatial resolution (~50 m), without the overfitting risks and resource demands of "black-box" models.

The dependent variable P_i denotes the estimated likelihood that cell i is in a built-up state. It is calibrated using a binary built-up versus non-built-up indicator derived from authoritative land use data for Shenzhen (2010 and 2020), rather than from observed multi-temporal land use transitions. Cells classified as built-up are treated as positive instances, while non-built-up cells serve as negative instances in model calibration. Thus, P_i represents development likelihood or transition suitability, rather than a directly observed probability of historical land conversion.

Independent variables included morphological attraction derived from the gravity model MA_i, accessibility to road networks A_{cc_i}, socio-economic indicators $Socio_i$, and policy-related factors such as zoning classification and development zones $Policy_i$. Morphological attraction was derived using the gravity model described earlier; accessibility was calculated based on proximity and connectivity to major roads and transit hubs; socioeconomic indicators were determined through aggregated population and density metrics; and policy factors were obtained from zoning classifications and planned development zones defined by Shenzhen's urban planning authority. The model's structure is defined as:

$$P_i = \beta_0 + \beta_1 MA_i + \beta_2 A_{cc_i} + \beta_3 Socio_i + \beta_4 Policy_i + \beta_5 MA_i^2 + \beta_6 MA_{cc_i}^2 + \ldots + \varepsilon_i \tag{5}$$

where β_0 to β_n are coefficients, and ε_i is the error term.

To mitigate potential multicollinearity introduced by polynomial expansion, we applied ridge regression—a regularization technique that penalizes the magnitude of coefficients to reduce overfitting. The ridge regression objective minimizes:

$$\frac{min}{\beta} \sum_i (P_i - \widehat{P}_i)^2 + \alpha \sum_j \beta_j^2 \tag{6}$$

where α determines the degree of regularization. Larger α values shrink coefficients more strongly toward zero, improving model robustness. To determine the optimal α value, we applied 10-fold cross-validation. The dataset was randomly partitioned into ten equal subsets; in each iteration, nine subsets were used to train the model, while the remaining subset was used for validation. This process was repeated ten times, and the average performance across folds was used to identify the α value that minimized prediction error. This approach improves generalizability and guards against overfitting by evaluating the model on unseen data during training.

Model performance was assessed using Residual Standard Error (RSE) and Adjusted R-squared for explanatory power, and the Area Under the ROC Curve (AUC-ROC) for classification accuracy. Following validation, the model projected urban change probabilities for 2035 scenarios, updating independent variables to reflect anticipated demographic, infrastructural, and policy developments, ultimately generating a high-resolution forecast of built-up area expansion.

3.3.4. Future urban projection with land use pattern allocations

The validated polynomial regression model was applied to predict urban change probabilities for each spatial cell under projected conditions for 2035. These probabilities were classified into binary outcomes using a threshold determined by Shenzhen's master plan, identifying cells likely to transition to built-up areas. This threshold calibration, based on official targets outlined in Table A.2, ensured the projected total area of urban expansion closely matched Shenzhen's policy goal of adding 30 km^2 of newly developed area by 2035, thereby aligning projections precisely with strategic planning objectives.

Subsequently, an official land use mask derived from Shenzhen's authoritative planning datasets was applied to constrain potential developments, ensuring simulations respected zoning regulations, ecological preservation zones, and areas designated as non-developable. Once the official mask delineated feasible built-up areas, a systematic method was employed to assign specific urban land use patterns, such as residential, commercial, and industrial, to these zones. Allocation involved probabilistic modeling complemented by spatial placement rules to ensure realistic and cohesive urban development. The probability $P_{i,k}$ of cell i being assigned to a specific land use type k was calculated as:

$$P_{i,k} = \frac{exp\left(\beta_{k0} + \sum_j \beta_{kj} X_{ij}\right)}{\sum_l exp\left(\beta_{l0} + \sum_j \beta_{lj} X_{ij}\right)} \tag{7}$$

where β_{k0} represents the intercept for land use type k, β_{kj} are coefficients for explanatory variables X_{ij} (e.g., proximity to existing land uses, accessibility to major roads, policy compliance), and l indexes all possible land use types. These variables were clearly defined and calculated in previous sections, particularly highlighting unified accessibility metrics and morphological attraction from the gravity model.

Detailed spatial placement rules guided iterative allocation, including: (1) clustering near existing land uses; (2) optimizing

accessibility to major roads and transit nodes; (3) adhering to environmental constraints and ecological preservation; and (4) enforcing zoning regulations and development priorities stipulated in Shenzhen's Master Plan (Table A.2). Specifically, the zoning mask operationalized the *"Three Zones and Three Control Lines"* framework: the ecological control line restricted development in protected corridors, the permanent basic farmland line safeguarded agricultural land, and the urban development boundary capped total built-up land at 1125 km^2 by 2035. In parallel, predictor variables were updated to reflect anticipated 2035 conditions: population growth was aligned with the official forecast of 17.6 million permanent residents, accessibility layers were revised to incorporate the planned 1000 km metro network, and socioeconomic indicators and policy variables were adjusted to match zoning and redevelopment priorities outlined in the Master Plan. These updates ensured consistency between projected dynamics and statutory planning assumptions. Collectively, the constraints and updated predictors embody the plan's principle of *"growth within boundaries"*, redirecting expansion toward compact, transit-oriented, and policy-compliant areas. The iterative allocation prioritized areas with higher transition probabilities first, while simultaneously respecting these boundaries, thereby enhancing spatial continuity, preserving ecological and agricultural functions, and ensuring consistency with Shenzhen's long-term sustainability goals. The final result was a comprehensive and cohesive high-resolution urban land use projection, harmoniously integrated with both existing patterns and policy constraints.

4. Results

4.1. Fine-grained urban land use patterns

The integration of UrbanCLIP with high-resolution street view imagery produced a fine-grained spatial distribution of urban land use patterns across Shenzhen, classified into eight primary categories: residential, commercial, industrial, education, healthcare, civic and governmental, outdoors and natural spaces, and sports and recreation. Fig. 2 illustrates the spatial heterogeneity of these patterns, highlighting the complex interplay of functions within Shenzhen's urban land use. Residential areas dominate suburban districts, interwoven with commercial hubs that emphasize mixed-use developments. Commercial zones are prominently clustered along key transit corridors and urban centers, reflecting Shenzhen's transit-oriented development strategy. Industrial areas, in contrast, are concentrated on the urban periphery, aligning with strategic zoning policies aimed at minimizing conflicts with residential and recreational areas.

The zoomed-in areas (Spots 1–4) provide deeper insights into localized land use patterns. Spot 1 reveals dense residential and commercial clusters in central urban districts, supporting vibrant mixed-use developments. Spot 2 showcases the juxtaposition of healthcare and educational facilities alongside residential neighborhoods, ensuring accessibility to essential services. Spot 3 highlights a predominantly civic and governmental zone interspersed with natural and recreational areas, indicative of a balanced land use approach. Spot 4, situated at the city's outskirts, features a mix of industrial areas and outdoor spaces, demonstrating the city's strategic balance between economic activities and environmental sustainability. These urban land uses, mapped at a ~ 50-m resolution (interval), underscore the efficacy of UrbanCLIP in extracting detailed urban functions while minimizing the need for labeled datasets. The mapped patterns provide the empirical basis for the SDM and the planning analyses that follow.

4.2. SDM results

The SDM framework effectively integrated polynomial regression, KDE, and attraction power analysis to simulate Shenzhen's future urban land use changes. The polynomial regression analysis (see Figs. A.2–A.5 in Supplementary Materials) evaluated the relative contributions of population attraction, accessibility, and other urban factors to land use transitions.

Population attraction emerged as the strongest predictor across all models, underscoring its critical role in driving urban land use changes. In the first-degree polynomial model (Fig. A.2), population attraction exhibited a highly significant positive relationship with land use change ($p < 0.001$; Estimate = 0.499), but the fit was limited by the model's linear assumptions (Adjusted $R^2 = 0.7348$). By incorporating higher-order terms, the second-degree model (Fig. A.3) captured complex nonlinear dynamics, with the quadratic term of population attraction ($p < 0.001$; Estimate = 1.95) enhancing the model's ability to reflect spatial clustering and density thresholds. This refinement significantly improved model fit (Adjusted $R^2 = 0.9504$) and highlighted the intricate, non-linear interactions underlying urban growth.

Accessibility metrics including proximity to education, healthcare, and commercial centers generally showed weaker statistical significance across all models when analyzed individually ($p > 0.3$ in first-degree models). This can be attributed to Shenzhen's relatively uniform spatial distribution of these facilities, which reduces their differential impact on land use transitions. However, their role became important in enabling functional clustering and neighborhood stability, as reflected in the slight positive effect of first-degree education accessibility ($p < 0.001$; Estimate = 0.661) and the second-degree term ($p < 0.001$; Estimate = −1.156). This suggests a localized influence, where accessibility indirectly supports urban transformation through interactions with other factors.

At higher polynomial degrees (Fig. A.4 and Fig. A.5), accessibility metrics such as proximity to healthcare ($p < 0.001$; Estimate = 0.356 for the quadratic term) and commercial centers ($p < 0.001$; Estimate = −0.464) contributed to the model but with diminishing significance. This indicates that their direct impact on land use transitions becomes more diffuse as spatial dynamics are captured through more complex interactions.

The refined polynomial regression model (Fig. A.5) demonstrated the best balance between complexity and predictive power, achieving an Adjusted R^2 of 0.9882. This model incorporated selective polynomial degrees for key predictors to account for their non-linear effects without overfitting. The integration of KDE and attraction power analysis further supported these findings by highlighting areas with high spatial clustering, consistent with the patterns predicted by the polynomial regression.

Fig. 3 complements these findings by visualizing the spatial dynamics of urban attraction power and change probabilities across Shenzhen. Fig. 3a displays the modeled change probabilities for the year 2035, highlighting regions of high urbanization likelihood. Central Shenzhen, particularly the Futian and Luohu districts, shows the highest probabilities, driven by dense population clusters, well-established commercial zones, and robust transit networks. These areas are also marked by the concentration of mixed-use developments, underscoring their role as urban growth hotspots. In contrast, peripheral regions such as the eastern and northern outskirts exhibit lower change probabilities, reflecting weaker functional intensity and limited infrastructure connectivity.

Fig. 3b–j illustrate the spatial distribution of attraction power for various urban functions, offering insights into the dynamics driving Shenzhen's urbanization. Residential areas (Fig. 3b) demonstrate strong clustering effects in suburban zones with high public transit accessibility, serving as growth anchors due to their stable population bases. Commercial areas (Fig. 3c) are concentrated along key transit corridors such as Shennan Boulevard and Shenzhen Bay, where accessibility and commercial activity synergize to reinforce economic hubs. Industrial zones (Fig. 3d) exhibit high attraction power on the urban periphery, reflecting zoning policies designed to balance economic development with central urban livability. Meanwhile, education and healthcare facilities (Fig. 3e and f) are evenly distributed across Shenzhen, ensuring

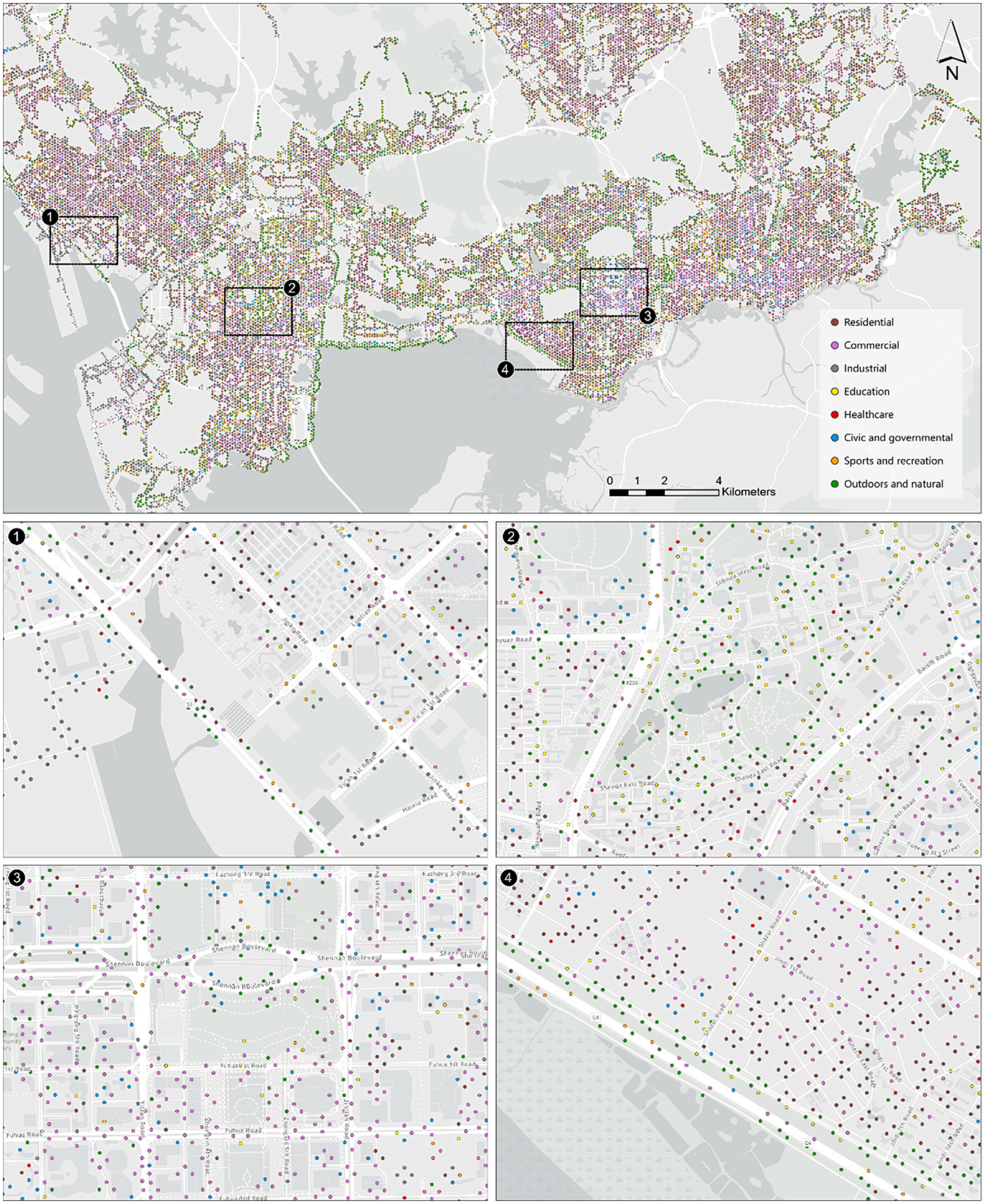


Fig. 2. Fine-grained spatial distribution of urban land use functions across Shenzhen. Spots 1–4 highlight local functional heterogeneity.

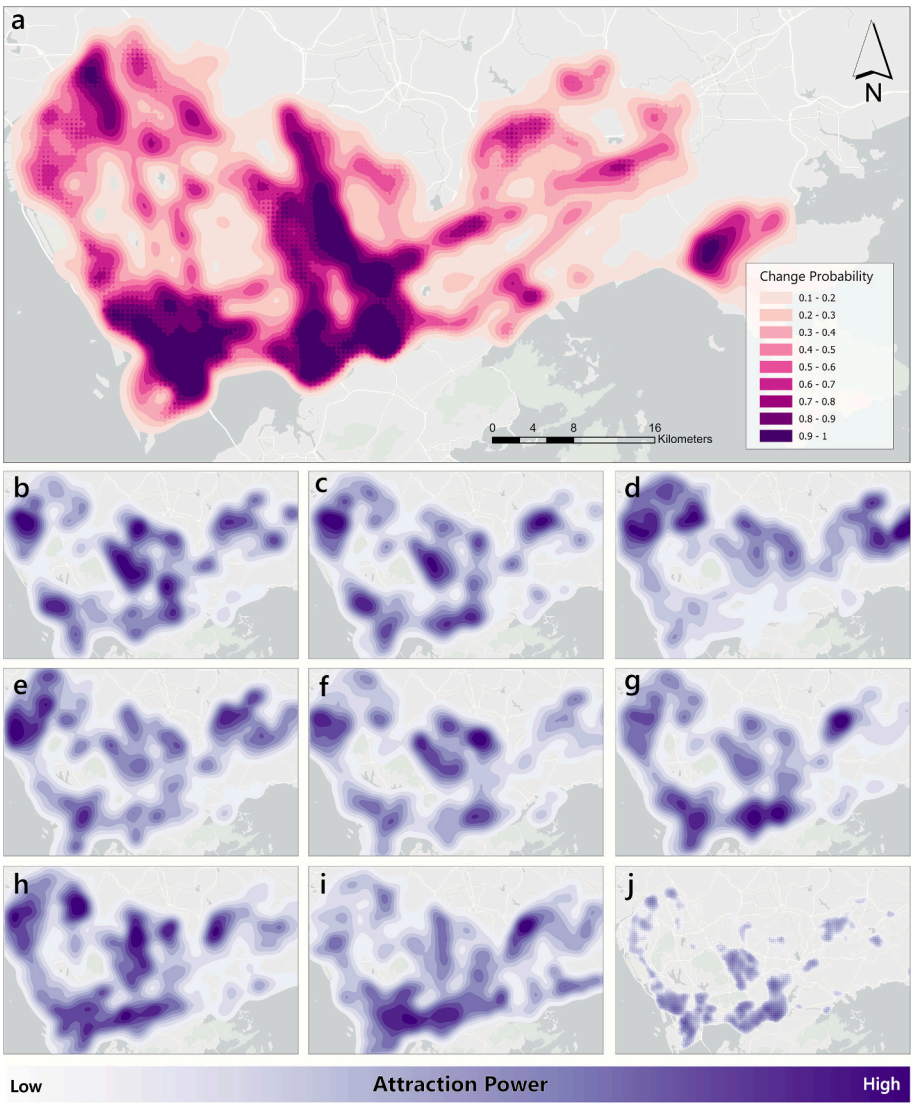


Fig. 3. Urban land use change probability and attraction power in Shenzhen. (a) highlighting future urban change probabilities for 2035; (b) to (j) detailing attraction power for residential (b), commercial (c), industrial (d), education (e), healthcare (f), civic and governmental (g), sports and recreation (h), outdoor and natural (i), and integrated population attraction (j).

equitable service access and contributing to overall functional diversity. Civic and governmental functions (Fig. 3g) are strategically located in administrative centers and sub-centers, enhancing the accessibility of essential public services. Sports and recreational facilities (Fig. 3h) are predominantly co-located with residential neighborhoods, fostering community engagement and quality of life. Finally, outdoor and natural areas (Fig. 3i) exhibit high attraction power in ecological preservation zones, aligning with Shenzhen's sustainability priorities and enhancing environmental resilience. Together, these land use patterns highlight the interconnectedness of urban growth drivers, emphasizing the importance of accessibility, clustering, and land use diversity in shaping urban dynamics. The combined attraction power (Fig. 3j) integrates these patterns into a cohesive representation, highlighting population attraction as a key driver. Central Shenzhen emerges as a dominant hub with concentrated high-value zones, while attraction intensity diminishes outward. This pattern mirrors the city's development trajectory, underscoring the critical role of urban cores in driving regional growth.

These findings validate the SDM framework, demonstrating its ability to capture the multifaceted interactions shaping Shenzhen's urban growth.

4.3. *Prediction of future urban land use*

Ridge regression, with its capacity to regularize and balance feature weights, proved instrumental in identifying critical drivers of urban transitions. Fig. 4 showcases the ridge regression model's performance, with the optimal α parameter determined through cross-validation. The cross-validation curve indicates that an α value of 1000 minimizes the mean squared error (MSE), reflecting an appropriate trade-off between model complexity and prediction accuracy.

Fig. 4a highlights the relative importance of features influencing urban change probabilities. Population density emerged as the dominant factor, consistent with Shenzhen's urban dynamics, where densely populated areas are primary drivers of land use transitions. Transportation accessibility, including proximity to transit hubs and road networks, ranked as the second most influential factor, reflecting the city's emphasis on transit-oriented development. Residential clustering further supports urban transitions, acting as anchors for neighborhood development. Education and healthcare, while essential for urban livability, had comparatively lower coefficients, suggesting an indirect role in shaping immediate land use changes.

The feature correlation matrix in Fig. 4e offers deeper insights into

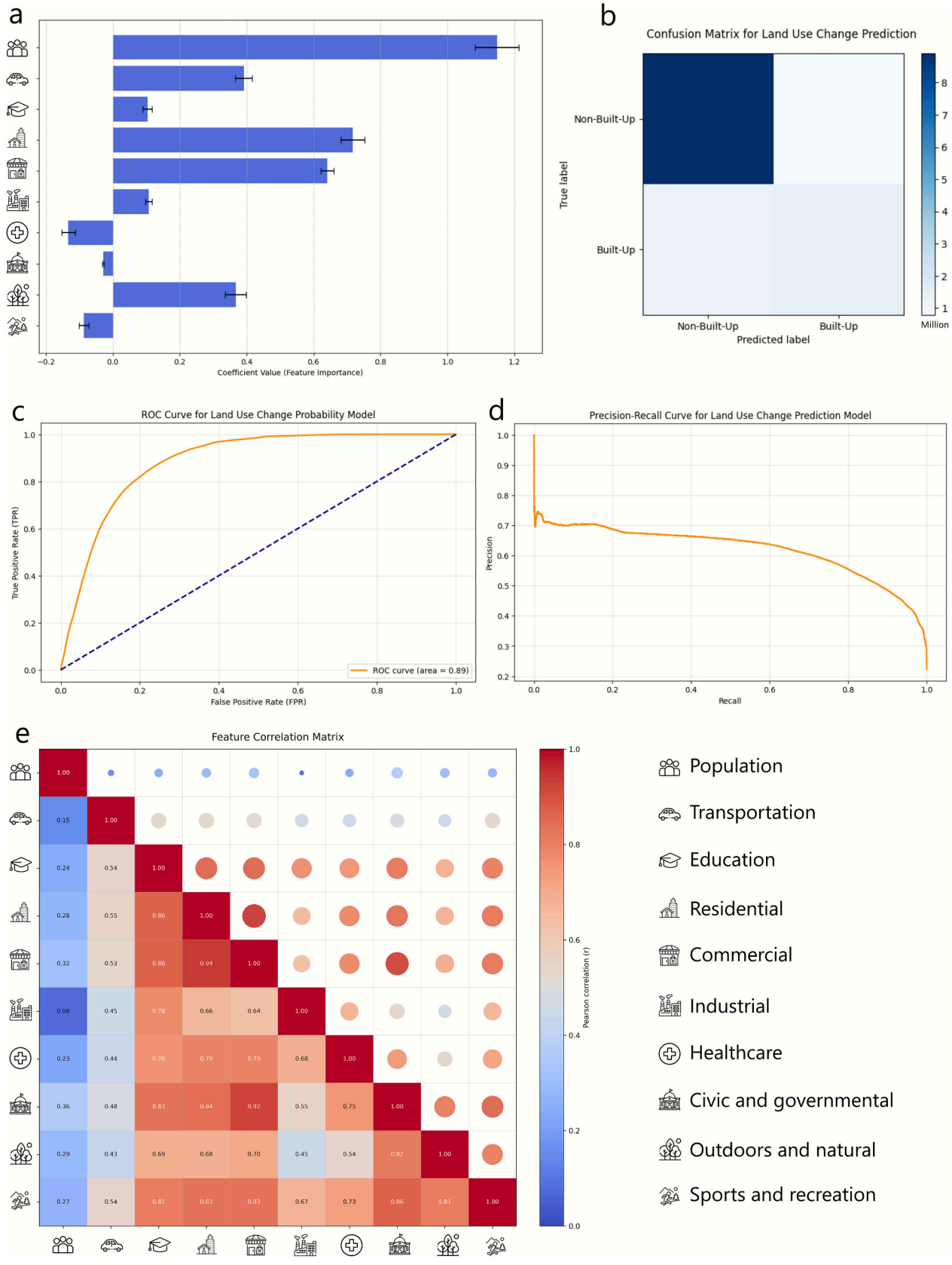


Fig. 4. Ridge regression model and feature importance for land use change prediction. (a) Coefficient values for feature importance; (b) Confusion matrix for land use change prediction; (c) ROC-AUC curve; (d) Precision-recall curve; (e) Feature correlation curve.

the relationships among urban variables. Strong positive correlations were observed between population and residential features (correlation coefficient = 0.94), emphasizing their intertwined influence on urbanization patterns. Similarly, transportation accessibility exhibited strong correlations with commercial features (0.85), highlighting the role of transit corridors in driving commercial development. Industrial features showed moderate correlations with transportation (0.75), consistent with zoning policies that locate industries along accessible peripheries. The matrix also revealed weaker relationships between outdoor and natural spaces and other features, reflecting their role as isolated yet critical components of Shenzhen's sustainability agenda.

Interestingly, correlations among civic and governmental, healthcare, and educational facilities were moderately high (0.79–0.83), underscoring their clustering tendencies in well-planned mixed-use areas. This clustering not only improves accessibility to essential services but also promotes balanced land use, fostering cohesive urban development. These interdependencies among features highlight the predictive framework's capability to effectively capture the intricate dynamics of Shenzhen's urban interactions.

Fig. 4b–d further validate the model's predictive capabilities. The confusion matrix (Fig. 4b) reveals strong performance in distinguishing between built-up and non-built-up areas, with over 89 % of non-built-up areas and 75 % of built-up areas correctly classified. While some misclassification occurred, this was limited to areas with overlapping functional influences, such as mixed-use zones near the urban periphery. The ROC curve (Fig. 4c) achieved an AUC of 0.890, underscoring the model's high sensitivity and specificity, while the precision-recall curve (Fig. 4d) demonstrates its reliability across varying thresholds, particularly for identifying high-probability urbanization zones.

The cross-validation results in Fig. 5 illustrate the model's performance in determining the optimal regularization parameter α for ridge regression. The MSE is plotted against the logarithmic scale of α, providing insights into how varying levels of regularization impact model accuracy. The red line represents the mean MSE, while the shaded gray area denotes the confidence interval (± 1 standard deviation), indicating the variability across folds during cross-validation.

The curve exhibits a characteristic pattern where the MSE remains relatively stable at lower regularization values ($\log(\alpha) < 2$), indicating the model's capacity to handle the complexity of features without overfitting. However, as α increases beyond a certain threshold ($\log(\alpha) \approx 3$), the MSE begins to rise, indicating over-regularization, where important feature contributions are excessively penalized, reducing the model's predictive power.

The optimal α, marked by the blue dot ($\log(\alpha) = 3$), minimizes the MSE while maintaining a balance between bias and variance. This value ensures the model achieves robust predictions without overfitting or underfitting, as evidenced by the narrow confidence interval at this point. The stability of the MSE across a wide range of α values highlights the model's resilience to variations in regularization, ensuring reliable application for predicting future urban land use changes. This robust regularization framework allows for the effective integration of complex features, such as population attraction and accessibility, into the predictive modeling process.

Fig. 6 illustrates the projected spatial distribution of future urban fabrics in Shenzhen by 2035, highlighting the functional allocations of residential, commercial, industrial, and other urban uses. The map provides an overview of Shenzhen's anticipated urban growth, with new development concentrated along key corridors and urban cores. The citywide map, combined with detailed Spots 1–4, reveals the diverse and strategic spatial organization underpinning the city's expansion plan.

Across Shenzhen, central and southern districts emerge as major growth hubs, driven by their connectivity and established infrastructure. The Futian and Luohu districts (Spots 2 and 3), already functioning as central business areas, are projected to see further commercial development, supported by nearby residential and civic uses. The Nanshan district in the south demonstrates a more balanced distribution of residential, commercial, and recreational functions, reflecting its role as a mixed-use zone. Northern regions, particularly in the Bao'an district, exhibit suburban expansion with a focus on residential clustering and integrated natural spaces, signaling the city's push to decentralize growth while enhancing livability.

Spot 1, located in western Shenzhen, highlights a dense agglomeration of residential developments complemented by recreational and green spaces. This mixed-use pattern suggests efforts to build compact neighborhoods that reduce commuting times and enhance accessibility. In contrast, Spot 2 in the city center reveals an emphasis on administrative and commercial activities, underpinned by the clustering of civic and governmental and healthcare facilities. This area is expected to

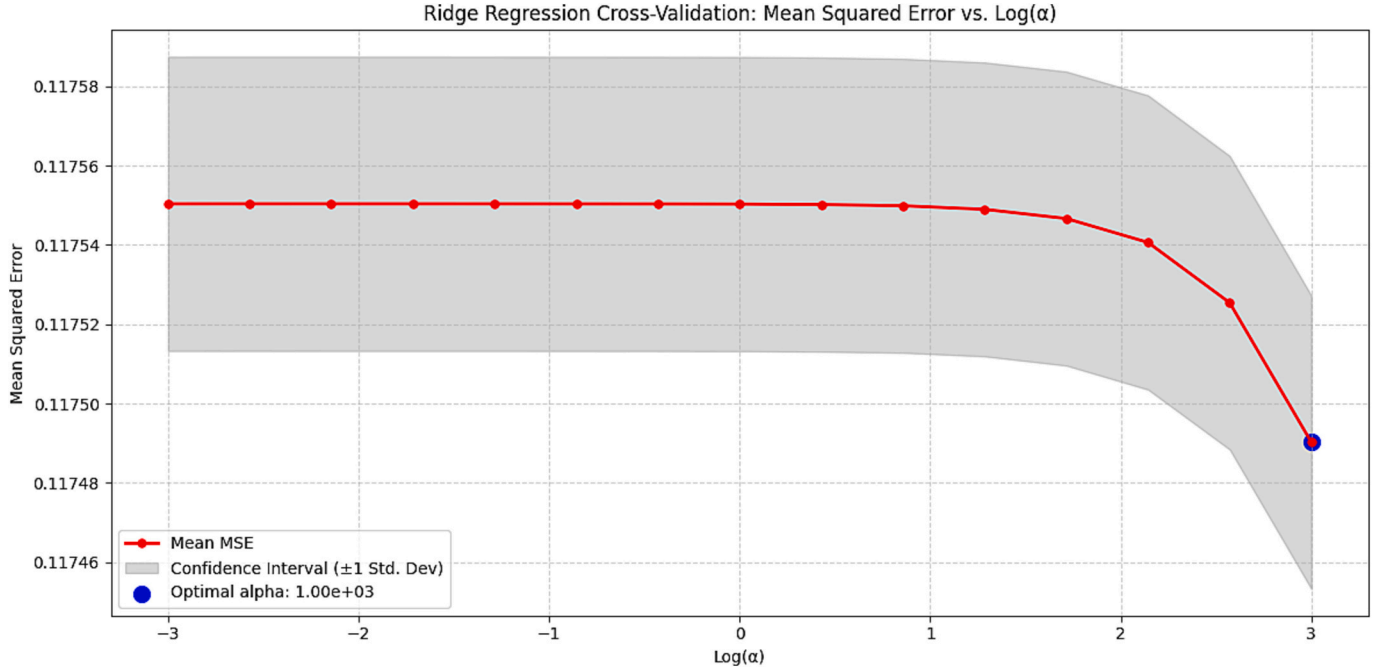


Fig. 5. Ridge regression cross-validation.

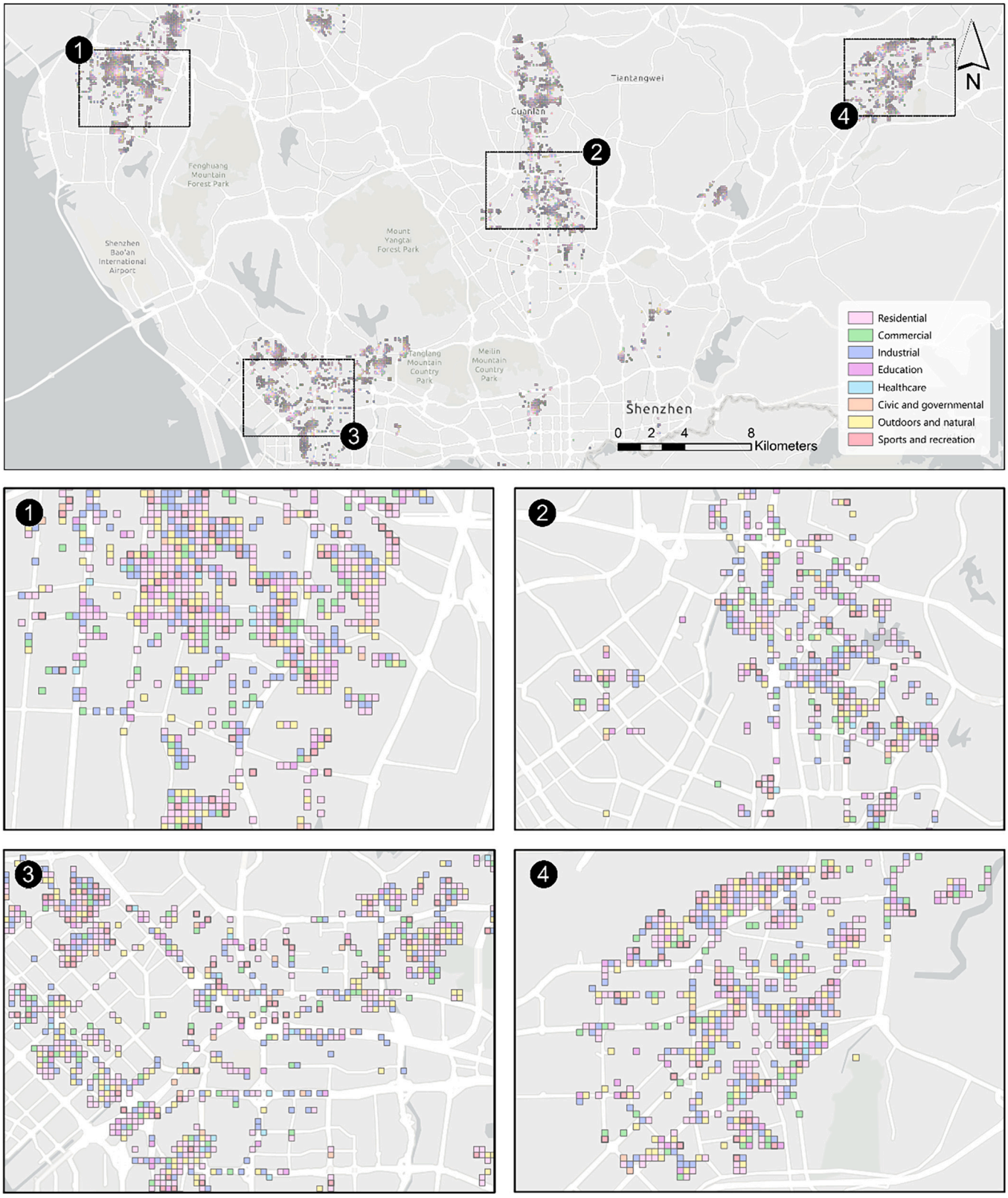


Fig. 6. Projected urban fabrics in Shenzhen for 2035. Spots 1–4 highlight key urban functional patterns.

solidify its status as a service-oriented hub.

Spot 3 in southern Shenzhen focuses on large-scale residential expansion, accompanied by healthcare and recreational functions. This combination aims to create vibrant communities while maintaining urban quality of life. Spot 4, located in the northeastern peripheral regions of Shenzhen, stands out for its suburban growth characterized by a mix of low-density residential areas and significant outdoor and natural spaces. This pattern demonstrates the city's efforts to balance expansion with ecological preservation, ensuring sustainable development in less urbanized areas.

Overall, the projections underscore Shenzhen's strategic approach to urban growth, where mixed-use developments dominate central areas, and suburban regions maintain a balance of residential and green spaces. The deliberate clustering of urban functions such as healthcare, education, and recreation within high-growth zones ensures accessibility, while the preservation of green and outdoor spaces reflects Shenzhen's sustainability commitments. These insights offer a roadmap for policymakers to prioritize resource allocation and infrastructure investments, aligning urban growth with ecological and social objectives. By 2035, Shenzhen's urban landscape is poised to become more resilient, livable, and strategically organized.

5. Discussions

5.1. Model performance and spatial outcomes

The integration of VLMs with SDM substantially enhances both the precision and granularity of urban land use predictions. Using UrbanCLIP, SVI-based functional inference eliminated the need for manually labeled datasets. This capability is particularly valuable in rapidly evolving urban contexts like Shenzhen, where continuously updating labeled data is both costly and time-consuming.

At a ~ 50-m spatial resolution, the resulting maps reveal fine-grained patterns of functional clustering. Residential and commercial zones are shown to interweave along transit corridors, industrial areas are strategically concentrated at peripheral nodes, and recreational and green spaces are well-distributed within residential neighborhoods. These spatial signatures align with Shenzhen's planned but dynamic development trajectory, emphasizing functional diversity, mixed-use accessibility, and transit-oriented intensification.

By embedding these enriched, high-resolution functional layers into the machine learning–enhanced SDM, we captured complex, non-linear interactions among population density, accessibility, morphological attraction, and policy constraints. The validated model achieved a ROC–AUC of 0.890, a PR–AUC of 0.574, and a Brier score of 0.112, indicating robust discrimination, calibration, and precision in identifying high-probability transition zones.

To quantify the value of incorporating SVI-derived fine-grained function signals, we present an ablation-style comparison in (see Table A.5 in Supplementary Materials). The baseline specification represents a conventional potential-based formulation that uses population kernel density and road accessibility (KDE-only), while the full specification additionally incorporates UrbanCLIP-derived functional clustering surfaces. This ablation is designed to isolate the marginal predictive gain attributable to street-level semantic information within the same modeling framework, rather than to serve as a cross-paradigm benchmark against fundamentally different land use mapping systems. Comparative results show that the full model improves discrimination and precision (ROC–AUC: 0.890 vs. 0.835; PR–AUC: 0.574 vs. 0.479) and yields better calibration (Brier score: 0.112 vs. 0.134).

Satellite imagery- and points of interest (POI)-based land use inference provide useful reference points for interpreting these gains. Remote sensing-based mapping offers consistent large-scale (global) coverage and reliable delineation of built-up morphology. However, fine-grained functions (both spatially and semantically) and mixed-use configurations remain challenging to infer from remote sensing alone, as functional semantics are only weakly encoded in spectral-textural signals and are sensitive to local labeling conventions (Bai et al., 2023). POI-based mapping captures place semantics more directly and can perform well when enriched with spatial-context modeling (Huang et al., 2022, 2023; Yao et al., 2023), but it is also heavily influenced by the data completeness of POIs (e.g., commercial areas are well represented whereas industrial facilities are not), taxonomy inconsistencies, and spatial aggregation choices (Hu & Han, 2019). In this context, our SVI-based, zero-shot inference is complementary rather than substitutive, i.e., it introduces fine-grained street-level semantic evidence with high spatial continuity, which helps resolve functional distinctions within built-up fabrics. In addition, the coupling with SDM translates these signals into development likelihood surfaces and policy-constrained scenario projections. The findings and interpretations can inform recent multimodal efforts that combine multiple data modalities to enhance urban land use function mapping (Balsebre et al., 2024; Chen et al., 2024; Zhou et al., 2024), e.g., in terms of how to align the intrinsic information content of each modality with its suitable mapping granularities.

Spatial comparisons (see Section 6 in Supplementary Materials) further illustrate how the additional SVI-derived semantic signals sharpen localized probability patterns beyond the KDE-only baseline. While the KDE-only baseline captured broad growth trends, it oversimplified local variations and failed to detect emerging hotspots. In contrast, the full model provided sharper, localized probability clusters in key redevelopment corridors, central business districts, and transit-oriented nodes, while simultaneously reducing noise in peripheral low-relevance areas. These targeted improvements demonstrate that the integration of functional and visual cues not only improves predictive accuracy but also enhances spatial fidelity and interpretability, yielding actionable insights for planners at the block and corridor levels.

5.2. Planning applications at multiple spatial scales

The fine-grained functional maps, growth-likelihood surfaces, and scenario-based projections provide an empirical basis for planning applications at both metropolitan and neighborhood scales. Importantly, the outputs support spatial diagnosis and scenario design by identifying where growth pressures concentrate and how functional patterns co-locate. The framework does not directly estimate downstream socio-economic outcomes such as commuting time reduction, “urban vitality”, or public health impacts, which would require additional behavioral and network-based evaluation. The implications below are therefore framed in terms of spatial alignment, regulatory feasibility, and policy leverage.

At the citywide scale, the functional maps show that residential and commercial functions are strongly co-located along major transit corridors and within central districts, especially in Futian and Nanshan. When interpreted together with the modeled growth-likelihood surface and the projected 2035 urban fabrics, this corridor-oriented structure provides a clear spatial basis for identifying priority areas where intensification pressure is most likely to emerge. This evidence supports the practical use of transit-oriented redevelopment and corridor-based zoning checks as targeted responses to spatially concentrated growth pressure.

At the neighborhood scale, the mapped functional mosaics highlight localized configurations where residential areas are interwoven with education, healthcare, and recreational functions, as well as areas where such functions appear more weakly represented. Combined with the growth-likelihood surface, these patterns can be used as a screening layer to identify neighborhoods where service demand may increase under future intensification and where facility planning or accessibility improvements may be needed. In parallel, the functional patterns also indicate that industrial functions tend to concentrate at peripheral nodes, suggesting that planning can treat peripheral industrial clusters and their buffers as distinct management zones for land use compatibility and logistics support.

Ecological and regulatory considerations are most clearly illustrated by the Master Plan scenario comparison. In the unconstrained simulation, expansion spills into ecological corridors and farmland and exceeds the official development cap, whereas the plan-constrained scenario redirects growth inward and reinforces compact intensification patterns in central and corridor locations. This contrast demonstrates that planning instruments are not merely boundary conditions but active levers that reshape spatial trajectories. The spatial difference between constrained and unconstrained simulations can therefore be used to identify

where policy protection has the greatest effect, and where conflict pressure would be highest in the absence of constraints.

The zoning framework in the Shenzhen Master Plan 2035 plays a decisive role in shaping these outcomes. In unconstrained simulations, expansion spilled into ecological corridors and farmland, exceeding the official development cap of 1125 km^2. By contrast, the plan-constrained scenario redirected growth inward, producing compact intensification in central and southern districts and reinforcing transit-oriented redevelopment corridors. These results illustrate how planning instruments are not merely boundary conditions but active levers that structure urban futures across scales.

Finally, while coarser models may support strategic goals, our ~50-m outputs reveal block-level clusters, corridor-scale gradients, and neighborhood transitions that are not visible in coarse-resolution products. This level of detail supports practical tasks such as zoning consistency checks, candidate redevelopment-node screening, and infrastructure sequencing in areas where the model indicates consistently higher growth likelihood.

5.3. Policy recommendations and implications

This subsection links the model outputs to planning implications in a transparent and evidence-based way. Rather than presenting broad policy claims, we anchor each implication in the spatial evidence produced by the framework, including the fine-grained functional maps, the growth-likelihood surface, the projected 2035 urban fabrics, and the constrained versus unconstrained scenario comparison under the Shenzhen Master Plan 2035. Because the framework focuses on spatial patterns of functional co-location, development likelihood, and policy-induced redirection of growth, it does not directly estimate downstream outcomes such as commuting time reduction, "urban vitality," or public health impacts. Where such outcomes are discussed, they should be interpreted as potential co-benefits supported by the broader literature, not as effects demonstrated by this study. Table 1 summarizes the resulting evidence-to-policy mapping. It provides a compact set of planning directions that are directly supported by the study outputs, together with a practical implementation focus and an explicit scope note clarifying what the model does and does not claim.

A first implication concerns zoning and land use controls in high-growth corridors and nodes. The results show strong co-location of residential, commercial, and recreational functions in central and corridor settings, and these areas also emerge as high-likelihood zones for future intensification. In practical terms, this suggests that zoning regulations should be reviewed to ensure that compatible mixed-use redevelopment is not unintentionally constrained in locations where the model indicates both strong functional coupling and high development pressure. Rather than asserting specific benefits that are not modeled here, the evidence supports a narrower and more defensible recommendation: align regulatory controls with the observed and projected functional coupling so that redevelopment can proceed without avoidable regulatory friction.

A second implication concerns infrastructure sequencing and public-service provisioning. The growth-likelihood surface and projected 2035 fabrics indicate that expansion pressure is spatially concentrated rather than uniform, with hotspot clusters forming along redevelopment corridors and connected nodes. This provides a basis for using the probability outputs as a screening layer to prioritize the timing and placement of transit capacity upgrades and essential public facilities in neighborhoods expected to intensify, improving coordination between likely growth pressure and investment rollout. The practical advantage is not a guaranteed socio-economic outcome, but a more spatially targeted allocation logic grounded in modeled hotspots.

Third, the scenario comparison provides the clearest evidence for environmental and regulatory implications. In the unconstrained simulation, expansion spills into ecological corridors and farmland and exceeds the statutory development cap, whereas the plan-constrained scenario redirects growth inward and reinforces compact intensification patterns. This contrast identifies where policy protections have the greatest effect and where development–conservation conflict pressure would be highest in the absence of controls. From a planning perspective, the spatial difference between scenarios can be used to prioritize corridor continuity measures, buffer strengthening, and enforcement monitoring in the most sensitive locations, using the model as a spatial indicator of conflict pressure rather than as a quantified ecological-impact assessment.

Finally, the results suggest that the planning instruments and the modeling workflow can be used together as a scenario-testing tool. Because modifying constraint layers and comparing outcomes produces clear spatial differences in simulated trajectories, planners can evaluate alternative boundary placements, corridor protection designs, and redevelopment priorities through iterative scenario design. The 50 m outputs further support implementation by revealing block-level heterogeneity and corridor-scale gradients that are not visible at coarser resolution, enabling zoning consistency checks, candidate redevelopment-node screening, and map-based communication of spatial options. These uses are directly aligned with what the model outputs quantify: functional co-location patterns, growth likelihood, and policy-induced redirection of development.

Table 1
Evidence-lined planning implications derived from model outputs.

Recommendation	Evidence	Implementation focus	Scope note
Zoning review for mixed-use compatibility in hotspot corridors and nodes	Functional co-location surfaces; growth-likelihood hotspots; projected 2035 intensification (Fig. 2; Fig. 3a; Fig. 6)	Review zoning in hotspots to ensure compatible mixed use is permitted; refine sub-district controls where misaligned	Downstream outcomes such as commuting-time change or "vitality" are not estimated here
Infrastructure and public-facility sequencing guided by hotspots	Concentrated growth-likelihood hotspots; projected corridor/node intensification (Fig. 3a; Fig. 6)	Prioritize transit capacity, station-area upgrades, and key facilities in hotspots; align phasing with projected pressure	Infrastructure and service outcomes require additional network and demand evaluation
Strengthen ecological and farmland protection in conflict-pressure zones	Unconstrained vs plan-constrained difference highlights potential encroachment zones (Fig. A.6)	Target enforcement, corridor continuity, buffers, and monitoring where scenario differences are largest	Identifies conflict pressure, not ecological impact magnitudes
Use constraints as scenario levers for plan evaluation	Statutory controls redirect simulated growth trajectories (Fig. A.6)	Adjust constraint layers to test alternatives and compare spatial consequences	Planning support tool; not an optimization or causal policy-effect estimate
Operationalize ~50-m outputs for implementation and communication	Block-level heterogeneity and corridor gradients visible at ~50 m (Fig. 2; Fig. 6)	Use maps for zoning consistency checks, redevelopment-node screening, and evidence-based communication	Participation outcomes are not evaluated empirically here

Note: Recommendations are derived from modeled spatial patterns of functional co-location, growth likelihood, and constraint-induced redirection of development; downstream socio-economic outcomes are outside the inferential scope of this study.

5.4. Transferability and computational considerations

The generalizability of this approach beyond Shenzhen largely hinges on the transferability of UrbanCLIP, which was intentionally designed to be adaptable across diverse geographic and socio-economic contexts. This adaptability is partially achieved through a comprehensive urban taxonomy that includes commonly seen urban objects in many cities. The model's generalizability was validated in Huang et al. (2024), where UrbanCLIP's performance in London and Singapore exceeded its performance in the primary test area of Shenzhen. However, we acknowledge it can be further enhanced through local adaptations. For instance, one can adjust the overall land use categories, modify the list of UOTs to reflect local characteristics, or alter prompt templates to suit different environmental contexts (e.g., a prompt de-emphasizing trees is less useful in an arid region).

In terms of computational efficiency, the proposed method is highly efficient. The text encoding process requires minimal time, while image encoding proceeds at a speed of ~42 images/s with an Nvidia RTX 4090 Laptop GPU. The subsequent SDM is also efficient. Kernel density estimation, applied as Gaussian filtering on a rasterized population surface with the 50 m grid resampled to 10 m for stable neighborhood operations, was completed in under 10 min together with gravity calculations. Polynomial feature expansion combined with ridge regression, tuned over 15 alpha values from 10^{-3} to 10^{3} with cross-validation and parallelized across CPU cores, was finished in about 25 min on the Shenzhen feature matrix. These tasks were performed on an Intel Core i9-12900K CPU with 32 GB RAM. Since the pipeline scales linearly with the number of images, the workflow is well suited for large metropolitan applications once the initial datasets are prepared.

6. Conclusions

This study introduced an innovative methodology that combines VLMs with machine learning-enhanced SDM to predict fine-grained urban land use in Shenzhen. By leveraging UrbanCLIP's zero-shot learning capabilities, we successfully classified SVIs into detailed urban function categories without relying on extensive labeled datasets. This approach enabled the creation of high-resolution urban functional maps, revealing intricate spatial relationships and functional interactions within the city.

Integrating these fine-grained urban land use into our predictive modeling framework significantly enhanced the accuracy and depth of urban change predictions. The model effectively accounted for factors such as morphological attraction, accessibility, neighborhood composition, and policy constraints, providing valuable insights for urban planners and policymakers. These findings support more informed decision-making, facilitating strategic planning and sustainable urban development in rapidly evolving cities and regions.

While our methodology presents meaningful advancements in predictive urban modeling, limitations exist, including data quality constraints, potential biases in SVI-based classification, and reliance on static policy assumptions that may affect outcomes. Future research could address these uncertainties by integrating multi-modal geospatial data and exploring more robust machine learning techniques to improve generalizability and predictive performance.

CRediT authorship contribution statement

Zipan Cai: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrew Karvonen:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation. **Cong Cong:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Weiming Huang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2026.102416.

Data availability

Data will be made available on request.

References

An, L., Grimm, V., Sullivan, A., Turner, B. L., II, Malleson, N., Heppenstall, A., … Tang, W. (2021). Challenges, tasks, and opportunities in modeling agent-based complex systems. *Ecological Modelling, 457*, Article 109685. https://doi.org/10.1016/j.ecolmodel.2021.109685

Araldi, A., & Fusco, G. (2019). From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis. *Environment and Planning B: Urban Analytics and City Science, 46*(7), 1243–1263. https://doi.org/10.1177/2399808319832612

Bahers, J.-B., Athanassiadis, A., Perrotti, D., & Kampelmann, S. (2022). The place of space in urban metabolism research: Towards a spatial turn? A review and future agenda. *Landscape and Urban Planning, 221*, Article 104376. https://doi.org/10.1016/j.landurbplan.2022.104376

Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., & Liu, B. (2023). Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs. *ISPRS Journal of Photogrammetry and Remote Sensing, 201*, 193–208. https://doi.org/10.1016/j.isprsjprs.2023.05.006

Balsebre, P., Huang, W., Cong, G., & Li, Y. (2024). City foundation models for learning general purpose representations from OpenStreetMap. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 87–97). https://doi.org/10.1145/3627673.3679662

Batty, M. (2024). Digital twins in city planning. *Nature Computational Science, 4*(3), 192–199. https://doi.org/10.1038/s43588-024-00606-7

Cai, Z. (2025). Evolving from rules to learning in urban modeling and planning support systems. *Urban Science, 9*(12), 508. https://doi.org/10.3390/urbansci9120508

Cai, Z., Kwak, Y., Cvetkovic, V., Deal, B., & Mörtberg, U. (2023). Urban spatial dynamic modeling based on urban amenity data to inform smart city planning. *Anthropocene, 42*, Article 100387. https://doi.org/10.1016/j.ancene.2023.100387

Cai, Z., Wang, B., Cong, C., & Cvetkovic, V. (2020). Spatial dynamic modelling for urban scenario planning: A case study of Nanjing, China. *Environment and Planning B: Urban Analytics and City Science, 47*(8), 1380–1396. https://doi.org/10.1177/2399808320934818

Cavicchia, R., & Cucca, R. (2022). Urban densification and its social sustainability. In *The Palgrave Encyclopedia of urban and regional futures* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-030-51812-7_156-1.

Chang, N.-B., Hossain, U., Valencia, A., Qiu, J., & Kapucu, N. (2020). The role of food-energy-water nexus analyses in urban growth models for urban sustainability: A review of synergistic framework. *Sustainable Cities and Society, 63*, Article 102486. https://doi.org/10.1016/j.scs.2020.102486

Chen, M., Li, Z., Huang, W., Gong, Y., & Yin, Y. (2024). Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 319–328). https://doi.org/10.1145/3637528.3671918

Chen, S., & Biljecki, F. (2023). Automatic assessment of public open spaces using street view imagery. *Cities, 137*, Article 104329. https://doi.org/10.1016/j.cities.2023.104329

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., … Arcucci, R. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA Journal of Automatica Sinica, 10*(6), 1361–1387. https://doi.org/10.1109/JAS.2023.123537

Clarke, K. C. (2021). Cellular automata and agent-based models. In *Handbook of regional science* (pp. 1751–1766). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-662-60723-7_63.

Cohen, B. (2006). Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability. *Technology in Society, 28*(1–2), 63–80. https://doi.org/10.1016/j.techsoc.2005.10.005

Fan, C., Xu, J., Natarajan, B. Y., & Mostafavi, A. (2023). Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality. *Computer-Aided Civil and Infrastructure Engineering, 38*(14), 2013–2029. https://doi.org/10.1111/mice.12972

Hariram, N. P., Mekha, K. B., Suganthan, V., & Sudhakar, K. (2023). Sustainalism: An integrated socio-economic-environmental model to address sustainable development and sustainability. *Sustainability, 15*(13), 10682. https://doi.org/10.3390/su151310682

He, S., Yu, S., Li, G., & Zhang, J. (2020). Exploring the influence of urban form on land-use efficiency from a spatiotemporal heterogeneity perspective: Evidence from 336 Chinese cities. *Land Use Policy, 95*, Article 104576. https://doi.org/10.1016/j.landusepol.2020.104576

Herold, M., Couclelis, H., & Clarke, K. C. (2005). The role of spatial metrics in the analysis and modeling of urban land use change. *Computers, Environment and Urban Systems, 29*(4), 369–399. https://doi.org/10.1016/j.compenvurbsys.2003.12.001

Ho, Y., Li, L., & Mostafavi, A. (2024). Integrated vision language and foundation model for automated estimation of building lowest floor elevation. *Computer-Aided Civil and Infrastructure Engineering*. https://doi.org/10.1111/mice.13310

Hu, Y., & Han, Y. (2019). Identification of urban functional areas based on POI data: A case study of the Guangzhou economic and technological development zone. *Sustainability, 11*(5), 1385. https://doi.org/10.3390/su11051385

Huang, W., Cui, L., Chen, M., Zhang, D., & Yao, Y. (2022). Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science, 36*(10), 1905–1930. https://doi.org/10.1080/13658816.2022.2040510

Huang, W., Wang, J., & Cong, G. (2024). Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science, 38*(7), 1414–1442. https://doi.org/10.1080/13658816.2024.2347322

Huang, W., Zhang, D., Mai, G., Guo, X., & Cui, L. (2023). Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing, 196*, 134–145. https://doi.org/10.1016/j.isprsjprs.2022.11.021

Janowicz, K., Mai, G., Huang, W., Zhu, R., Lao, N., & Cai, L. (2025). GeoFM: How will geo-foundation models reshape spatial data science and GeoAI? *International Journal of Geographical Information Science, 39*(9), 1849–1865. https://doi.org/10.1080/13658816.2025.2543038

Kang, J., Körner, M., Wang, Y., Taubenböck, H., & Zhu, X. X. (2018). Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006

Karvonen, A., Cvetkovic, V., Herman, P., Johansson, K., Kjellström, H., Molinari, M., & Skoglund, M. (2021). The 'New Urban Science': towards the interdisciplinary and transdisciplinary pursuit of sustainable transformations. *Urban transformations, 3*, 1–13. https://doi.org/10.1186/s42854-021-00028-y

Kutty, A. A., Wakjira, T. G., Kucukvar, M., Abdella, G. M., & Onat, N. C. (2022). Urban resilience and livability performance of European smart cities: A novel machine learning approach. *Journal of Cleaner Production, 378*, Article 134203. https://doi.org/10.1016/j.jclepro.2022.134203

Li, H., Zhao, Y., & Zheng, F. (2020). The framework of an agricultural land-use decision support system based on ecological environmental constraints. *Science of the Total Environment, 717*, Article 137149. https://doi.org/10.1016/j.scitotenv.2020.137149

Li, X., Wen, C., Hu, Y., Yuan, Z., & Zhu, X. X. (2024). Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine, 12*(2), 32–66. https://doi.org/10.1109/MGRS.2024.3383473

Liang, H., Zhang, J., Li, Y., Wang, B., & Huang, J. (2024). Automatic estimation for visual quality changes of street space via street-view images and multimodal large language models. *IEEE Access, 12*, 87713–87727. https://doi.org/10.1109/ACCESS.2024.3408843

Liao, J., Chen, X., & Du, L. (2023). Concept understanding in large language models: An empirical study. In *ICLR 2023 tiny papers*.

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Hong, Y. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science, 31*(8), 1675–1696. https://doi.org/10.1080/13658816.2017.1324976

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., & Li, T. (2022). CLIP4Clip: An empirical study of CLIP for end to end video CLIP retrieval and captioning. *Neurocomputing, 508*, 293–304. https://doi.org/10.1016/j.neucom.2022.07.028

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., … Lao, N. (2024). On the opportunities and challenges of foundation models for GeoAI (vision paper). *ACM Transactions on Spatial Algorithms and Systems, 10*(2), 1–46. https://doi.org/10.1145/3653070

Marey, A., Wang, L. L., Goubran, S., Gaur, A., Lu, H., Leroyer, S., & Belair, S. (2024). Forecasting urban land use dynamics through patch-generating land use simulation and Markov chain integration: A multi-scenario predictive framework. *Sustainability, 16*(23), 10255. https://doi.org/10.3390/su162310255

Meresa, G. A., Mitiku, A. B., & Weldemichael, A. T. (2024). Dynamics and predictability of land use/land cover change using artificial neural network-based cellular automata (ANN-CA): The case of the Upper Awash River Basin, Ethiopia. In A. M. Melesse, M. M. Deribe, & E. B. Zeleke (Eds.), *Land and Water Degradation in Ethiopia: Climate and Land Use Change Implications* (pp. 25–41). Cham: Springer. https://doi.org/10.1007/978-3-031-60251-1_3.

Miao, S., Huang, J., Bai, D., Qiu, W., Liu, B., Geiger, A., & Liao, Y. (2025). *Efficient depth-guided urban view synthesis* (pp. 90–107). https://doi.org/10.1007/978-3-031-73404-5_6

Murphy, K. (2012). The social pillar of sustainable development: A literature review and framework for policy analysis. *Sustainability: Science, Practice and Policy, 8*(1), 15–29. https://doi.org/10.1080/15487733.2012.11908081

Pijanowski, B. C., Brown, D. G., Shellito, B. A., & Manik, G. A. (2002). Using neural networks and GIS to forecast land use changes: A land transformation model. *Computers, Environment and Urban Systems, 26*(6), 553–575. https://doi.org/10.1016/S0198-9715(01)00015-1

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). *Learning transferable visual models from natural language supervision* (pp. 8748–8763). PMLR.

Ren, Y., Lü, Y., Comber, A., Fu, B., Harris, P., & Wu, L. (2019). Spatially explicit simulation of land use/land cover changes: Current coverage and future prospects. *Earth-Science Reviews, 190*, 398–415. https://doi.org/10.1016/j.earscirev.2019.01.001

Rode, P., Floater, G., Thomopoulos, N., Docherty, J., Schwinger, P., Mahendra, A., & Fang, W. (2017). *Accessibility in cities: Transport and urban form* (pp. 239–273). https://doi.org/10.1007/978-3-319-51602-8_15

Shi, Y., Qi, Z., Liu, X., Niu, N., & Zhang, H. (2019). Urban land use and land cover classification using multisource remote sensing images and social media data. *Remote Sensing, 11*(22), 2719. https://doi.org/10.3390/rs11222719

Shukla, A., Jain, K., Ramsankaran, R., & Rajasekaran, E. (2021). Understanding the macro-micro dynamics of urban densification: A case study of different sized Indian cities. *Land Use Policy, 107*, Article 105469. https://doi.org/10.1016/j.landusepol.2021.105469

Silva, E., & Wu, N. (2012). Surveying models in urban land studies. *Journal of Planning Literature, 27*(2), 139–152. https://doi.org/10.1177/0885412211430477

Vivanco Cepeda, V., Nayak, G. K., & Shah, M. (2023). Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems, 36*, 8690–8701.

Wang, J., Bretz, M., Dewan, M. A. A., & Delavar, M. A. (2022). Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. *Science of the Total Environment, 822*, Article 153559. https://doi.org/10.1016/J.SCITOTENV.2022.153559

Wen, R., & Li, S. (2021). A review of the use of geosocial media data in agent-based models for studying urban systems. *Big Earth Data, 5*(1), 5–23. https://doi.org/10.1080/20964471.2020.1810492

Wilkerson, M. L., Mitchell, M. G. E., Shanahan, D., Wilson, K. A., Ives, C. D., Lovelock, C. E., & Rhodes, J. R. (2018). The role of socio-economic factors in planning and managing urban ecosystem services. *Ecosystem Services, 31*, 102–110. https://doi.org/10.1016/j.ecoser.2018.02.017

Wu, A. N., Stouffs, R., & Biljecki, F. (2022). Generative adversarial networks in the built environment: A comprehensive review of the application of GANs across data types and scales. *Building and Environment, 223*, Article 109477. https://doi.org/10.1016/j.buildenv.2022.109477

Wu, H., Li, Y., Lin, A., Fan, H., Fan, K., Xie, J., & Luo, W. (2024). A review of crowdsourced geographic information for land-use and land-cover mapping: Current progress and challenges. *International Journal of Geographical Information Science, 38*(11), 2183–2215. https://doi.org/10.1080/13658816.2024.2379468

Wu, M., Huang, Q., Gao, S., & Zhang, Z. (2023). Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. *International Journal of Applied Earth Observation and Geoinformation, 125*, Article 103591.

Xia, C., Zhang, A., Wang, H., Zhang, B., & Zhang, Y. (2019). Bidirectional urban flows in rapidly urbanizing metropolitan areas and their macro and micro impacts on urban growth: A case study of the Yangtze River middle reaches megalopolis, China. *Land Use Policy, 82*, 158–168. https://doi.org/10.1016/j.landusepol.2018.12.007

Xu, Z., & Zhao, S. (2024). Fine-grained urban blue-green-gray landscape dataset for 36 Chinese cities based on deep learning network. *Scientific Data, 11*(1), 266. https://doi.org/10.1038/s41597-023-02844-2

Yao, Y., Zhu, Q., Guo, Z., Huang, W., Zhang, Y., Yan, X., Dong, A., Jiang, Z., Liu, H., & Guan, Q. (2023). Unsupervised land-use change detection using multi-temporal POI embedding. *International Journal of Geographical Information Science, 37*(11), 2392–2415. https://doi.org/10.1080/13658816.2023.2257262

Yin, J., Dong, J., Hamm, N. A. S., Li, Z., Wang, J., Xing, H., & Fu, P. (2021). Integrating remote sensing and geospatial big data for urban land use mapping: A review. *International Journal of Applied Earth Observation and Geoinformation, 103*, Article 102514. https://doi.org/10.1016/j.jag.2021.102514

Zhou, H., Huang, W., Chen, Y., He, T., Cong, G., & Ong, Y.-S. (2024). Road network representation learning with the third law of geography. *Advances in Neural Information Processing Systems, 37*, 11789–11813. https://doi.org/10.52202/079017-0376

Zhou, Z., Lei, Y., Zhang, B., Liu, L., & Liu, Y. (2023). Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11175–11185).