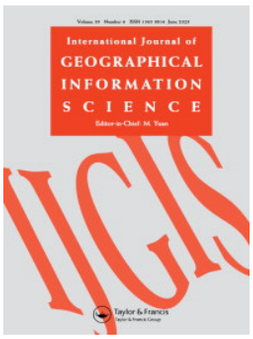

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

LandGPT: a multimodal large language model for parcel-level land use classification with multi-source data

Geyuan Zhu, Mi Tang, Yueheng Ma, Zhihui Hu, Chenglong Yu, Xiang Zhang, Huanjun Hu, Qingfeng Guan & Yao Yao

To cite this article: Geyuan Zhu, Mi Tang, Yueheng Ma, Zhihui Hu, Chenglong Yu, Xiang Zhang, Huanjun Hu, Qingfeng Guan & Yao Yao (20 May 2025): LandGPT: a multimodal large language model for parcel-level land use classification with multi-source data, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2506533

To link to this article: https://doi.org/10.1080/13658816.2025.2506533

Published online: 20 May 2025.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2506533

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

LandGPT: a multimodal large language model for parcel-level land use classification with multi-source data

Geyuan Zhu[a,b], Mi Tang[a], Yueheng Ma[a], Zhihui Hu[a], Chenglong Yu[a,b], Xiang Zhang[a,b], Huanjun Hu[c], Qingfeng Guan[a,d] and Yao Yao[a,b,d,e,f,g]

[a]School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei Province, China; [b]LocationMind Institution, LocationMind Inc., Chiyoda, Tokyo, Japan; [c]School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China; [d]National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, Hubei province, China; [e]Hitotsubashi Institute for Advanced Study, Hitotsubashi University, Kunitachi, Tokyo, Japan; [f]Faculty of Engineering, Reitaku University, Kashiwa, Chiba, Japan; [g]Ministry of Land and Resources of China, Key Laboratory of Urban Land Resources Monitoring and Simulation, Shenzhen, China

ABSTRACT

Actual land parcels vary significantly in size and complexity. Previous studies were limited by existing technical methods for fine-grained land use classification. The emergence of multimodal large language models offers new techniques for image classification, but their application in land use classification remains unexplored. This study presents LandGPT, a multimodal large language model trained on the CN-MSLU-100K dataset, covering fine-grained land use classification of irregular parcels. This study proposes a trans-level discrimination framework to improve LandGPT's ability to classify fine-grained land use. Under this framework, LandGPT achieves a discrimination accuracy of 89.7% and a Kappa coefficient of 0.85 for fine-grained land use categories, outperforming state-of-the-art models with a 48.33% accuracy improvement. In some challenging categories, the improvement reaches nearly 1500%. This study finds that training with multi-source remote sensing image data improved LandGPT's accuracy by 15.79% compared to single-image data. This study explores Prompt engineering based on LandGPT. The optimal prompt paradigm offers fine-grained categories and guides the model for accurate classification, reducing errors from LLM hallucinations. This study pioneeringly explores the application of large language models in the land use domain and offers a new solution for fine-grained land use classification.

ARTICLE HISTORY
Received 7 January 2025
Accepted 12 May 2025

KEYWORDS
Land-use classification; LandGPT; LLM tuning; prompt engineering; multi-source geospatial data

1. Introduction

An accurate and detailed classification of land use is regarded as an indispensable foundation for urban planning and sustainable development. Significant references are provided to reflect regional socioeconomic contributions, thereby establishing a critical

CONTACT Yao Yao yaoy@cug.edu.cn; yaoy@csis.u-tokyo.ac.jp

basis for analyzing the impacts of land use changes on ecological environments (Fang *et al*. 2022; Lyu *et al*. 2022). Under the influence of human activities, urban land use exhibits diverse patterns, including residential areas, commercial zones, and public service facilities (Xia *et al*. 2020). With the acceleration of urbanization processes, land use patterns have become increasingly diverse, and their interactions exhibit growing complexity (Koroso *et al*. 2021; Yan *et al*. 2024). Therefore, designing a more precise land use classification model is crucial to better serve urban planning and promote sustainable development.

Land use classification units are a critical component in the study of land use classification. These units are generally divided into two types: regular plot units and irregular plot units. The partitioning of regular plot units employs a grid-based unit partitioning (GBUP) method. This method achieves partitioning by dividing raw remote sensing images into grid units of equal size (Cao *et al*. 2020; Lu *et al*. 2022; Yao *et al*. 2022). Irregular plot units are partitioned using a block-based unit partitioning (BBUP) method. The BBUP method segments raw remote sensing images based on road network data, creating units of varying scales (Shen and Karimi, 2016; Du *et al*. 2020; Zhu *et al*. 2022). The diversity in shape and scale renders the processing of irregular parcel units considerably more complex and challenging. Conversely, regular parcel units may be characterized by the presence of multiple land use categories, which can induce biases in land use classification. This indicates that the adoption of irregular parcel units as land use classification units is imperative for enhancing classification accuracy.

Scene classification is considered an essential application in remote sensing technology, wherein the identification and categorization of scenes in remote sensing images are automatically performed using computer algorithms. This process holds significant importance in fields such as land monitoring, environmental protection, and urban planning. In recent years, numerous land use scene classification models have been proposed, with most existing studies focusing on the classification of scenes in regular plots. For instance, Yao *et al*. (2022) proposed a land use model combining temporal power data with remote sensing data, employing convolutional neural networks to extract features and classify grid-based data. Lu *et al*. (2022) presented a deep learning-based land use classification model within a unified framework, extracting deep semantic features from remote sensing images and POI via convolutional neural networks.

In the context of irregular plot scene classification, a land classification framework was proposed by Du *et al*. (2020), which integrates remote sensing imagery and social sensing data. This model extracts features through context-based image segmentation, and Latent Dirichlet Allocation (LDA) is employed in conjunction with Support Vector Machines (SVM) for classification. However, such research is often limited by the lack of high-quality irregular plot datasets that can serve as a foundation for research. Due to the pronounced spatial heterogeneity of land use data, the transferability of models developed from specific datasets across diverse regions presents significant challenges (Wu *et al*. 2024). Moreover, the imbalance among land use categories introduces notable obstacles in the training of machine learning models (Wasikowski and Chen, 2010). Land use scenarios are significantly influenced by socioeconomic factors.

Even under similar natural conditions, land use types may be altered due to local influences such as policies and the economy. Land use scenarios are also subject to variability over different time spans. These characteristics significantly elevate the learning cost associated with land use classification models.

Due to the intricacy of land use scene characteristics, the use of only remote sensing images for land use classification is easily affected by the visual resemblance of land use (Zhou *et al.* 2020). Therefore, integrating socioeconomic attributes and natural physical attributes, as well as understanding these interactions, are essential for accurate land use identification. This approach enables the capture of potential dynamics and interdependencies among regions more effectively.

Multimodal large language models (Gao *et al.* 2024; Lee, 2024; Chen et al. 2024a), capable of processing and understanding information from various modalities such as text, images, and videos, have emerged as a novel technology in fields like natural language processing, computer vision, and human-computer interaction. Multimodal large language models employ flexible image processing strategies and can accommodate input images of varying resolutions, offering an innovative technical approach to addressing irregularly shaped plot units. Additionally, multimodal large language models excel in capturing image details, primarily pre-trained on large-scale image datasets, which makes them more effective in handling class imbalances in image classification. Thus, multimodal large language models demonstrate potential in more refined fine-grained (refers to a more precise and detailed division of classification categories, for example, further subdividing 'Agriculture and Nature' into 'Mountain', 'Water', etc.) land use scene classification tasks. However, the application of multimodal large language models in the land use domain remains in its infancy, and their potential has yet to be fully explored. Hence, studying the effective integration of multimodal large language models with multi-source land data and applying them to challenging fine-grained land use classification tasks has become a prominent frontier scientific issue that demands further investigation and discussion.

To address these issues, multimodal large language model technology is applied in this study to land use classification for the first time. A multimodal large language model called LandGPT is explicitly developed for the fine-grained classification task of irregular land parcels. LandGPT is built using the large-scale land use classification dataset CN-MSLU-100K (Wu *et al.* 2024) and adopts InternVL2 (Chen *et al.* 2024a) as the base model. CN-MSLU-100K includes remote sensing image data of irregular parcels across 81 cities in China, significantly reducing the impact of spatial heterogeneity in land use data on model training. As a leading open-source vision model, InternVL2 empowers LandGPT with robust image detail-capturing capabilities.

Based on the LandGPT model, this study proposes a 'trans-level discrimination' framework to enhance fine-grained land use classification accuracy using existing high-precision first-level classification results. A Prompt design paradigm precisely for integrating multi-source data such as remote sensing images and POI is also introduced. This study also employs LandGPT to perform high-precision fine-grained land use classification mapping for selected cities in China.

2. Methodology

Figure 1 shows the flowchart of this study. LandGPT adopts the architecture of widely used open-source MLLMs, specifically the 'Vision Transformer (ViT)-Feedforward Neural Network Layer (MLP)-Large Language Model (LLM)' configuration for image feature extraction, as mentioned in various existing studies (Dong *et al.* 2024; H. Liu *et al.* 2024; Steiner *et al.* 2024). The overall technical roadmap is divided into three main parts: (1) Collection and processing of training data. Training data is divided into image data and non-image data, which are preprocessed appropriately before being used to train the model. (2) Tokens for image and non-image data are generated using dynamic high-resolution processing techniques and the innovative Prompt assembly method designed in this study. These tokens are then fed into the InternLM2-Chat model for training. (3) The LoRA fine-tuning method is used to complete the fine-tuning of the LandGPT model for fine-grained land use classification.

2.1. Dataset and data processing strategy

This study trained LandGPT using 18 datasets, categorized into 15 pretraining datasets (about 110GB) and three fine-tuning datasets (about 10GB). Table 1 summarizes these datasets, which cover various multimodal tasks such as visual question answering, image recognition and annotation, domain-specific image data identification, and POI and Temporal Population data related to data fusion tasks. This diverse dataset

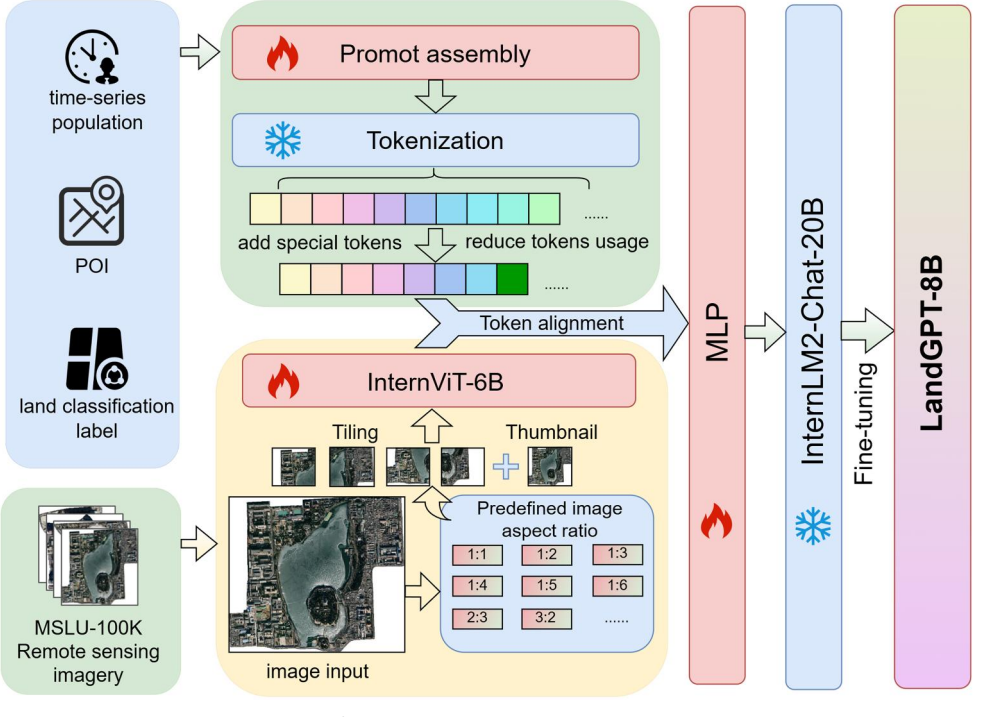

Figure 1. The model structure of LandGPT. 🔥: It indicates that this part was redesigned or retrained. ❄: It indicates that the model in this part was frozen, and the pre-trained results were directly adopted.

Table 1. Datasets used for training LandGPT.

Stage type	Image type	Dataset
Pretraining stage	Visual question answering	DVQA (Kafle *et al.* 2018), GeoQA+ (Cao and Xiao, 2022), GQA (Hudson and Manning, 2019), OCR-VQA (Mishra *et al.* 2019), TestVQA (Singh *et al.* 2019)
	Image recognition and annotation	COCO (Lin *et al.* 2014), VG (Krishna *et al.* 2017)
	Domain-specific image	AI2D (Kembhavi *et al.* 2016), ChatQA (Masry *et al.* 2022), DocVQA (Mathew *et al.* 2021), Synthdog-en (Kim *et al.* 2022), Web-celebrity (Liu *et al.* 2015), Web-landmark (Weyand *et al.* 2020), Wikiart (Saleh and Elgammal, 2015)
Fine-tuning stage	Land use classification multi-source data	CN-MSLU-100K (Wu *et al.* 2024), Amap Poi, Tencent Temporal Population

combination ensures that LandGPT performs reliably in land-use category classification tasks and adapts to diverse tasks requiring language and visual elements. The pre-trained dataset will be incorporated into the model during fine-tuning to improve model robustness.

2.1.1. High-quality land use classification dataset CN-MSLU-100K

Table 1 highlights that multi-source data for land-use classification serves as a vital foundation for training the LandGPT model in land-use classification tasks.CN-MSLU-100K (Wu *et al.* 2024) is a key component for model training at the image data level. This dataset is built from high-resolution visible light remote sensing images sourced from Google Earth Engine (GEE). It integrates multi-source satellite and aerial imagery, including Landsat, Quick Bird, IKONOS, and SPOT5. The temporal range of the images spans from 2019 to 2022, with a spatial resolution of 3 meters. CN-MSLU-100K provides LandGPT with hundreds of thousands of high-precision remote-sensing images of irregular parcels, laying a foundation for fine-grained land-use classification.

The classification criteria of CN-MSLU-100K are based on the 'Urban Land Use Classification and Planning Construction Land Standard' (GB 50137-2011) and the Alibaba Geographical Map POI dataset. The dataset covers 81 major cities in China, with a total area of approximately 983,215 square kilometers. The size of sampled parcels ranges from 598.34 square meters to 64,690,413.16 square meters. In this experiment, 86,240 samples from CN-MSLU-100K were selected, accounting for approximately 72% of the original dataset. The training set accounted for 80%, and the test set accounted for 20%. The dataset includes five first-level land-use classifications: Residential Districts, Commercial Zones, Industrial Land, Public Services, and Agriculture and Nature. It also has 18 second-level classifications: High-rise Residential Buildings, Urban Villages, Rural Architecture and Farmland, Business Tower, Commercial Entertainment, Office Campus, Commercial Market, Shopping Center and Commercial Street, Industrial Park and Factory, Party and Government Institutions, Non-profit Public Institutions (Museum; Stadium; Hospital), Educational and Research Institutions, Parks and Squares, Mountain, Forestland and Grassland, Water, Farmland, and Wasteland. In our subsequent discussion, we collectively refer to these 18 second-level categories as fine-grained categories. For clarity, these categories will be represented in abbreviated form in the following figures, with specific mappings provided in Table 2. Compared to the overall CN-MSLU-100K dataset, the Second-level categories

Table 2. Test accuracy comparison (%) between LandGPT under the Trans-level Discrimination framework and Swin transformer (Liu *et al.* 2021), Resnet50 (He *et al.* 2016), UniRepLKnet (Ding *et al.* 2024) and BDFnet (Wu *et al.* 2024).

SecondLevel	Swin	Resnet	UniRepLKnet	BDFnet	LandGPT	LandGPT†
Tower*	2.99	2.49	0	6.97	46.6	**72.88**
Entertainment*	0	0	0	0	21.95	**70.59**
Market*	6.58	2.19	1.32	9.21	54.96	**76.39**
Education*	44.25	23.98	46.39	51.07	76.79	**97.26**
Farmland	23.25	23.98	29.66	25.64	45.44	**84.57**
Forest*	59.32	50.54	57.66	62.24	72.39	**79.49**
Buildings*	86.89	80.87	86.80	87.94	91.26	**95.8**
Factory*	80.58	84.25	79.54	81.28	93.77	**100**
Mountain	0	0	0	0	57.4	**77.68**
Public*	4.42	2.6	22.92	6.77	62.11	**67.21**
Office*	13.62	10.57	15.45	16.67	30.38	**82.35**
Park*	14.48	8.69	24.72	12.25	58.23	**91.6**
Government*	4.11	0	0.68	3.42	50.57	**77.5**
Rural*	84.5	87.73	83.34	85.85	79.39	**98.29**
Shop*	11.72	1.67	10.04	10.46	64.57	**70.59**
Village*	20.77	8.95	24.6	17.57	**59.91**	56.2
Wasteland	11.78	8.22	4.11	8.49	27.70	**36.67**
Water	24.26	34.55	18.54	21.28	65.28	**69.31**
Total	59.32	54.83	58.86	61.17	75.95	**89.7**

*Indicates the use of second-level category abbreviations. †Indicates the use of the Trans-level discrimination framework. Bold numbers indicate the highest accuracy in the current category. *Tower: Business Tower, Entertainment: Commercial Entertainment, Market: Commercial Market, Education: Educational and Research Institutions, Forest: Forestland and Grassland, Buildings: High-rise Residential Buildings, Factory: Industrial Park and Factory, Public: Non-profit Public Institutions (Museum; Stadium; Hospital), Office: Office Campus, Park: Parks and Squares, Government: Party and Government Institutions, Rural: Rural Architecture and Farmland, Shop: Shopping Center and Commercial Street, Village: Urban Villages.

Villas and High-end Residences and Rural Homestead under the Residential Districts First-level category, as well as the Construction Site Second-level category under the Industrial Land First-level category, were removed. These excluded Second-level categories exhibited significant semantic similarity to High-rise Residential Buildings, Rural Architecture and Farmland, and Industrial Park and Factory within the corresponding First-level categories and were characterized by a relatively small number of samples. During the LandGPT model training phase, manual inspection identified inaccurate labels within these categories, leading to their eventual removal. Furthermore, the first-level categories of transportation facilities and unknown land were excluded from the original dataset. These categories were inconsistent with the core land use theme of this study and predominantly comprised geographically elongated images (e.g. railways and highways), which held limited relevance for land use analysis.

2.1.2. Socioeconomic attribute dataset

The socioeconomic attribute data used in this study includes POI data and Temporal Population data. POI data is sourced from the Amap Open Platform and contains approximately 60.90 million records collected through web crawler techniques. It covers 23 first-level categories and 261 second-level categories, effectively reflecting the spatial structure of urban functions. Tencent Temporal Population data contains user density data for 7 May 2019, with a temporal resolution of 1 hour and a spatial resolution of approximately 1100 meters. These non-image datasets, combined with remote sensing image data, further enhance the overall classification performance of the model.

2.1.3. The combination of multi-source heterogeneous data

Figure 2 shows the format into which fine-tuning data was preprocessed during the training of the LandGPT model. Each remote-sensing image of an irregular parcel is integrated with its corresponding POI information and Temporal Population data into the Prompt structure, ensuring consistency and accuracy between each image and its related non-image data. The unified integration of remote sensing images with their corresponding multi-source data significantly enhances LandGPT's effectiveness and performance in multi-source data fusion tasks. For detailed methods and explanations of the Prompt design, refer to Section 2.2.1.

2.2. Model design

2.2.1. Extraction of non-image data prompt design

The prompt structure is used to structure non-image data integration in LandGPT. Figure 2 shows that the question integrates three main types of non-image data: land-use classification candidate categories, the POI information associated with the current parcel image, and the corresponding Temporal Population data. First, the core task of LandGPT is explicitly defined as selecting the most suitable category from the listed land-use classification candidate categories. Next, the specific formats of POI

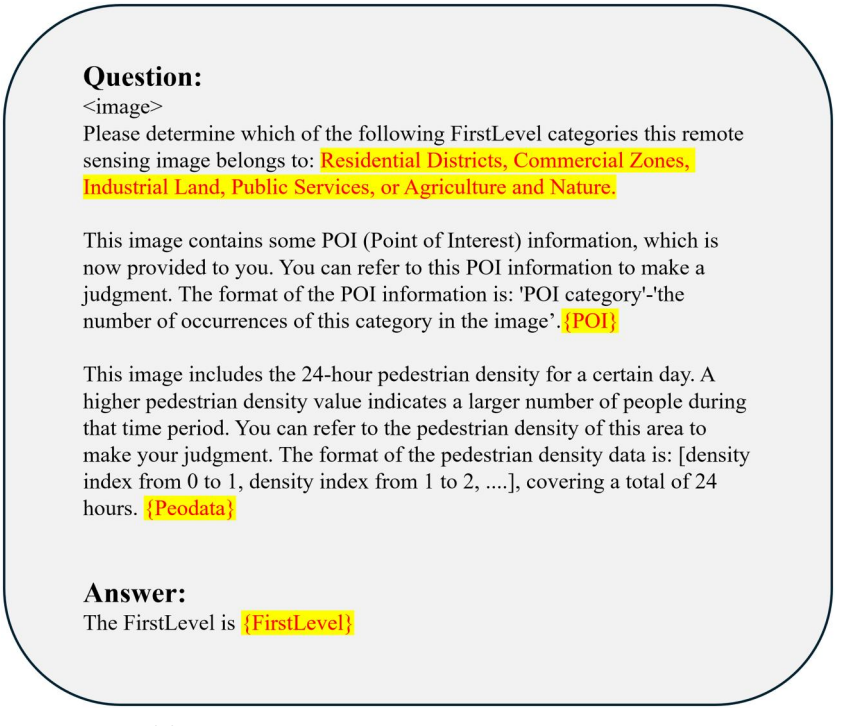

Figure 2. Illustration of fine-tuning data composition. We demonstrate the multi-source data integration employed during the first-level land use classification training of the LandGPT model, where {POI} and {density data} represent the Point of Interest (POI) and Temporal Population data associated with < image>. At the same time {FirstLevel} indicates the first-level land use category of < image>.

information and Temporal Population data are detailed. This approach helps the large language model better interpret the input data structure and combine it with knowledge learned during the pretraining stage to facilitate deep multi-source data integration. This approach has also been validated in applying various fine-tuning models (Vielzeuf *et al.* 2018; Guo *et al.* 2024; Jin *et al.* 2023). After structuring non-image data, tokenizers are used to map prompts to numerical tokens. Measures like adding special token markers are employed to reduce token usage. This approach converts natural language into numbers to lower the cost of model understanding and speed up model training (Kudo and Richardson, 2018).

2.2.2. Dynamic high-resolution images of irregular plots

Inspired by UReader (Ye *et al.* 2023) and the Internvl model (Chen *et al.* 2024a, 2024b, 2024c), LandGPT does not use traditional sampling methods when processing irregular parcel images. Instead, it adopts a dynamic resolution strategy that effectively adapts to varying resolutions and aspect ratios of input images. This approach leverages the flexibility of splitting images into tiles, enhancing LandGPT's ability to process detailed information.

As shown in Figure 3, this process mainly includes the following steps:

2.2.2.1. Match the input image to a predefined aspect ratio. For an image I of size $X \times Y$, its aspect ratio is calculated using the formula $\alpha = \frac{X}{Y}$. The visual encoder of LandGPT for image data only supports images with a fixed input size of $Z \times Z$ (where Z =448). Thus, the most appropriate predefined aspect ratio is determined while minimizing distortion. Additionally, the number of segmented image tiles is constrained within the predefined range $[n_{\min}, n_{\max}]$. To derive the optimal resizing parameters, we define a set Γ for determining the best target aspect ratio:

$$\Gamma = i/j | 1 \leq i, j \leq n, i \times j \in [n_{\min}, n_{\max}] \tag{1}$$

The optimal aspect ratio r_{best} determined by minimizing the difference between the original aspect ratio r and the target aspect ratios r_{target}, given by:

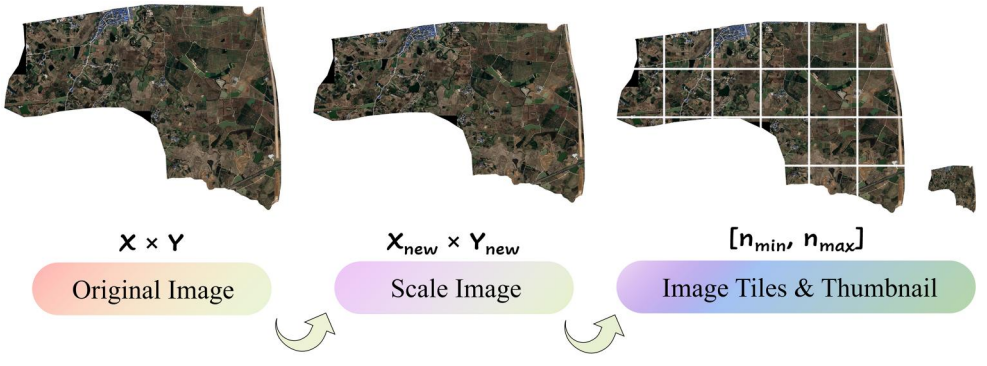

Figure 3. Illustration of processing irregular parcel images through the dynamic high-resolution approach. The input, sized $X \times Y$, involves identifying the optimal aspect ratio i and j for resizing. Ultimately, it is processed into image tiles within the predefined range $[n_{\min}, n_{\max}]$ in quantity and forcibly scaled into a 448*448 thumbnail.

$$r_{\text{best}} = arg\min_{r_{\text{target}} \in \Gamma} \left| r - r_{\text{target}} \right| \quad (2)$$

If multiple aspect ratios exhibit identical differences (e.g. 1:2 and 2:4), priority is assigned to the aspect ratio that ensures the resized area does not exceed twice the original image's area. This approach partially mitigates the risk of low-resolution images suffering from visual quality degradation caused by excessive scaling.

2.2.2.2. The image is resized and further segmented into tiles. Once the optimal aspect ratio is determined, the image is resized to a new dimension $X_{new} \times Y_{new}$, calculated as:

$$X_{new} = i_{best} \times Z, Y_{new} = j_{best} \times Z \quad (3)$$

Here, i_{best} and j_{best} correspond to the parameters associated with r_{best}. Subsequently, the resized image is further segmented into tiles of dimensions $Z \times Z$, with the total number of tiles n given by:

$$n = i_{best} \times j_{best} \quad (4)$$

2.2.2.3. Create thumbnails. If the number of generated tiles n exceeds 1, the original image I is resized to a square dimension of $Z \times Z$ to produce additional thumbnails. The generated thumbnails, combined with the tiles, are used as inputs for the LandGPT model to facilitate global context awareness. If only a single tile is generated, thumbnail creation is unnecessary, and the dynamic high-resolution strategy will automatically bypass this step.

2.2.3. Strong vision encoder

This study employs the Internvl-6B model (Chen et al. 2024a) as the visual encoder module for the LandGPT model. Unlike existing pre-trained visual encoder models (Radford *et al.* 2021; Cherti et al. 2023; Chen *et al.* 2024b), Internvl-6B increases the training image resolution from a fixed 224×224 to a dynamic 448×448 as described in Section 2.2.2 and adopts the 448×448 size as the base tile. This design allows Internvl-6B to perform exceptionally well in handling high-resolution images or images from non-internet sources, such as remote-sensing images of irregular land parcels. Additionally, Internvl-6B enhances model robustness, OCR parsing ability, and high-resolution image processing capabilities by expanding the scale of the pretraining dataset, improving data quality, and increasing diversity. After being processed by Internvl-6B, the converted tokens are aligned with the non-image data introduced in Section 2.2.1. These tokens, at the $<$image$>$tag in Figure 2, go through the MLP layer for the subsequent training.

2.3. Fine-tuning

The LoRA fine-tuning method was used to fine-tune the base model InternLM2 (Chen *et al.* 2024a) of LandGPT. The MLP layers of InternLM2 were unfrozen, while the LLM layers were kept frozen. Table 3 lists some fine-tuning parameters. The model was trained using 8*NVIDIA 4000 Ada GPUs.

Table 3. Selected hyperparameter configurations utilized for fine-tuning.

Parameter	Value	Parameter	Value
trainable params	37,748,736	freeze_llm	True
force_image_size	448	freeze_mlp	False
learning_rate	4e-5	use_llm_lora	16
max_seq_length	4096	optimizer	Adam W (Loshchilov and Hutter 2017)

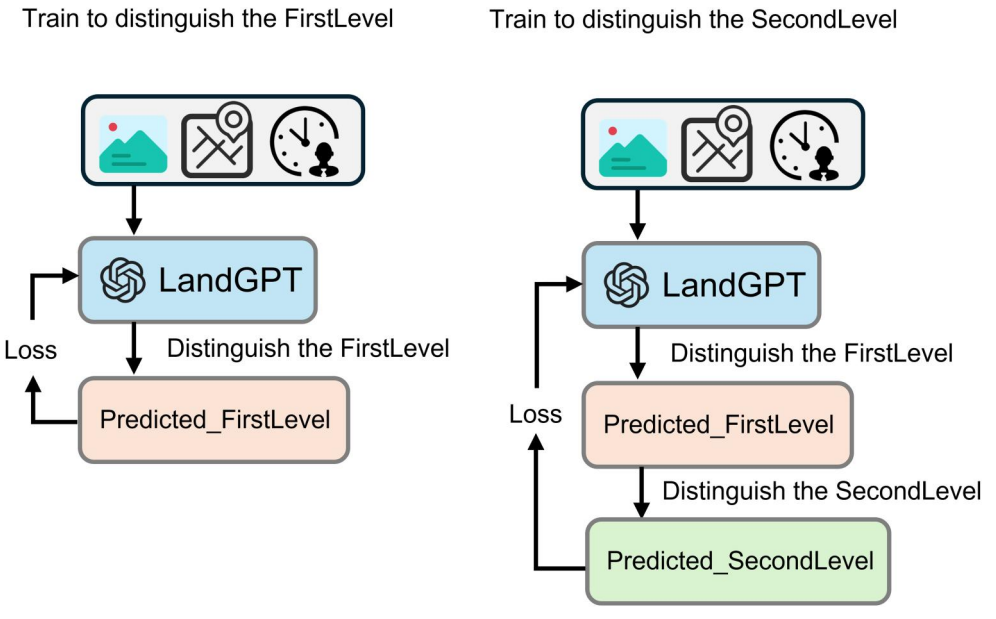

Figure 4. Progressive fine-tuning strategy. In training LandGPT to classify second-level categories, the model first predicts first-level categories, followed by second-level categories, and then calculates the loss.

Figure 4 shows that this study designed an innovative, progressive fine-tuning training strategy. Differing from traditional land use classification models (Saleh and Elgammal, 2015; Yao *et al.* 2022; Zhu *et al.* 2022), LandGPT does not directly predict second-level categories for irregular parcels. Instead, it prioritizes identifying their first-level categories, further refining the classification to second-level categories as needed.

For tasks involving up to 18 different land-use second-level categories, a direct second-level classification requires the model to select one from 18 categories. However, this approach first directs the model to identify the Trans-level category, limiting the choice to 5 Trans-level categories. Then, it selects the Second-level category based on the identified Trans-level category. Since each Trans-level category maps to a maximum of 5 second-level categories, this design effectively transforms the complex '18-choose-1' problem into two simplified '5-choose-1' problems. Experiments confirm that this strategy significantly improves the classification accuracy of LandGPT. This study incorporates a self-correction mechanism to avoid failure in second-level classification caused by errors in Trans-level classification. Results indicate that the self-

correction mechanism allows the model to correct the mistakes during the second-level classification stage, even if Trans-level classification fails. The self-correction mechanism significantly enhances the model's stability and reliability.

2.4. Trans-level discrimination framework

Based on the progressive fine-tuning strategy from Section 2.3 and the flexible instruction mechanism of large language models, this study further proposes a 'Trans-level Discrimination' framework. This framework aims further to enhance the classification accuracy of land-use second-level category tasks to predict the SecondLevel based on the real FirstLevel. Figure 4 shows that during the second-level category classification training, the Predicted_FirstLevel passed to the next-level classification task and was replaced with True_FirstLevel, using actual Trans-level labels during training. Experimental results show that this method fully enhances LandGPT's ability to differentiate between images of various land-use second-level categories while avoiding error propagation caused by biases in Trans-level predictions. Using this framework, the model can directly leverage high-accuracy Trans-level classification results for precise second-level classification without repeating Trans-level classification, thereby reducing computational overhead in practical scenarios.

3. Results

3.1. Land use classification results and comparison

We conducted comparative experiments on the LandGPT model alongside widely used deep learning methods to assess their overall performance in second-level land use classification tasks. To ensure the fairness of the comparative experiments, the multi-source data fusion method proposed by Yao *et al.* (2025) was referred to, and multi-source data fusion training was conducted with remote sensing images, POI, and Temporal Population data on commonly used deep learning models. The comparative experiments utilized a total of 10,000 test set images for accuracy evaluation. Table 2 presents the evaluation results, demonstrating that LandGPT exhibited clear superiority across all 18 second-level categories within the Trans-level Discrimination framework. For second-level categories such as Mountain, Business Tower, Commercial Market, Non-profit Public Institutions (Museum, Stadium, Hospital), and Party and Government Institutions, LandGPT showcased exceptional performance, achieving approximately 1,500% improvement in accuracy over other models. LandGPT also excelled in second-level categories such as Water, Urban Villages, and Parks and Squares, which are traditionally challenging for conventional models to differentiate. The evaluation results further emphasize LandGPT's exceptional image recognition capabilities.

Figure 5 shows that the Trans-level Discrimination framework significantly enhanced the performance of LandGPT. In the Second-level categories of Commercial Entertainment, Farmland, and Office Campus, the integration of the Trans-level Discrimination framework improved LandGPT's accuracy by approximately 200%. Meanwhile, in categories such as Party and Government Institutions, Commercial Market, and Mountain, this framework also improved LandGPT's accuracy by about

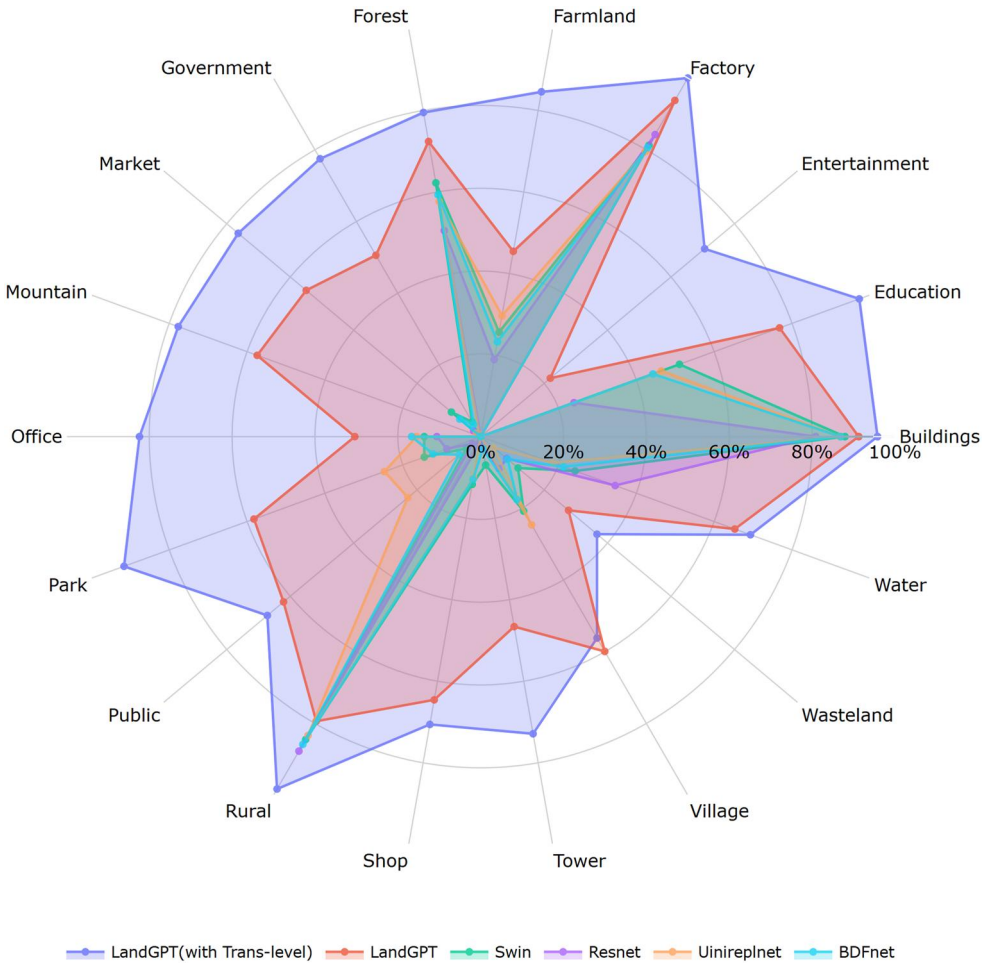

Figure 5. Radar chart of second-level land use classification accuracy across different models.

30%. Compared to all other models, including Swin, which achieved over 60% accuracy in only 3 Second-level categories, LandGPT, with the support of the Trans-level Discrimination framework, completed over 60% accuracy in 16 Second-level categories. Figure 5 clearly illustrates the effectiveness of the Trans-level Discrimination framework, which enhances LandGPT's classification performance in several domains, including commerce, government services, public facilities, education, and nature.

Figure 6 shows that the kappa coefficient of the LandGPT model for land-use Second-level classification tasks reached 0.85 under the 'Trans-level Discrimination' framework. Even without this framework, the kappa coefficient still reached 0.72. These results surpass BDFnet (kappa coefficient 0.54), ResNet (kappa coefficient 0.51), Swin Transformer (kappa coefficient 0.55), and UniRepLKnet (kappa coefficient 0.54). Figure 6 illustrates that LandGPT achieved significant improvements in land-use classification for commercial facilities, public service facilities, and natural areas compared to traditional models.

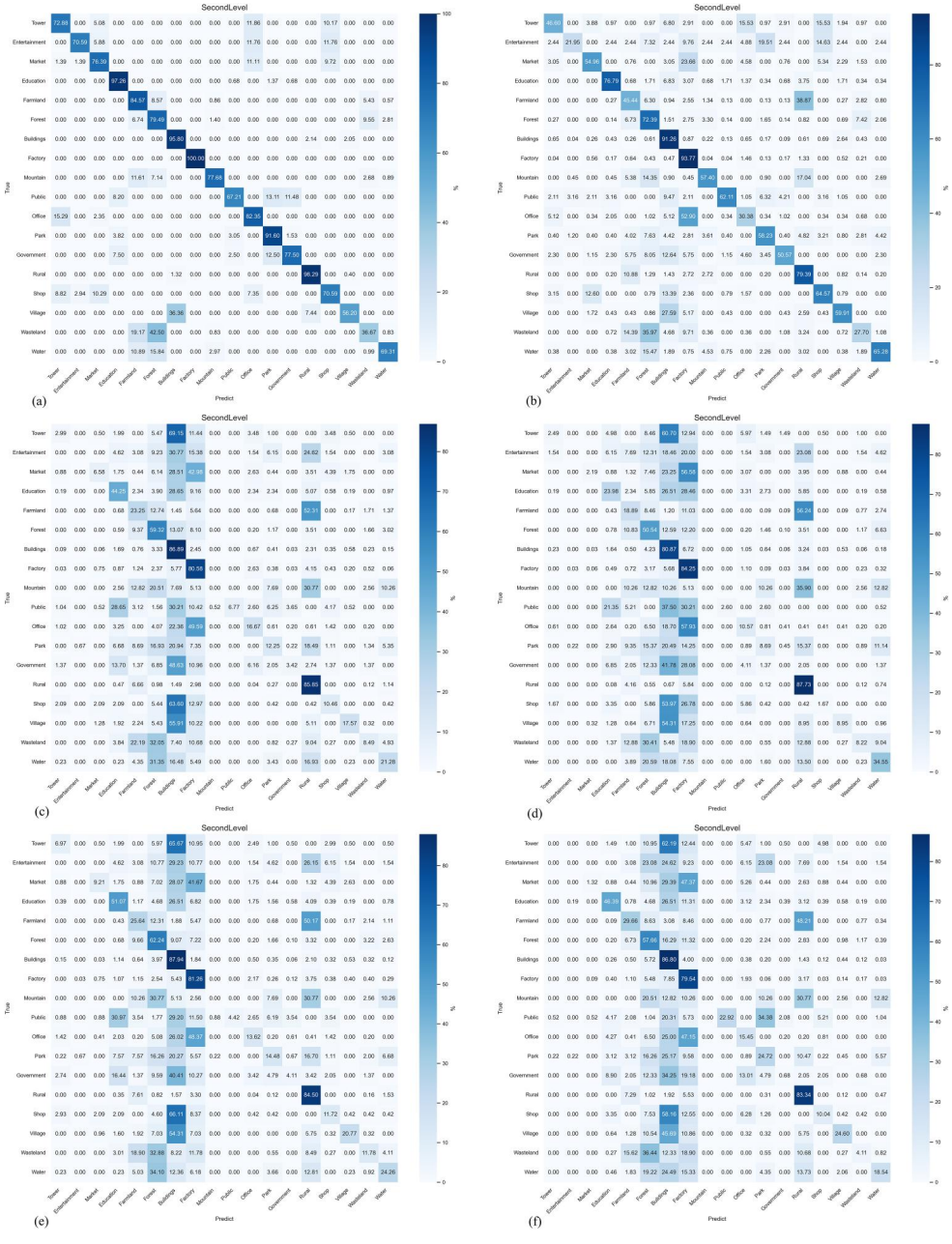

Figure 6. Confusion matrix for second-level land use classification among various models, with data represented in percentages. The corresponding models for different labels are as follows: (a): LandGPT (with Trans-level), (b): LandGPT, (c): BDF_net, (d): Resnet, (e): Swin-transformer, (f): UniRepLKnet.

3.2. Ablation experiment

This study conducted a comparative test between single data source and multi-source data fusion for classification performance to validate the effectiveness of multi-source data fusion. Table 4 shows that for the single remote sensing image data source, LandGPT achieved a test accuracy of 66.3% and a kappa coefficient of 0.62 in the land-use Second-level classification task. When POI and Temporal Population data were integrated, the classification performance improved significantly, with test accuracy increasing by 15.79–76.77% and the kappa coefficient rising to 0.73. These experimental results demonstrate that LandGPT can effectively integrate multi-source data to enhance its performance in land-use Second-level classification.

Multi-source data fusion often encounters the challenge of missing data, such as the absence of POI data or Temporal Population data in specific irregular plots. This study designed a set of targeted experiments to evaluate the performance of LandGPT under data-missing scenarios. We randomly selected 900 test set images from CN-MSLU-100K, including 300 images missing Points of Interest data, 300 images missing Temporal Population data, and 300 images with complete data, and employed the LandGPT model and Swin Transformer model to classify their second-level land use categories. The results are presented in Table 5. In the scenario of missing Temporal Population data, LandGPT achieved a classification accuracy of 75.67%, showing only a 0.43% decline relative to the complete data scenario. When Points of Interest data were missing, its classification accuracy stood at 70%, exhibiting a 7.8% reduction compared to the entire data scenario. Comparatively, the Swin Transformer model recorded a testing accuracy of 54% under conditions of missing Temporal Population data representing a significant 14% decline compared to the entire data scenario. In the case of missing Points of Interest data, its testing accuracy was 52%, dropping significantly by 17%. The results indicate that LandGPT continues to demonstrate high classification accuracy in multi-source missing data environments, significantly surpassing traditional deep learning models in terms of flexibility and robustness.

Table 4. Accuracy table for LandGPT data fusion.

Accuracy metrics	Remote sensing images	Remote sensing images + POI	Remote sensing images + Temporal population	Remote sensing images + POI + Temporal population
Test accuracy	66.3%	74.25% (↑11.99%)	68.88% (↑3.9%)	76.77% (↑15.79%)
Kappa	0.62	0.70	0.64	0.73

Table 5. Test accuracy table of LandGPT and swin-transformer under data-missing scenarios in multi-source data fusion.

Model	Remote sensing images + POI + Temporal population	Remote sensing images + POI	Remote sensing images + Temporal population
LandGPT	76%	75.67% (↓0.43%)	70% (↓7.8%)
Swin	63%	54% (↓14%)	52% (↓17%)

3.3. The impact of prompt engineering on LLM output

Guided by the progressive fine-tuning strategy, we designed two input paradigms with sequential progression. Figure 7 shows that Final input Q2 includes the production of the model for Final input Q1, which is then used for further Second-level category prediction based on the First-level category. The result is used for further Second-level category prediction based on the First-level category. A self-correction statement was added to the prompt to avoid errors in first-level category prediction that affect second-level predictions. Specific changes in Final input Q2 are marked in red. Experimental results showed that adding the self-correction statement increased the test accuracy by 2% and the kappa coefficient by 0.1 on a test set of 10,000 images. This result demonstrates that reasonable Prompt design enables LandGPT to acquire a certain level of self-correction ability, thereby further improving classification performance.

To implement our proposed 'Trans-level Discrimination' framework, one only needs to replace the Predicted_FirstLevel variable in LandGPT's Final Input Q2 with the True_FirstLevel variable and remove the red-marked self-correction statements in both Final Input Q1 and Q2, enabling LandGPT to perform trans-level discrimination. This design simplifies and enhances the efficiency of operating the 'Trans-level Discrimination' framework.

3.4. Typical case analysis

We selected some representative test images, which were chosen from several easily confused second-level categories. The final test results have been presented in Figure 8. All traditional deep learning models were found to perform poorly in distinguishing specific categories. In the identification of 'Urban Villages', 'Non-profit Public Institutions (Museum; Stadium; Hospital)' and 'Business Tower' success in distinguishing these categories from 'High-rise Residential Buildings' was achieved solely by LandGPT. In the 'Mountain' category, farmland was not effectively distinguished from mountainous areas by traditional deep learning models. In the 'Rural Architecture and Farmland' category, insufficient information was captured by traditional deep learning models due to random sampling, resulting in the identification of either farmland or buildings, without combining the two. Typical case analysis diagrams clearly demonstrate that LandGPT has achieved high performance in image parsing and detail capturing. Partial discussions of the results illustrated in Figure 8 have been provided in Section 4.1.

3.5. Mapping land use in typical big cities of China

Figure 9 shows that this study used the LandGPT model combined with irregular parcel images and POI (Point of Interest) data from various cities to generate the fine-grained land-use classification maps of key cities. The model inputs irregular plot images, and POI data in the format shown in Figure 2 and outputs the category to which different plots belong. Figure 9 shows that urban land use in densely populated developed areas is primarily focused on commercial, residential, educational, and

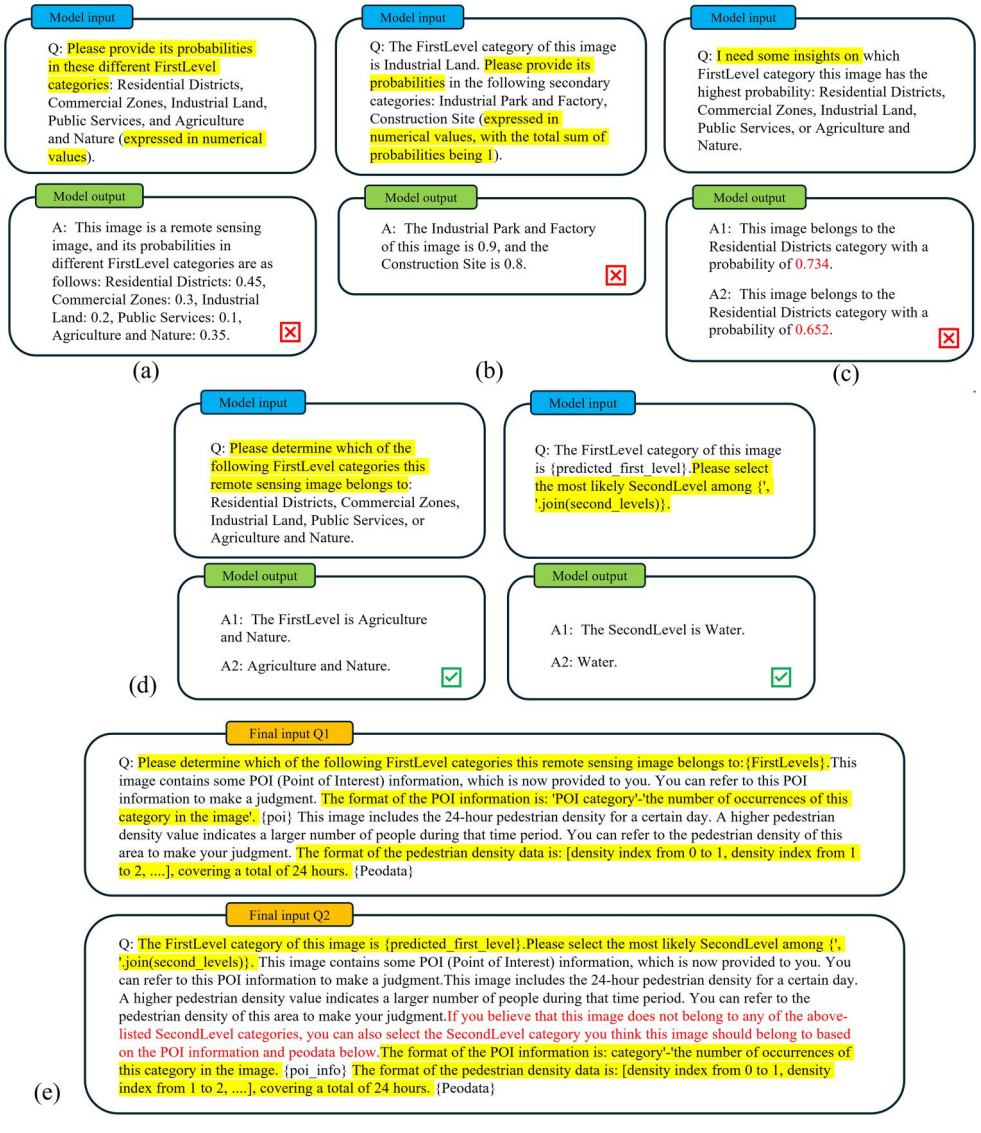

Figure 7. Effects of different prompt designs on model outputs. In (a), (b), and (c), three methods were attempted to query the LandGPT model about the probabilities of different second-level categories for an irregular plot image, but the LandGPT model failed to produce mathematically consistent answers. Specifically, queries in (c) exhibited unstable responses. (d) forms the basis of our finalized prompt composition, whereas (e) fully illustrates the input method of our ultimate prompt design.

public facilities purposes. In economically less developed regions, land is allocated mainly for agricultural activities or remains undeveloped. The eastern plain areas, due to favorable natural conditions, generally have the potential for large-scale urban expansion and industrial development, with relatively well-developed urban infrastructure. In western mountainous areas, more land is usually designated for nature reserves or ecological purposes. In southern coastal areas, land use exhibits a distinctive distribution along coastlines.

Image					
Real SecondLevel	Village*	Mountain	Public*	Rural*	Tower*
Swin	Buildings* ☒	Farmland ☒	Education* ☒	Buildings* ☒	Buildings* ☒
ResNet	Buildings* ☒	Farmland ☒	Buildings* ☒	Farmland ☒	Buildings* ☒
UniRepLKNet	Buildings* ☒	Forest* ☒	Buildings* ☒	Farmland ☒	Buildings* ☒
BDFnet	Buildings* ☒	Farmland ☒	Education* ☒	Farmland ☒	Buildings* ☒
LandGPT	Village* ☑	Mountain ☑	Public* ☑	Rural* ☑	Tower* ☑

Figure 8. Example analysis result diagram. Images featuring category characteristics were selected from Urban Villages, Mountain, Non-profit Public Institutions (Museum; Stadium; Hospital), Rural Architecture and Farmland, and Business Tower for analysis. In this test, LandGPT did not utilize the Trans-level framework.

The fine-grained land use classification map visually shows the spatial distribution of functional urban areas. For educational functions, Beijing's academic facilities are mainly located in Haidian District, Wuhan in Wuchang District, and Nanjing near the Zhongshan area. For industrial functions, Shanghai's industrial areas are primarily in Baoshan, Jinshan, and Pudong New Districts. Wuhan's are in Hanyang and Dongxihu Districts. Hefei's are concentrated at the junction of Yaohai and Luyang Districts. The fine-grained land use classification map, divided by irregular plots, enables a more precise identification of urban functional areas and serves as necessary guidance for urban planning.

4. Discussion

4.1. Technical contributions

For fine-grained land-use classification tasks (corresponding to the 18 second-level categories mentioned above), LandGPT, incorporating a pre-trained multimodal large language model, was proposed in this study. Compared to previous studies, significant improvement was demonstrated in the classification of finer fine-grained land-use category. According to experimental results, the pre-trained multimodal large language model, leveraging learning capabilities from large-scale pre-trained datasets, was found to be significantly better at capturing image details than traditional deep learning models. A dynamic, high-resolution strategy was employed in LandGPT, enabling adaptation to images of different sizes and resolutions with minimal performance loss without relying on conventional lossy sampling methods. On a test set of 10,000 images, test accuracy of 89.70% under the Trans-level Discrimination framework was

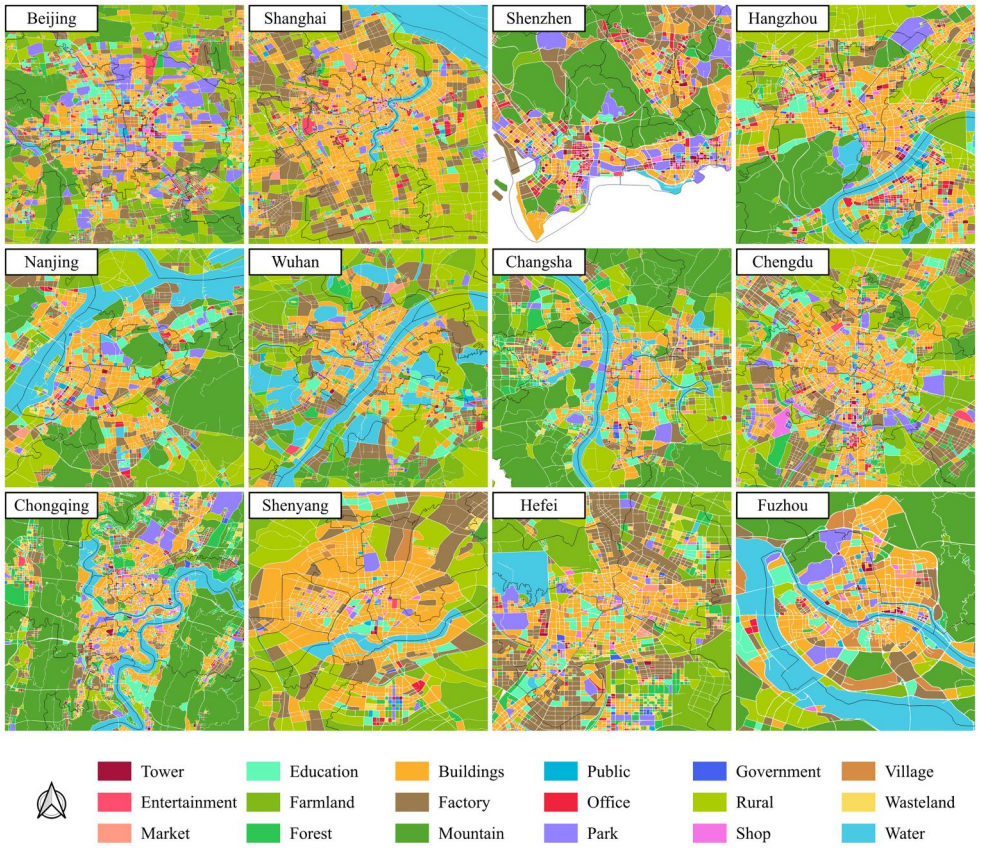

Figure 9. Fine-grained land use classification map for selected major cities in China.

achieved by LandGPT, representing an improvement of approximately 49.5% over Swin and other comparative models.

This study first discovers that multi-source data fusion remains critical in performing land use scenario classification tasks by multimodal large language models, and it addresses the issue of integrating remote sensing image data with other multi-source data in such models. Large language models leverage strong natural language understanding and the ability to process multi-source data of various formats and contents directly, significantly improving LandGPT's learning capability on diverse data sources, with notable advantages in handling POI data. Traditional deep learning models typically rely on embedding techniques to vectorize POI data before inputting it into neural networks, a process that inevitably leads to semantic loss. In contrast, LandGPT supports direct input of POI information, enabling the model to capture semantic information embedded in POI data through tokenization, thereby achieving more efficient multi-source data fusion, mitigating the negative impact of data loss during fusion, and effectively enhancing model robustness. The performance of LandGPT in multi-source data fusion offers substantial reference value for identifying socioeconomic attributes based on irregular land unit cells.

A 'Trans-level Discrimination' fine-grained classification framework is introduced in land use scenario classification for the first time in this study. Trans-level discrimination

leverages existing high-accuracy first-level land use classification results, significantly enhancing LandGPT's capability in fine-grained land use classification to achieve human expert-level performance. This framework design is found to better exploit LandGPT's potential in distinguishing subtle differences between various image types and demonstrates its excellent fine-grained classification performance. This discovery provides a significant reference value for land use analysts.

Prompt design is critical for result accuracy. Poorly designed prompts increase the difficulty of training classification tasks, while effective prompt design plays a key role in optimizing multimodal significant language model performance. A series of experiments are conducted to determine the optimal prompt design strategy for enabling LandGPT to perform fine-grained l land use classification tasks. Figure 7 shows the corresponding experiment results. Various methods are attempted to allow for the model to generate probability values for different fine-grained land use classifications. However, the model exhibits some instability in numerical outputs and shortcomings in understanding specific calculations. This issue is hypothesized to relate to the 'hallucination' phenomenon in large language models, where such models fail to handle operations between numbers accurately (White *et al.* 2023; Marvin *et al.* 2024). Consequently, a final prompt design scheme is adopted, involving concise task requirement summaries, detailed supplementation of multi-source data format information, and explicit requests to generate unique solutions.

Compared to traditional deep learning models, LandGPT demonstrates significant advantages in complex land feature recognition tasks and mitigates the influence of imbalanced data distributions. Figure 6 shows that most traditional models cannot distinguish High-rise Residential Buildings from other types of building infrastructure. For instance, up to 55.91% of Urban Villages are misclassified as High-rise Residential Buildings. This phenomenon may be because traditional models struggle to differentiate buildings in images and comprehend the semantic information of various POIs. During the training of conventional models, they may favor assigning similar types to categories with higher image frequencies in an effort to minimize penalty functions. Within the CN-MSLU-100K dataset, High-rise Residential Buildings (20884 sheets), Forestland and Grassland (6916 sheets), and Rural Architecture and Farmland (14148 sheets) are classified into second-tier categories containing higher sample volumes. The misclassification rate of other categories into these expands considerably.

In contrast to zero-shot models in the field of remote sensing, such as SenCLIP (Jain *et al.* 2025) and RemoteClip (F. Liu *et al.* 2024), LandGPT requires a higher initial training cost but achieves superior fine-grained land use classification accuracy within its specific training domain, coupled with remarkable ease of use. Zero-shot models, by leveraging the shared embedding space of images and text, are characterized by inherent flexibility that enables them to generalize to novel categories without explicit training for specific classes. However, this generalization capability is often associated with a degree of accuracy loss, particularly in fine-grained tasks where subtle distinctions between categories are of critical importance. LandGPT is focused on achieving high accuracy within its training domain through supervised fine-tuning. While this fine-tuning process is computationally intensive, it enables LandGPT to acquire nuanced details and patterns within the training data, thereby excelling in its

specialized area. Furthermore, LandGPT is characterized by significant user-friendliness. In the secondary classification mapping experiment of major Chinese cities conducted in this study, LandGPT successfully completed the secondary classification of land parcels in Wuhan within 20 minutes during the inference process, with memory resource occupation being controlled within 16GB. This renders LandGPT a powerful tool for detailed and accurate land use analysis. Leveraging the technological characteristics of large language models, LandGPT is equipped to support natural language interaction, thereby enabling users to flexibly manipulate the model through natural language for efficient land use scene classification. This feature significantly reduces the complexity of land use classification, thereby increasing its accessibility to non-professional users.

4.2. Limitations and future works

Although the CN-MSLU-100K dataset has high image quality and annotation accuracy, the significant quality gap with other multi-source datasets restricts LandGPT's performance in large-scale land use classification tasks. Figure 9 demonstrates that irregular parcels in China's primary city datasets are divided by road networks. This leads to the generation of ultra-high-resolution irregular parcels in sparsely networked areas, affecting LandGPT's classification results. River parcels are particularly conspicuous. Figure 9 shows that river disconnection is observed in Shanghai, Hangzhou, Nanjing, Wuhan, and Shenyang to varying degrees. The phenomenon occurs because densely populated areas have dense road networks, enabling more accurate segmentation of river parcels. However, in suburban areas, rivers are often grouped with larger land parcels, preventing the model from entirely assigning such parcels to the Water category, leading to visually observable river disconnections. Cities like Changsha, Chongqing, and Fuzhou experience less noticeable river disconnections due to prosperous urban development along riversides.

Certain confusing categories limit the performance improvement of the Trans-level framework in classification tasks. For instance, the second-level category, 'Urban Village' mixes features of high-density residential areas and rural architecture. Additionally, incorporating surrounding urban landscapes during plot division increases the difficulty of the model in learning unique features and achieving precise classifications. Even if the model is directed to assign this second-level category to the 'Residential Districts' first-level category, it is still prone to be confused with 'High-rise Residential Buildings' or 'Rural Architecture and Farmland' under the same first-level category. Table 3 shows that the performance of the 'Village*' category with the Trans-level framework does not significantly exceed performance without this framework.

Dynamic resolution technology enables models to handle input images of different sizes with flexibility, which is advantageous in practical applications. Nonetheless, dynamic resolution adjustment inevitably introduces some level of image distortion. Although this distortion may be visually imperceptible, its precise impact on subsequent task performance has remained challenging to quantify. The task of defining and minimizing this distortion, as well as assessing its specific effects on model accuracy, remains a complex and challenging problem.

Future studies will continue to enhance LandGPT's capabilities in the land use domain. On the one hand, higher-quality data will be constructed by introducing LandGPT and used for continuous model learning. On the other hand, further integration of multi-source geographical information will enable LandGPT to identify parcel-level socioeconomic attributes, advancing its evolution towards a geographic intelligence agent.

5. Conclusion

Focusing on the issue of land use scenario classification, LandGPT, a multimodal large language model, is introduced and designed explicitly for fine-grained land use classification tasks for the first time. The application of multimodal large language models in the domain of land use classification is marked by LandGPT for the first time. Various challenges encountered by conventional models in extracting irregular parcel features and performing fine-grained land use classification tasks have been effectively addressed by LandGPT. A testing accuracy of 89.7% in fine-grained land use classification was achieved by LandGPT, indicating nearly a 50% performance improvement over traditional models. More efficient and accurate analytical tools for urban planning are thereby offered, serving as a valuable reference for land use analysis.

Another significant contribution of this study lies in the finding that carefully designed prompt strategies can significantly boost the multimodal large language model's performance in multi-source data extraction. In the process of multi-source data fusion, the challenge posed by missing data has long been identified as a significant obstacle in related fields. By employing carefully designed prompt strategies, LandGPT has been demonstrated to exhibit strong robustness and adaptability, even in the presence of missing data.

Fine-grained level land use classification maps for several major Chinese cities were produced using LandGPT, resulting in more detailed land use category data. Essential references are provided for professional practitioners in related fields. To further advance the field, LandGPT is planned to be shared through the community, and its wide application is aimed to be promoted.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China [42171466], the Key Laboratory of Earth Surface System and Human-Earth Relations, Ministry of Natural Resources of China [LBXT2023YB03], the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources [KF-2023-08-04] and the 'CUG Scholar' Scientific Research Funds at China University of Geosciences (Wuhan) [2022034].

Notes on contributors

Geyuan Zhu is a graduate student at China University of Geosciences (Wuhan). His research interests are intelligent agriculture, and large language model.

Mi Tang is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining and land use classification.

Yueheng Ma is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining, data-centric urban modeling.

Zhihui Hu is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining and geospatial foundation modelling.

Chenglong Yu is a graduate student at China University of Geosciences (Wuhan). His research interests are GeoAI and Large Language Model.

Xiang Zhang is a graduate student at China University of Geosciences (Wuhan). His research interests are GeoAI and human mobility.

Huanjun Hu is a lecturer at Wuhan Polytechnic University, chief technology officer of Hubei Huajian Xiyuan Agriculture and Animal Husbandry Technology Co. LTD. His research interests are spatio-temporal data mining, intelligent agriculture, and large language model.

Qingfeng Guan is a professor at China University of Geosciences (Wuhan). His research interests are high-performance spatial intelligence computation and urban computing.

Yao Yao is a professor at China University of Geosciences (Wuhan), Hitotsubashi University, and Reitaku University. He has previously served as a Project Researcher at the University of Tokyo. His research interests include spatiotemporal big data mining, social geographic computing, and urban geographic information systems.

Data and codes availability statement

Both the LandGPT model and its training code are publicly available. The LandGPT model is available for download: https://huggingface.co/zhou777/LandGPT, and the code can be downloaded: https://doi.org/10.6084/m9.figshare.28143191. The CN-MSLU-100K dataset used in the paper can be downloaded here: https://doi.org/10.17605/OSF.IO/YAENR.

References

Cao, J. and Xiao, J., 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. *In*: N. Calzolari, *et al.*, eds. *Proceedings of the 29th international conference on computational linguistics,* Gyeongju, Republic of Korea. International Committee on Computational Linguistics,1511–1520.

Cao, R., *et al.*, 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163, 82–97.

Chen, Z., *et al.*, 2024a. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12), 220101.

Chen, Z., *et al.*, 2024b. InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Chen, Z., *et al.*, 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271.

Cherti, M., *et al.*, 2023. Reproducible scaling laws for contrastive language-image learning. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 2818–2829.

Ding, X., *et al.*, 2024. UniRepLKNet: a universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5513–5524.

Dong, X., *et al.*, 2024. InternLM-XComposer2: mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420.

Du, S., *et al.*, 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. *GIScience & Remote Sensing*, 57 (3), 411–430.

Fang, Z., *et al.*, 2022. Impacts of land use/land cover changes on ecosystem services in ecologically fragile regions. *The Science of the Total Environment*, 831, 154967.

Gao, Z., *et al.*, 2024. Mini-InternVL: a flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *Visual Intelligence*, 2 (1), 1–17.

Guo, L., *et al.*, 2024. MKGL: mastery of a three-word language. *Advances in Neural Information Processing Systems*, 37, 140509–140534.

He, K., *et al.*, 2016. Deep residual learning for image recognition. *In: Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hudson, D.A., and Manning, C.D., 2019. GQA: a new dataset for real-world visual reasoning and compositional question answering. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jain, P., *et al.*, 2025. SenCLIP: enhancing zero-shot land-use mapping for sentinel-2 with ground-level prompting. *In: 2025 IEEE/CVF winter conference on applications of computer vision (WACV)*. IEEE, 5656–5665.

Jin, M., *et al.*, 2023. Time-LLM: time series forecasting by reprogramming large language models. arXiv preprint arXiv:2310.01728.

Kafle, K., *et al.*, 2018. DVQA: understanding data visualizations via question answering. *In: Proceedings of the IEEE conference on computer vision and pattern recognition*, 5648–5656.

Kembhavi, A., *et al.*, 2016. A diagram is worth a dozen images. *In*: B. Leibe, *et al.*, eds. *Computer vision – ECCV 2016*. Cham: Springer International Publishing, 235–251.

Kim, G., *et al.*, 2022. OCR-free document understanding transformer. *In*: S. Avidan, *et al.*, eds. *Computer vision – ECCV 2022*. Switzerland, Cham: Springer Nature, 498–517.

Koroso, N.H., *et al.*, 2021. Urbanization and urban land use efficiency: evidence from regional and Addis Ababa satellite cities, Ethiopia. *Habitat International*, 117, 102437.

Krishna, R., *et al.*, 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123 (1), 32–73.

Kudo, T. and Richardson, J., 2018. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

Lee, H.L., *et al.*, 2024. LLaVA-NeXT: improved reasoning, OCR, and world knowledge. LLaVA. https://llava-vl.github.io/blog/2024-01-30-llava-next/

Lin, T.-Y., *et al.*, 2014. Microsoft COCO: common objects in context. *In: Computer vision – ECCV 2014: 13th European conference, proceedings, part v 13,* 6–12 September 2014 Zurich, Switzerland. Springer, 740–755.

Liu, F., *et al.*, 2024. RemoteCLIP: a vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–16.

Liu, H., *et al.*, 2024. Improved baselines with visual instruction tuning. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, Z., *et al.*, 2015. Deep learning face attributes in the wild. *In: Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Liu, Z., *et al.*, 2021. Swin transformer: hierarchical vision transformer using shifted windows. *In*: *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I., and Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv: 1711.05101.

Lu, W., *et al.*, 2022. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sensing of Environment*, 270, 112830.

Lyu, Y., *et al.*, 2022. Mapping trade-offs among urban fringe land use functions to accurately support spatial planning. *The Science of the Total Environment*, 802, 149915.

Marvin, G., *et al.*, 2024. Prompt engineering in large language models. *In*: I. J. Jacob, *et al.*, eds. *Data intelligence and cognitive informatics*. Singapore: Springer Nature, 387–402.

Masry, A., *et al.*, 2022. ChartQA: a benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244.

Mathew, M., Karatzas, D., and Jawahar, C.V. 2021. DocVQA: a dataset for VQA on document images. *In: Proceedings of the IEEE/CVF winter conference on applications of computer visio*n, 2200–2209.

Mishra, A., *et al.*, 2019. OCR-VQA: visual question answering by reading text in images. *In*: 2019 *International conference on document analysis and recognition (ICDAR)*. IEEE, 947–952.

Radford, A., *et al.*, 2021. Learning transferable visual models from natural language supervision. *In: International conference on machine learning*. PmLR, 8748–8763.

Saleh, B., and Elgammal, A., 2015. Large-scale classification of fine-art paintings: learning the right metric on the right feature. arXiv preprint arXiv:1505.00855.

Shen, Y., and Karimi, K., 2016. Urban function connectivity: characterisation of functional urban streets with social media check-in data. *Cities*, 55, 9–21.

Singh, A., *et al.*, 2019. Towards VQA models that can read. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Steiner, A., *et al.*, 2024. PaliGemma 2: a family of versatile VLMs for transfer. arXiv preprint arXiv: 2412.03555.

Vielzeuf, V., *et al.*, 2018. CentralNet: a multilayer approach for multimodal fusion. *In: Proceedings of the European conference on computer vision (ECCV) workshops*.

Wasikowski, M., and Chen, X., 2010. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22 (10), 1388–1400.

Weyand, T., *et al.*, 2020. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2575–2584.

White, J., *et al.*, 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382.

Wu, H., *et al.*, 2024. DCAI-CLUD: a data-centric framework for the construction of land-use datasets. *International Journal of Geographical Information Science*, 38 (11), 2379–2402.

Xia, C., *et al.*, 2020. Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: a case study of five Chinese megacities. *Landscape and Urban Planning*, 193, 103669.

Yan, X., *et al.*, 2024. A multimodal data fusion model for accurate and interpretable urban land use mapping with uncertainty analysis. *International Journal of Applied Earth Observation and Geoinformation*, 129, 103805.

Yao, Y., *et al.*, 2022. Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 106, 102664.

Yao, Y., *et al.*, 2025. Explainable mapping of the irregular land use parcel with a data fusion deep learning model. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1–15.

Ye, J., *et al.*, 2023. UReader: universal OCR-free visually-situated language understanding with multimodal large language model. arXiv preprint arXiv:2310.05126.

Zhou, W., *et al.*, 2020. SO–CNN based urban functional zone fine division with VHR remote sensing image. *Remote Sensing of Environment*, 236, 111458.

Zhu, Q., *et al.*, 2022. Knowledge-guided land pattern depiction for urban land use mapping: a case study of Chinese cities. *Remote Sensing of Environment*, 272, 112916.