Computers, Environment and Urban Systems 132 (2027) 102554

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

# Explainable parcel-level land-use classification with geometrically grounded multimodal reasoning

Geyuan Zhu [a,b], Yao Yao [a,b,c,*], Chenglong Yu [a,b], Xiang Zhang [a,b], Liangyang Dai [d], Zhihui Hu [e], Yueheng Ma [a], Huanjun Hu [f], Qingfeng Guan [a,c]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, Hubei province, China
[b] LocationMind Institution, LocationMind Inc., Chiyoda, Tokyo, Japan
[c] National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430078, Hubei province, China
[d] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, Hubei province, China
[e] Key Laboratory of Geographic Information Science (Ministry of Education), School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[f] School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430023, Hubei province, China

A B S T R A C T

Earth observation and geospatial analysis are increasingly vital for managing urban expansion and environmental changes. However, although large vision-language models show strong potential for interpreting complex landscapes, their black-box nature and tendency to hallucinate limit their use in land-use analysis and related environmental applications. In Earth observation contexts, they often produce explanations that are insufficiently grounded in physical visual evidence. To address this issue, we present the LandAI framework, which enables explainable land-use analysis through geometrically guided reasoning. By introducing GRPO-L1, a multimodal reinforcement learning algorithm, we encourage the model to generate coordinate-based visual references via bounding boxes before producing semantic conclusions. This structured reasoning process helps mitigate conflicts between visual signals and textual priors. In the land-use sample-size comparison, LandAI achieves state-of-the-art classification accuracy with a sample budget equivalent to 25% of the land-use training sample count reported for the comparison model, showing potential for data-scarce regions. Furthermore, LandAI can help identify human label noise in large-scale datasets. Our results show that geometrically guided reasoning can improve the transparency, data efficiency, and reliability of geospatial AI for land-use analysis.

## 1. Introduction

Deep neural networks and large language models are increasingly driving decisions in critical domains. However, the deployment of these black-box AI systems has introduced a critical tension between predictive performance and interpretability (Obermeyer et al., 2019; Rudin, 2019; Suzgun et al., 2025). While these data-driven models drive technological progress, their inherent opacity has raised concerns about transparency and accountability (Chakraborti et al., 2025). As Ribeiro et al. succinctly framed in "Why Should I Trust You?", stakeholders cannot confidently deploy a model without a mechanism to verify its decision logic (Ribeiro et al., 2016). Crucially, this verification requires more than merely plausible justifications; it demands grounded reasoning that strictly aligns the explanation with observable evidence rather than engaging in post-hoc rationalization. This challenge is particularly acute in safety-critical applications, where the demand has shifted from merely obtaining high-precision predictions to demanding transparent, verifiable reasoning processes. Consequently, the field of Explainable AI has evolved, moving beyond post-hoc feature attribution toward developing models with intrinsic interpretability (Mehdiyev et al., 2025).

Recent reviews have documented the rapid evolution of AI-driven systems from task-specific and rule-based models toward increasingly autonomous and adaptive architectures, while identifying persistent challenges in benchmarking, reliability, transparency, ethical risk assessment, and responsible real-world deployment (Jin et al., 2025). Similar concerns have been highlighted in remote sensing, where the widespread use of black-box learning models has increased the need for faithful explanations, systematic evaluation protocols, and explainability methods adapted to the distinctive characteristics of Earth-observation data (Höhl et al., 2024). These challenges underscore the importance of developing GeoAI models whose predictions are not only accurate but also transparently linked to verifiable spatial evidence.

This opacity is especially problematic in GeoAI and urban land-use analysis. Over the past decade, deep learning has revolutionized Earth observation(Bodnar et al., 2025; Chen et al., 2023; Wu et al., 2023; Yang

* Corresponding author at: School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, Hubei province, China.
*E-mail addresses:* geyuan_zhu@cug.edu.cn (G. Zhu), yaoy@cug.edu.cn (Y. Yao), yuchenglong01@cug.edu.cn (C. Yu), z.xiang@cug.edu.cn (X. Zhang), daily0@whu.edu.cn (L. Dai), zhihui.hu@stu.ecnu.edu.cn (Z. Hu), myhovo@cug.edu.cn (Y. Ma), huhuanjun@whpu.edu.cn (H. Hu), guanqf@cug.edu.cn (Q. Guan).

https://doi.org/10.1016/j.compenvurbsys.2026.102554
Received 24 April 2026; Received in revised form 20 October 2026; Accepted 28 October 2026
0198-9715/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

et al., 2026; Zhu et al., 2026; Wu et al., 2023), achieving land-use classification accuracy competitive with expert-annotated benchmarks (Brown et al., 2022; Maslov et al., 2025; Wu, Lee, et al., 2025; Yao, Yang, et al., 2025; Y. Yao et al., 2025). However, even as the field transitions toward multimodal foundation models, this predictive success continues to come at the cost of systemic inscrutability. For urban planners and environmental regulators, a mere classification label such as "Industrial Land" is insufficient; they require the rationale: specifically, identifying which auditable features, such as building morphology, road accessibility, or POI distribution, drive the decision. As noted by Reichstein et al., the leap in accuracy has not yielded a corresponding leap in "process understanding"(Reichstein et al., 2019). This gap prevents stakeholders from translating black-box predictions into actionable analytical insights for planning and management, highlighting a critical missing link in decision-making explainability(Janowicz et al., 2020; Liu & Biljecki, 2022).

To bridge this gap, a paradigm shift is emerging with the advent of multimodal large language models. These models demonstrate emerging reasoning capabilities, generating human-like, step-by-step explanations via Chain-of-Thought prompting. Wei et al.(Wei et al., 2022) demonstrated that Chain-of-Thought elicits latent reasoning capabilities, enabling models to decompose complex problems and articulate their "thinking process": a leap from "what" to "why." Subsequent innovations, such as OpenAI's o1(OpenAI, 2024) and DeepSeek-R1's GRPO training(Guo et al., 2025), have confirmed that optimizing reasoning paths is central to scaling model intelligence. Crucially, this paradigm has recently extended to the vision-language domain, where intermediate reasoning steps are explicitly modeled as mixed vision-language token chains(Alayrac et al., 2022; Gao et al., 2025; Huang et al., 2025; Wang et al., 2026). In remote sensing and urban analysis, recent multimodal large language models have further advanced this direction through instruction tuning, visual prompting, multisource imagery integration, and the coupling of domain-specific and general-purpose models (Zhan et al., 2025; Zhang et al., 2025; Zhou et al., 2025). These developments demonstrate the growing capacity of multimodal models to integrate heterogeneous geospatial information and support spatially informed interpretation across remote-sensing and urban tasks. Together, these advances establish a potential foundation for more verifiable, stepwise reasoning in visual and geospatial tasks.

However, applying this paradigm to geospatial analysis remains challenging. While recent efforts have introduced Chain-of-Thought to GeoAI (Feng et al., 2025; Li, Jiang, et al., 2024; Li, Xu, et al., 2024; Yerramilli et al., 2025), standard reasoning sequences do not inherently guarantee faithfulness in multimodal environments. Without rigorous geometric constraints, these models often rely on text-based prompt engineering or linguistic priors, yielding explanations that are detached from physical pixel-level evidence. This disconnection leads to critical hallucinations: a model might generate highly fluent, reasonable-sounding justifications that blatantly contradict the observable visual reality. Conversely, while state-of-the-art vision-language models have improved at grounding text to image regions (Comanici et al., 2025; Kuckreja et al., 2024; Peng et al., 2023; Yang et al., 2025), they generally perform simple referring comprehension rather than the complex, multi-step geometric reasoning required for Earth observation (Bell et al., 2025; Chen et al., 2024). The central challenge, therefore, is to bridge this divide: How can we compel a model to not only "think" in language but to rigorously anchor its reasoning steps in verifiable visual coordinates?

We address this as a grounded multimodal land-use reasoning problem in which semantic conclusions must be explicitly linked to verifiable visual evidence. To address this fundamental challenge in grounded explainable AI, we present LandAI, a framework that supports explainable land-use reasoning through explicit geometric grounding. Our approach follows a systematic evolution: We first constructed Geo-Base-Thinking-14 K, a large-scale Chinese geospatial Chain-of-Thought dataset, to activate the model's domain-specific logical reasoning. Building on this foundation, we introduce GRPO-L1, a multimodal reinforcement learning algorithm that enforces a strict cognitive progression: visual indexing, followed by geometric localization, and culminating in language reasoning (Fig. 1). As illustrated in Fig. 1, the framework consists of a training stage and an inference stage. In the training stage, Geo-Base-Thinking-14 K is used to activate domain-specific reasoning, after which the policy model is optimized against a reference model through reinforcement learning. In the inference stage, the model aligns multimodal inputs, including visual, POI, and temporal signals, and performs visual indexing and geometric grounding before generating a semantic conclusion. Here, the bounding box serves not merely as a detection output but as an auditable anchor to prevent semantic drift. By structurally rewarding this "look-then-reason" path, we establish a check-and-balance mechanism that penalizes intuitive guessing in favor of verified deduction. This mechanism is particularly important when textual priors conflict with visual evidence, as it encourages the model to remain anchored in localized image evidence rather than follow misleading textual cues. Our experiments demonstrate that this geometrically grounded paradigm fundamentally alters the model's learning mechanism: shifting from rote texture fitting to generalizable rule learning. As a result, LandAI achieves state-of-the-art accuracy in the land-use sample-size comparison with a sample budget equivalent to 25% of the land-use training sample count reported for the comparison model, while exhibiting robust resistance to visual-textual conflicts. Furthermore, we show that this explicit explainability can support human-in-the-loop data quality assessment. By generating transparent reasoning chains linked to physical coordinates, the model can efficiently identify and help correct human label noise within large-scale datasets. Together, these results suggest that explicit geometric grounding is a promising framework for reliable and interpretable geospatial reasoning.

## 2. Methods

### 2.1. Data construction and curation

#### 2.1.1. Geo-Base-Thinking-14 K dataset

Geo-Base-Thinking-14 K was constructed to address the scarcity of domain-specific chain-of-thought data in geospatial science. The corpus was primarily collected from the National Postgraduate Entrance Examination in Geography and GIS across more than 60 Chinese universities and supplemented with canonical geography textbooks. These sources were selected to provide broad topic coverage and relatively high-quality disciplinary reasoning examples. We employed a generate-and-filter distillation pipeline to construct high-quality reasoning chains. For each of the approximately 14,700 collected questions, we used DeepSeek-R1 to generate four distinct chain-of-thought reasoning paths. To encourage diversity, the first three paths were generated without explicit constraints, whereas the fourth was conditioned on the reference answer to improve logical alignment.

Quality control was implemented through a hierarchical LLM-as-a-judge framework. A panel of three commercial large language models, GPT-4o, Qwen-Max, and Doubao-1.5-pro, was first used to establish scoring criteria based on four dimensions: Accuracy, Completeness, Clarity, and Compliance. Because directly applying this multi-model evaluation procedure to the full corpus would be computationally expensive, we adopted a two-stage strategy. We first selected a representative subset of reasoning chains for full jury evaluation. The resulting consensus scores were then used as supervision signals to fine-tune a Qwen2.5-14B model to reproduce the jury's evaluation behavior. The distilled evaluator, named LandAI-Evaluation, achieved more than 80% consistency with the commercial jury when identifying optimal and sub-optimal reasoning paths. We subsequently applied this distilled evaluator to the full corpus and retained the highest-scoring reasoning paths for model training.

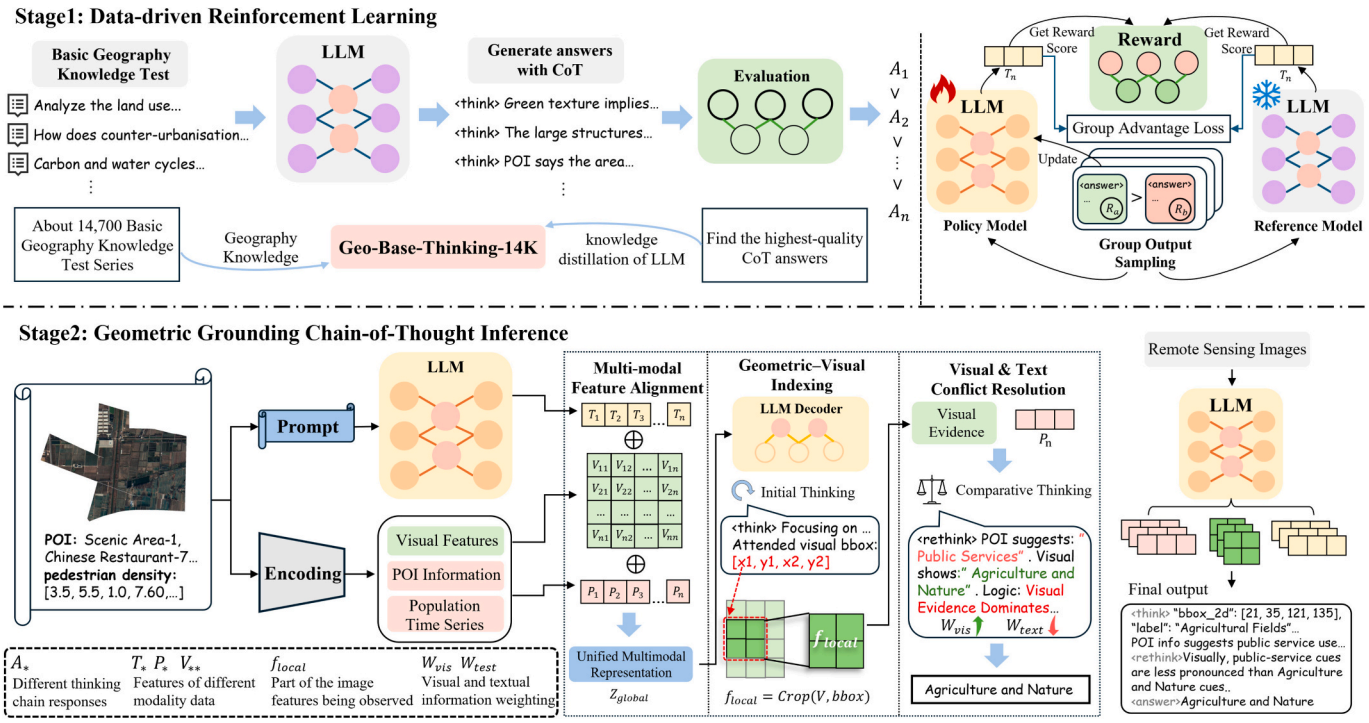



**Fig. 1.** Overview of the LandAI framework for geometrically grounded land-use reasoning. (a) Training stage. Geo-Base-Thinking-14 K is used to activate domain-specific reasoning, followed by reinforcement learning with GRPO-L1. (b) Inference stage. The model aligns multimodal inputs and performs visual indexing and geometric grounding before generating the final land-use label. When visual and textual cues conflict, the framework encourages predictions to remain grounded in localized visual evidence.

### 2.1.2. *General reasoning corpus for cognitive activation*

The final training corpus for LandAI-Base combines Geo-Base-Thinking-14 K with three open-source reasoning datasets, resulting in approximately 334,000 samples in total. These supplementary datasets were included to strengthen complementary reasoning abilities relevant to geospatial analysis.

**OpenR1-Math-96 K.** We adopted the default split of OpenR1-Math-220 K, which contains approximately 96,000 mathematical problems with reasoning traces generated by DeepSeek-R1. This subset was used to strengthen multi-step logical and quantitative reasoning abilities relevant to spatial analysis.

**Chinese-Data-R1-110 K.** This dataset, aggregated from Chinese STEM examinations and logic problems, contains approximately 110,000 samples. It was included to improve reasoning fluency in Chinese while maintaining logical consistency.

**OpenThoughts-114 K.** This dataset contains approximately 114,000 reasoning chains spanning biology, chemistry, and physics. We incorporated it to broaden the model's general scientific reasoning capacity and support transfer of general analytical patterns to geospatial tasks.

### 2.1.3. *Dataset characterization*

We further characterized Geo-Base-Thinking-14 K by comparing its reflective token distribution with several representative open-source chain-of-thought corpora, including OpenR1-Math-96 K, OpenThoughts-114 K, and Chinese-Data-R1-110 K. We focused on self-reflective markers associated with planning, hesitation, and self-correction during reasoning, such as think, hmm, maybe, and alternatively. These markers should not be interpreted as a direct measure of reasoning quality; instead, they provide a useful proxy for the degree to which a corpus preserves explicit reflective processes within chain-of-thought traces.

As shown in Table 1, Geo-Base-Thinking-14 K exhibits a reflective-token profile broadly comparable to that of ChineseData-R1 and higher than the two English-language corpora in several markers. This pattern is consistent with the design goal of the dataset, namely to preserve stepwise analytical structure and self-reflective reasoning for domain-specific adaptation. Together with the hierarchical LLM-as-a-judge filtering procedure described above, this characterization supports the use of Geo-Base-Thinking-14 K as a reasoning-oriented corpus for activating domain-specific cognition in LandAI-Base.

**Table 1**
Comparison of reflective token frequencies between Geo-Base-Thinking-14 K and representative chain-of-thought corpora.

| Aha word | OpenR1-Math 96 K | OpenThoughts 114 K | ChineseData-R1 110 K | Geo-Base-Think 14 K |
|---|---|---|---|---|
| Think | 0.264 | 0.153 | 0.151 | 0.382 |
| hmm | 0.169 | 0.126 | 0.305 | 0.377 |
| maybe | 0.576 | 0.473 | 5.581 | 5.941 |
| Alternatively | 0.429 | 0.410 | 1.817 | 1.262 |

### 2.1.4. *Visual grounding corpora based on CN-MSLU subsets*

The proposed framework was evaluated on complex urban land-use classification tasks using the MSLU-100 K dataset, a large-scale archive encompassing over 100,000 irregular land-use parcels across 81 major Chinese cities. The dataset is characterized by the integration of multi-source heterogeneous data, including high-resolution remote sensing imagery (0.5 m–0.8 m), geospatial point-of-interest vectors, and population time series. To accommodate the distinct phases of the training and evaluation pipeline, we curated three functionally specialized subsets from this repository.

**CN-MSLU-10 K (baseline control).** This subset was obtained through random sampling while preserving the original long-tail class distribution. As reported for MSLU-100 K, real-world land-use categories are highly imbalanced, with residential parcels substantially outnumbering public service or commercial parcels. This subset contains

10,000 samples without class rebalancing and is used exclusively to train the ungrounded baseline model, LandAI-L1-Zero. It therefore serves as a reference setting for evaluating how standard multimodal learning performs under realistic class imbalance.

**CN-MSLU-20 K-CoT (cold start).** To reduce the effect of class imbalance and provide stronger initialization for grounded reasoning, we used a model-in-the-loop bootstrapping strategy. LandAI-L1-Zero was first applied to the full MSLU-100 K archive, and only samples for which predictions matched the original labels were retained. From this filtered pool, we performed stratified sampling to construct a class-balanced subset of approximately 20,000 samples across five primary categories: Residential, Commercial, Industrial, Public Services, and Agriculture & Nature. This subset also contains explicit chain-of-thought traces and was used for the cold-start stage of grounded training.

**CN-MSLU-RL (reinforcement learning).** For the GRPO-L1 stage, we constructed a fixed class-balanced pool of 20,000 samples randomly drawn from MSLU-100 K. To maintain the data-efficiency setting, no additional external data were introduced during later rounds. Instead, we used curriculum learning with weighted random sampling. Hard samples were defined as instances misclassified by the LandAI-L1-sft (the cold-start checkpoint). Across three training rounds, their sampling weights were progressively increased through oversampling with replacement while keeping the total number of training steps constant. The hard-to-easy ratio was adjusted from 1:2 in Round 1 to 2:1 in Round 3. This design increased exposure to ambiguous cases without enlarging the underlying dataset.

The dataset was randomly partitioned at the parcel level, with each parcel assigned exclusively to one data split to avoid parcel-level overlap.

### 2.2. LandAI architecture and base model initialization

#### 2.2.1. Multimodal architecture

The LandAI family is built on Qwen2.5-VL-7B-Instruct, a multimodal large language model designed for high-resolution visual processing. We selected this backbone because its Naive Dynamic Resolution mechanism processes images as variable-length visual token sequences instead of resizing them to a fixed resolution. This is particularly useful for parcel-level land-use analysis, where preserving fine-grained spatial details is important. The architecture combines a Vision Transformer-based visual encoder with a Transformer-based language decoder, allowing visual embeddings and textual tokens to be processed within a unified semantic space.

#### 2.2.2. LandAI-base

We used supervised fine-tuning on the composite reasoning corpus of approximately 334,000 samples to activate domain-specific reasoning in the backbone model. This stage was designed to equip the model with reasoning structures relevant to geographical deduction, such as spatial relations and process-based interpretation, before introducing more complex multimodal alignment tasks. The resulting model, LandAI-Base, serves as the common initialization for all subsequent experiments.

#### 2.2.3. LandAI-L1-zero

LandAI-L1-Zero was trained from LandAI-Base on CN-MSLU-10 K using the standard Group Relative Policy Optimization algorithm. Its reward signal considers answer correctness and textual fluency, without requiring bounding-box generation or coordinate-grounded reasoning. It therefore provides an approximate ungrounded reference for examining model behavior when explicit geometric-grounding information is absent.

#### 2.2.4. LandAI-L1-sft

Existing research demonstrates that models pre-trained with Chain-of-Thought traces perform significantly better in subsequent reinforcement learning stages. Accordingly, we utilized the CN-MSLU-20 K-CoT dataset to perform supervised fine-tuning on LandAI-Base, yielding the LandAI-L1-sft model. This cold-start checkpoint acts as a transitional bridge, equipping the model with the foundational "visual indexing and bounding box generation" output format required for parcel-level classification. LandAI-L1-sft subsequently serves as the initialization point and the reference policy ($\pi_{ref}$) for the GRPO-L1 reinforcement learning stage.

### 2.3. GRPO-L1 reinforcement learning framework

#### 2.3.1. Algorithmic formulation

Rather than representing a single modification to standard GRPO, the core designs of GRPO-L1 are jointly organized to enable the model to learn geometric-grounding signals progressively and reliably. The grounding-aware composite reward provides the primary reference signal by jointly evaluating the required reasoning structure, explicit coordinate generation, final-answer correctness, and reasoning efficiency. The asymmetric clipping strategy constrains destabilizing policy updates while preserving exploration, whereas the efficiency penalty suppresses excessively long and repetitive reasoning. The progressive curriculum strategy further increases the proportion of difficult samples across training rounds, allowing the grounding signal learned from relatively clear cases to be gradually refined on more ambiguous parcels.

Formally, given a multimodal input $x$ (comprising image $I$ and prompt $P$) and a reference policy $\pi_{ref}$ (LandAI-L1-sft), the algorithm samples a group of $G$ outputs $\{o_1, o_2, \ldots, o_G\}$ from the current policy $\pi_\theta$. The optimization objective maximizes the following surrogate function:

$$\mathscr{J}_{GRPO}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}\in B}\frac{1}{G}\Sigma_{i=1}^{G}\frac{1}{|o_i|}\Sigma_{t=1}^{|o_i|}\min\left[r_{i,t}, \widehat{A}_{i,t}, \mathrm{clip}_{\mathrm{asym}(r_{i,t})}\widehat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left(\pi_\theta \,\|\, \pi_{ref}\right) \tag{1}$$

Where $r_{i,t} = \frac{\pi_\theta\left(o_{i,t}\middle|q,o_{i,<t}\right)}{\pi_{\mathrm{old}}\left(o_{i,t}\middle|q,o_{i,<t}\right)}$ denotes the probability ratio between the new and old policies at token $t$. The advantage $\widehat{A}_{i,t}$ is computed using group-relative normalization of the total reward $R_i : \widehat{A}_{i,t} = (R_i\mu_R)/\sigma_R$, where $\mu_R$ and $\sigma_R$ represent the mean and standard deviation of rewards within the group. To accelerate the discovery of correct reasoning paths during the unstable early training phase, we implement an asymmetric clipping strategy, denoted as $\mathrm{clip}_{\mathrm{asym}}$, where the clipping range is defined by $\varepsilon_{plus} = 0.28$ and $\varepsilon_{minus} = 0.2$. The KL penalty coefficient is set to $\beta = 0.04$.

#### 2.3.2. Grounding-aware reward mechanism

The GRPO-L1 updates are driven by a composite reward function $R$, designed to enforce the cognitive trajectory of visual indexing, geometric localization, and linguistic reasoning. The total reward for an output $o$ is defined as:

$$R(o) = R_{\mathrm{format}} + R_{\mathrm{accuracy}} + R_{\mathrm{efficiency}} \tag{2}$$

**Format and Grounding Reward ($R_{\mathrm{format}}$):** This component enforces structural compliance with the proposed "visual indexing-first" paradigm. It is computed as the aggregation of a tag adherence score ($R_{tag}$) and a bounding box verification score ($R_{bbox}$). Leveraging the intrinsic object localization capabilities of the Qwen2.5-VL backbone, we operationalize the geometric reward based on explicit coordinate generation. Specifically, we assess whether the model outputs coordinate tuples in the standardized format $[x_1, y_1, x_2, y_2]$ within the <think> phase. This check evaluates coordinate-format compliance rather than the spatial correspondence between the generated boxes and relevant visual objects. Accordingly, $R_{\mathrm{format}}$ follows a tiered scoring logic: it attains its maximum value of 1.0 if and only if both the structural tags and the coordinate format are valid; partial compliance (e.g., correct structure but missing geometric coordinates) yields 0.5, while failure in both

results in 0. LandAI does not use expert-annotated bounding boxes during training. Instead, this reward design encourages the model to externalize the intrinsic visual-grounding capability of the Qwen2.5-VL backbone through explicit coordinate generation. By coupling this structural constraint with final-prediction accuracy, GRPO-L1 jointly favors responses that provide coordinate-grounded reasoning and correct semantic conclusions. This design enables visually grounded reasoning without requiring dense box-level supervision.

**Answer Accuracy Reward ($R_{\text{accuracy}}$):** This binary outcome reward evaluates the fidelity of the final decision. It assigns a value of 1 solely when the conclusion generated within the <answer> phase strictly matches the ground truth label, and 0 otherwise. This ensures that the grounded reasoning process converges to a correct classification outcome.

**Efficiency and Length Penalty ($R_{\text{efficiency}}$):** To mitigate the over thinking instability where models generate redundant reasoning loops without convergence, we introduce a dynamic penalty function defined as:

$$R_{\text{efficiency}} = \begin{cases} 0 & \text{if } L \le L_{\text{cache}} \\ -\dfrac{L - L_{\text{cache}}}{L_{max} - L_{\text{cache}}} & \text{if } L_{\text{cache}} < L < L_{max} \\ -1 & \text{if } L \ge L_{max} \end{cases} \tag{3}$$

Where $L$ is the sequence length, with $L_{\text{cache}} = 12,000$ and $L_{\max} = 20,000$. This imposes a linear decay on rewards as inference extends beyond the optimal window, encouraging concise and efficient reasoning paths.

For example, consider three candidate responses generated for the same commercial parcel. The first response produces valid bounding-box coordinates, predicts the correct land-use category, and provides a concise reasoning chain; it therefore receives the highest total reward. The second response predicts the correct category but does not generate the required coordinates, so it receives the answer accuracy reward but only a partial format and grounding reward. The third response generates correctly formatted coordinates but predicts an incorrect category, so it does not receive the answer accuracy reward. Through group-relative comparison, GRPO-L1 increases the probability of responses that simultaneously provide explicit visual indexing, correct classification, and efficient reasoning.

#### 2.3.3. Iterative curriculum training strategy

To mitigate the optimization instability characteristic of multimodal reinforcement learning and explicitly enhance model performance on difficult and complex samples, we formulated the training process as a Multi-Stage Dynamic Curriculum. In standard training settings, policies often collapse into local minima by exploiting textual shortcuts. To address this, we define the training dataset $\mathscr{D}$ as a union of easy samples $\mathscr{D}_{easy}$ and hard samples $\mathscr{D}_{hard}$. The subset $\mathscr{D}_{hard}$ is empirically identified as the set of instances $\{x_i, y_i\}$ where the supervised fine-tuned checkpoint $\pi_{sft}$ fails to predict the ground truth $y_i$, formally: $\mathscr{D}_{\text{hard}} = \{(x,y) \in \mathscr{D} | \text{argmax}\pi_{\text{sft}}(y|x) \neq y\}$. The training is structured into three distinct optimization rounds $k \in \{1,2,3\}$. For each round, we construct a sampling distribution $P_k(x)$ governed by a difficulty coefficient $\lambda_k$:

$$P_k(x) \propto \begin{cases} \lambda_k & \text{if } x \in \mathscr{D}_{\text{hard}} \\ 1 - \lambda_k & \text{if } x \in \mathscr{D}_{\text{easy}} \end{cases} \tag{4}$$

We implement this via Weighted Random Sampling with replacement to maintain a constant epoch size while shifting the distribution. The difficulty coefficient $\lambda_k$ follows a stepwise progression. Exploration Phase (Round 1, $\lambda_1 \approx 0.33$): The model is primarily exposed to clear, unambiguous samples to stabilize the visual indexing mechanism and the $R_{\text{format}}$ reward signal. Consolidation Phase (Round 2, $\lambda_2 \approx 0.50$): The ratio of hard samples is increased to force the policy to refine its geometric grounding on ambiguous parcels. Refinement Phase (Round 3, $\lambda_3 \approx 0.66$): The majority of the batch consists of hard samples, compelling the model to resolve complex boundary cases where visual features are subtle or conflicting. This curriculum design allowed the grounding mechanism learned on relatively simple cases to be progressively transferred to more complex urban scenes.

### 2.4. Evaluation protocols

#### 2.4.1. Visual-linguistic adversarial benchmark

We designed a counterfactual evaluation benchmark focused on cross-modal conflict to assess the contribution of geometric grounding beyond standard classification accuracy. We targeted a high-risk confusion pair, Agriculture versus Commercial, and constructed approximately 2000 samples by pairing satellite imagery with clear agricultural textures and misleading POI descriptions associated with commercial land use. This setup creates strong disagreement between visual evidence and textual priors.

We report vision adherence, defined as the proportion of cases in which the predicted class follows the visual ground truth despite misleading textual cues. We also report text-following errors, defined as the frequency with which the model follows the distractor text instead of the image. These metrics quantify whether the model genuinely prioritizes visual evidence when the two modalities conflict.

#### 2.4.2. Human-in-the-loop data refinement pipeline

To examine whether grounded reasoning can support dataset quality assessment, we designed a hierarchical verification pipeline for identifying label noise. We first evaluated LandAI-L1 on the full test set of approximately 13,000 images using a Pass@10 consistency protocol. An image was flagged as a potential label-error case when the model's predictions consistently diverged from the original label across 10 runs. For each flagged instance, we extracted the consensus prediction together with the associated visual indexing reasoning chain.

This output was then passed to a commercial foundation model, Qwen-VL-Max, which acted as a logic auditor. Rather than reinterpreting the remote sensing image from scratch, the auditor evaluated whether the grounded reasoning produced by LandAI-L1 was more plausible than the original dataset label. Only cases supported by the auditor were forwarded for final expert adjudication. This procedure reduces manual review effort by narrowing attention to high-confidence disagreement cases.

## 3. Results

We developed the LandAI family of models to address the critical challenges of interpretability and hallucination in multimodal geospatial intelligence. To rigorously isolate the impact of geometric grounding, we employed a staged training strategy. First, we obtained LandAI-Base by supervised fine-tuning on a composite dataset of 334 k chain-of-thought samples to activate domain-specific reasoning capabilities. This model serves as the common initialization. We then constructed LandAI-L1-Zero as an ungrounded baseline by training it using the standard GRPO algorithm without geometric constraints. Finally, we obtained LandAI-L1 via our proposed GRPO-L1 algorithm, which enforces explicit bounding box grounding. This controlled comparison allows us to determine how a structured cognitive path alters the underlying learning mechanism in resolving complex geographic problems.

### 3.1. Explicit geometric grounding enables data-efficient generalization

Before evaluating the multimodal grounding mechanisms, we first verified the foundational reasoning capacity of the initialized model. Table 2 presents the zero-shot performance of LandAI-Base-7B across general and geographic domains. Driven by the reasoning-centric fine-

**Table 2**
Zero-shot performance benchmarking of LandAI-Base-7B across general and geographic domains.

| Model | General field | | | | Geography field |
|---|---|---|---|---|---|
| | MMMU pro | MATH500 | HumanEval | AIME2024 | GeoTest2025 |
| | | *Closed-source Business MLLMs* | | | |
| (2025/09) Gemini 2.5 pro | 71.2 | 95 | – | 92 | 98.3 |
| (2024/11) GPT-4o | 51.9 | 74 | 90.2 | 9.3 | 92.1 |
| (2025/02) Claude-3.7 sonnet | – | 79.8 | – | 23.3 | 90 |
| | | *Open-source MLLMs* | | | |
| (2025/04) InternVL3-9B | 52.7 | 74.4 | 72.9 | 3.3 | 57.5 |
| (2024/12) InternVL2.5-8B | 45.9 | 56.8 | 68.2 | 3.3 | 44.1 |
| (2025/01)MiniCPM-o2.6-8B | 43.6 | 31.2 | 45.1 | 3.3 | 41.1 |
| (2025/01)Phi-4-multimodal-6B | 37.0 | 70.6 | 71.3 | 10 | 56.1 |
| (2025/01) Qwen2.5VL-32B | 49.5 | 75.2 | 65.2 | 16.7 | 67.7 |
| (2025/01) Qwen2.5VL-7B$_{baseline}$ | 41.2 | 66.2 | 67.3 | 3.3 | 46.7 |
| | | *Our MLLM* | | | |
| LandAI-Base-7B | **44.7** | **74.9** | **66.4** | **16.7** | **93.3** |

tuning on the Geo-Base-Thinking-14 K corpus, the model achieved a significant improvement in geographic reasoning (93.3% vs. 46.7% for the baseline Qwen2.5-VL-7B) on the GeoTest2025 benchmark. This benchmark specifically utilizes restricted questions from the 2025 National Postgraduate Entrance Examination to ensure strict data decontamination, as these materials post-date the pre-training corpus. Accuracy on this benchmark is assessed via the LandAI-Evaluation model. This indicates that LandAI-Base successfully acquired the essential logical infrastructure for spatial analysis, providing a robust starting point for subsequent reinforcement learning.

We next examined land-use classification under different training-set sizes. As shown in Fig. 2a, the training curves were obtained over three curriculum rounds. The LandAI-L1-sft starting checkpoint achieved an accuracy of 71.7%. LandGPT, a recently proposed multimodal model developed specifically for parcel-level land-use classification, is included here as the strongest task-specific baseline and was trained with approximately 80,000 samples(G. Zhu et al., 2025). Under the 20,000-sample setting, LandAI-L1 continued to improve across all three rounds and surpassed the LandGPT reference level at the beginning of Round 2. The 10,000-sample and 5000-sample variants also exhibited

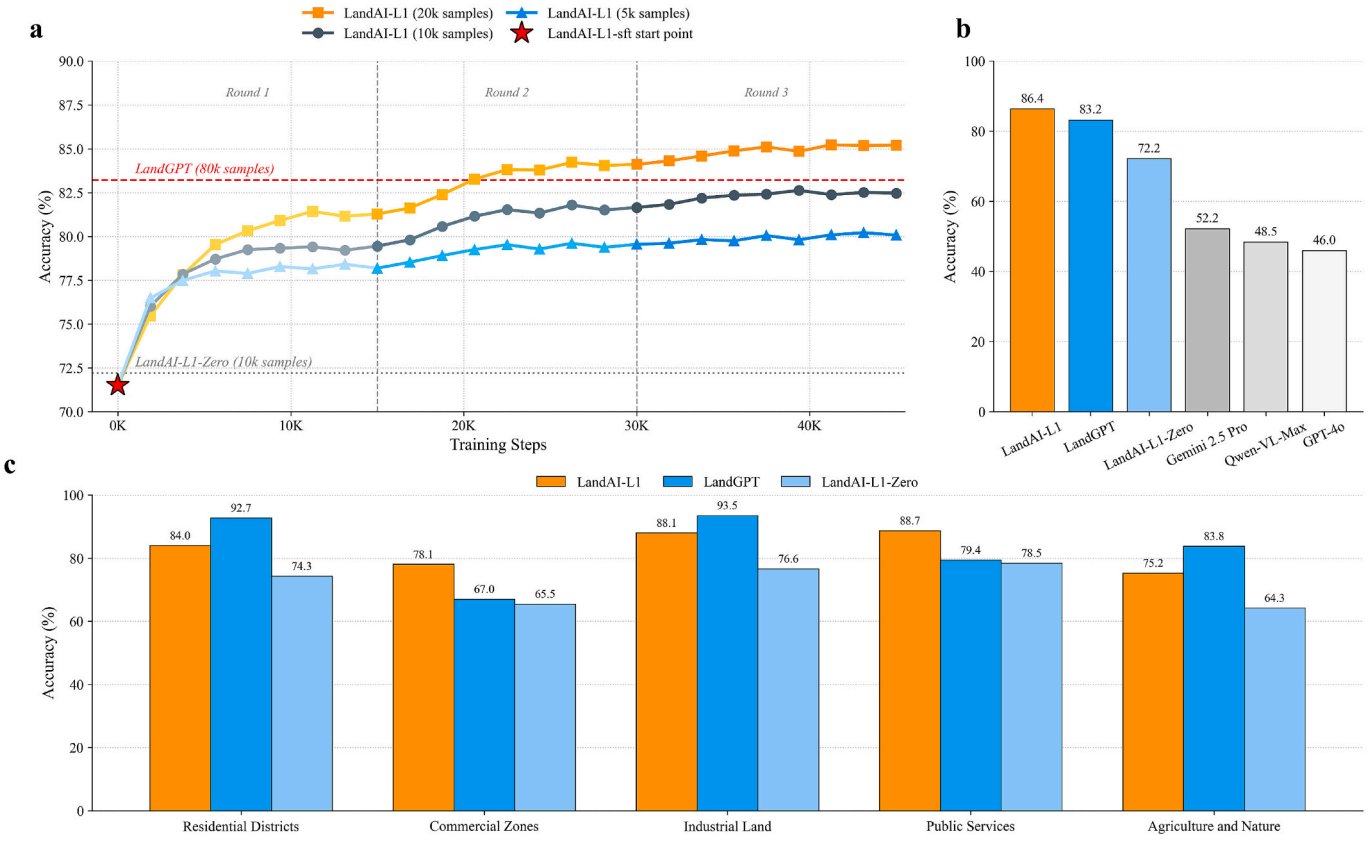



**Fig. 2.** Land-use classification performance and sample efficiency of LandAI-L1. a. Training accuracy under three curriculum-learning rounds for LandAI-L1 with different training-set sizes (5 k, 10 k, and 20 k samples). The dashed red line indicates the performance of LandGPT, a task-specific multimodal baseline trained with approximately 80 k samples. The dotted gray line represents the final test accuracy of LandAI-L1-Zero after standard GRPO training and serves as an approximate reference without explicit geometric-grounding information. b. Overall classification accuracy on the held-out parcel-level test set. LandAI-L1 is compared with LandGPT, LandAI-L1-Zero, and three general-purpose commercial multimodal models. c. Class-wise accuracy comparison across five land-use categories. LandAI-L1 shows improved overall performance and a more balanced accuracy distribution across categories than the comparison models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

progressive gains, although both remained below the 20,000-sample configuration. The dotted gray line in Fig. 2a marks the final test accuracy of LandAI-L1-Zero after standard GRPO training (72.2%). Notably, at the first evaluated checkpoint after reinforcement learning began, the LandAI-L1 variants had already exceeded this reference level, and the performance gap generally widened during subsequent optimization. This early separation indicates that the observed advantage cannot be explained simply by later training steps. Instead, it is consistent with the coordinated role of the grounding reward as the primary geometric reference signal and the supporting strategies that progressively stabilize and refine its optimization.

The overall comparison on the held-out parcel-level test set is shown in Fig. 2b. LandAI-L1 achieved the highest overall accuracy at 86.4%, followed by LandGPT at 83.2% and LandAI-L1-Zero at 72.2%. The three general-purpose commercial models performed substantially worse on this task, with accuracies of 52.2% for Gemini 2.5 Pro, 48.5% for Qwen-VL-Max, and 46.0% for GPT-4o. In combination with the training curves in Fig. 2a, this comparison highlights that the grounded model not only converged to a higher final accuracy, but also required substantially fewer training samples than the strongest task-specific baseline.

Fig. 2c further shows that the advantage of LandAI-L1 was not confined to a small subset of categories. Compared with LandAI-L1-Zero, LandAI-L1 improved accuracy in all five land-use classes, including Residential Districts (84.0% vs. 74.3%), Commercial Zones (78.1% vs. 65.5%), Industrial Land (88.1% vs. 76.6%), Public Services (88.7% vs. 78.5%), and Agriculture and Nature (75.2% vs. 64.3%). Relative to LandGPT, LandAI-L1 obtained higher accuracies in Commercial Zones and Public Services, while remaining competitive in Residential Districts, Industrial Land, and Agriculture and Nature. More importantly, the class-wise distribution of LandAI-L1 was more even than that of the comparison models: its accuracies ranged from 75.2% to 88.7%, compared with 67.0% to 93.5% for LandGPT and 64.3% to 78.5% for LandAI-L1-Zero. This narrower inter-class spread, together with the higher overall accuracy, suggests that LandAI-L1 provides the most balanced performance profile among the compared models and remains more robust across heterogeneous land-use categories.

### 3.2. Reasoning convergence and curriculum-based optimization

Explicit geometric grounding changed not only the final accuracy but also the training dynamics of multimodal reasoning. Fig. 3a compares the evolution of average chain-of-thought length for LandAI-L1 and the ungrounded baseline LandAI-L1-Zero during reinforcement learning. LandAI-L1 reduced its chain-of-thought length rapidly within the early training stage and then remained stable at a substantially lower level. In contrast, LandAI-L1-Zero showed a slower decline and remained much longer throughout training. The shaded region in Fig. 3a marks a period in which the ungrounded baseline entered an unstable plateau, with visible fluctuations and temporary rebounds in chain-of-thought length, whereas LandAI-L1 had already converged to a shorter and more stable reasoning pattern. This difference indicates that geometric grounding improves optimization stability and reduces redundant intermediate reasoning.

Fig. 3b compares curriculum-based training with a standard sampling strategy across three rounds. In the upper panel, the curriculum schedule maintained a consistent advantage in overall accuracy from Round 1 onward, and the gap widened further in later rounds. In the lower panel, the difference was more pronounced on hard samples. After the proportion of difficult cases increased in Round 2 and Round 3, the curriculum-based model improved substantially, whereas the standard strategy showed only limited gains. By the end of training, curriculum learning reached 49.8% accuracy on hard samples, while standard training reached 20.1%. Together, Fig. 3a and Fig. 3b indicate that geometric grounding and staged curriculum optimization jointly support faster convergence, more stable reasoning trajectories, and stronger performance on difficult parcels.

### 3.3. Multimodal conflict resolution and qualitative grounding analysis

In multimodal geospatial analysis, a critical challenge is modality imbalance, in which large vision-language models tend to prioritize linguistic probability over physical visual presence, thereby producing hallucinated interpretations under cross-modal conflict. To assess the role of GRPO-L1 in alleviating this imbalance, we selected Agriculture

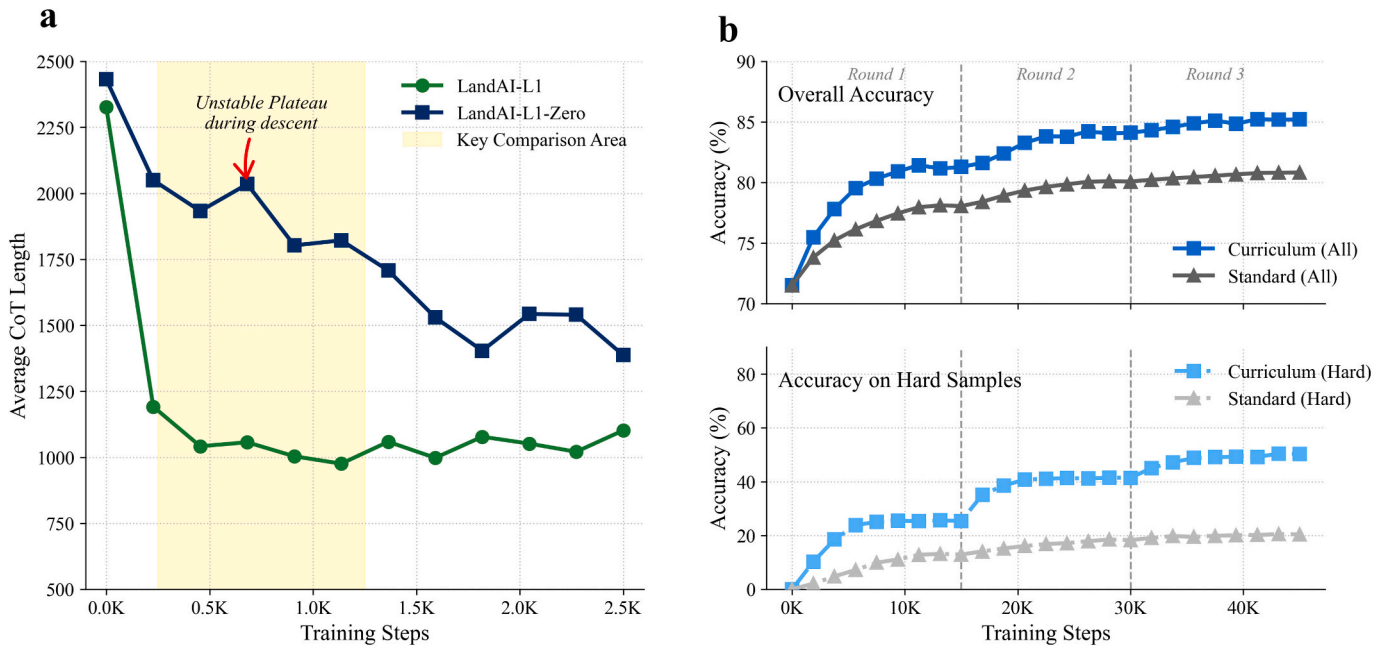



**Fig. 3.** Training dynamics of LandAI-L1 under geometric grounding and curriculum learning. a. Average chain-of-thought (CoT) length over training steps for LandAI-L1 and LandAI-L1-Zero. LandAI-L1-Zero exhibits an unstable plateau during early optimization, while LandAI-L1 reaches a shorter and more stable reasoning regime. b. Comparison between curriculum-based and standard training across three rounds. The upper panel reports overall accuracy, and the lower panel reports accuracy on hard samples. Hard samples are defined as instances misclassified by LandAI-L1-sft (the cold-start checkpoint).

and Nature and Commercial Zones as the test pair because they show the largest discrepancy between visual and textual characteristics in our dataset. Specifically, 2000 randomly sampled Agriculture and Nature images were paired with the POI and population time-series inputs from Commercial Zones parcels, creating a setting in which strong visual evidence was intentionally combined with misleading non-visual context. The decision distributions are summarized in Fig. 4a. LandAI-L1 achieved the highest vision-following rate at 37.0%, compared with 7.3% for LandAI-L1-Zero, 10.2% for LandGPT, and 29.0% for Gemini 2.5 Pro. Its text-following error rate was 25.5%, lower than that of Gemini 2.5 Pro (43.7%) and substantially lower than the proportion of misleading text-dominated decisions observed in the ungrounded models. These comparisons indicate that GRPO-L1 improves the model's resistance to misleading non-visual information relative to the evaluated baselines. The absolute vision-following rate also suggests that there remains room to further strengthen this capability under more challenging conflict settings.

Fig. 4b shows a representative case from this benchmark. In the auxiliary inputs, the POI and temporal signals contain cues that could support a commercial or residential interpretation, whereas the image itself is dominated by agricultural and natural surface patterns rather than dense built-up structures. The reasoning trace produced by LandAI-L1 explicitly recognizes this inconsistency: it first considers the contextual evidence, then re-examines the visible land-cover structure, and finally assigns the parcel to Agriculture and Nature because the image-grounded evidence does not support a commercial interpretation. This case illustrates how geometric grounding changes the decision process from text-led inference to image-grounded conflict resolution, providing a concrete explanation for the quantitative advantage observed in Fig. 4a.

To further assess whether the generated bounding boxes identify spatially meaningful evidence beyond coordinate-format compliance, three domain experts independently examined the remote-sensing images, the complete model-generated CoTs, and the overlaid bounding boxes from 445 samples. They evaluated whether each box spatially corresponded to the visual evidence described in the CoT and whether that evidence was relevant to the parcel-level land-use judgment. A score of 2 indicates a high-quality localization that accurately covers the relevant evidence described in the CoT. A score of 1 indicates a partially relevant localization in which the described evidence is relevant, but the generated box covers it only partially, imprecisely, or too broadly. A score of 0 indicates that the box does not correspond to the described evidence or mainly covers an unrelated region. The localization-quality rating was assessed independently of whether the final prediction matched the original dataset label. The final rating for each sample was determined by majority vote. Inter-rater agreement before consensus was quantified using ordinal Krippendorff's alpha, yielding $\alpha = 0.84$. According to the final ratings, 18.3% of the bounding boxes were classified as high-quality localizations, 46.2% as partially relevant localizations, and 35.5% as incorrect localizations. Overall, 64.5% contained relevant or partially relevant spatial information, with an average quality score of approximately 0.83 out of 2. These results indicate that the generated boxes can provide spatial references for inspecting the visual evidence described in the reasoning traces, although partially relevant boxes offer less precise support than high-quality localizations. The complete evaluation protocol and 18 representative cases, comprising six high-quality, six partially relevant, and six incorrect localizations, are provided in Supplementary Note 5 and Supplementary Fig. 2.

### 3.4. Explainable reasoning facilitates human-in-the-loop data correction

Large-scale parcel-level land-use datasets inevitably contain label noise, particularly in visually ambiguous or mixed-use scenes. Fig. 5 shows how the grounded reasoning trace produced by LandAI-L1 can be used to support a human-in-the-loop verification workflow for dataset refinement. The process begins with a Pass@10 consistency filter applied to the original dataset. Cases with repeated disagreement between the model prediction and the original label are retained as high-confidence discrepancies. For each retained case, the model output, including its chain-of-thought reasoning, is forwarded to a cross-model verification stage, after which only supported cases are sent for final expert review. This design converts the model from a passive classifier into an interpretable screening tool for identifying potentially mislabeled parcels.

The practical value of this workflow is reflected in both the reduction of manual screening effort and the quality of the identified cases. In the full test set of more than 13,000 parcels, 1026 cases were first flagged by

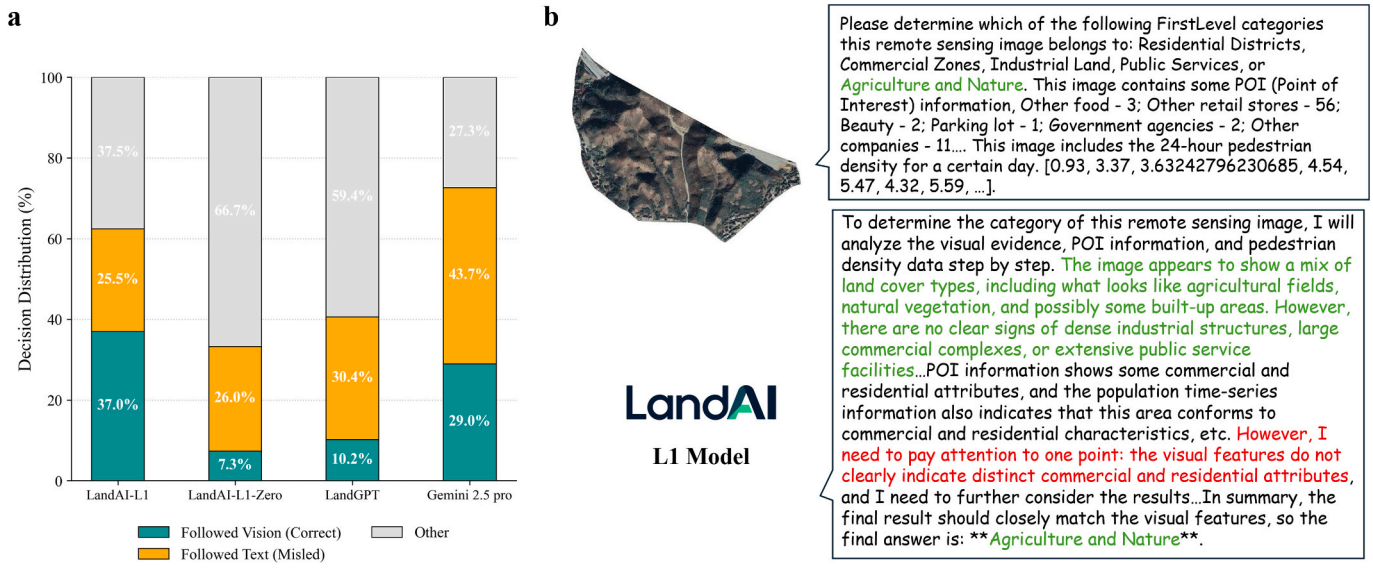



**Fig. 4.** Multimodal conflict benchmark and a representative grounded reasoning case. a. Decision distribution when Agriculture and Nature images are paired with misleading Commercial Zones contextual inputs. Green indicates vision-following correct predictions, orange indicates text-following errors, and gray indicates other outcomes. b. Representative case from the conflict benchmark. LandAI-L1 assigns the parcel to Agriculture and Nature despite misleading non-visual cues. Green text marks reasoning statements supporting the correct image-grounded decision, and red text marks the key correction step where conflicting contextual cues are overridden. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

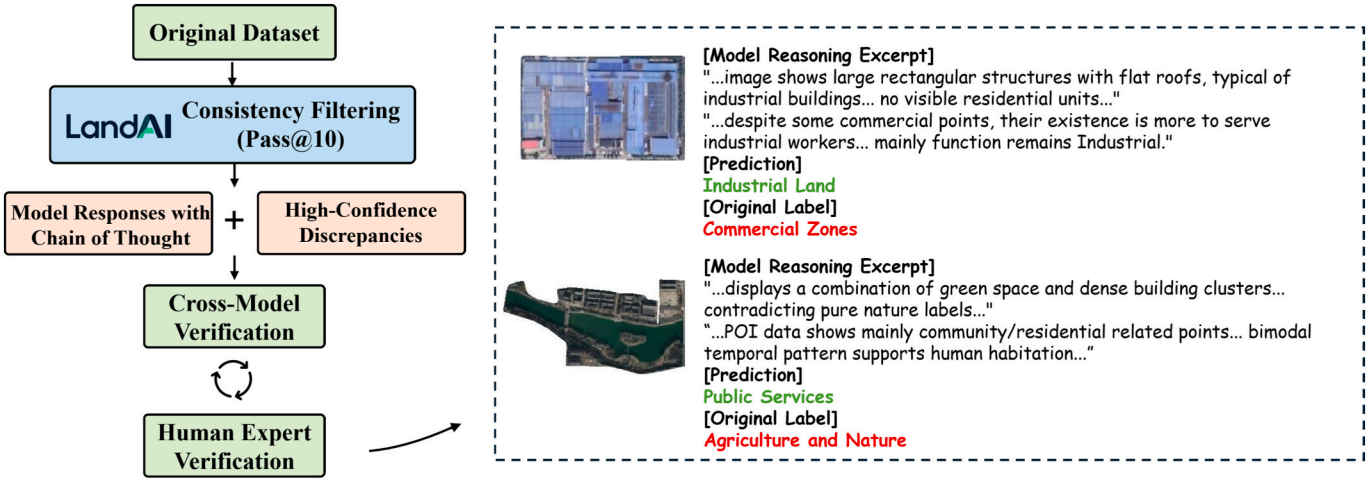



**Fig. 5.** Human-in-the-loop workflow for label-noise detection and representative correction cases. Left: The verification pipeline begins with Pass@10 consistency filtering on the original dataset. Model responses with chain-of-thought reasoning are combined with high-confidence discrepancies, followed by cross-model verification and final human expert review. Right: Representative cases identified by the workflow. The excerpts show how LandAI-L1 challenges the original labels using image-grounded reasoning. Green text indicates the model prediction, and red text indicates the original dataset label. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the consistency filter. Cross-model verification produced a set of high-probability candidates; after removing technically invalid cases, 297 samples were submitted for final expert review, which confirmed 292 mislabeled parcels, accounting for approximately 3% of the test set. The examples in Fig. 5 illustrate the type of discrepancy isolated by the framework: in one case, the model identifies large flat-roofed buildings consistent with industrial land despite an original label of commercial zones; in another, it recognizes dense building clusters and supporting temporal patterns that are more consistent with public services than with agriculture and nature. These cases show that the reasoning trace provides evidence that can be inspected directly, making the review process more focused and substantially reducing the burden of blind manual auditing.

## 4. Discussion

Geometrically guided reasoning improves the interpretability of parcel-level land-use classification by linking semantic conclusions to localized visual evidence. Unlike traditional interpretability methods that rely on post-hoc pixel-level attribution(Höhl et al., 2024; Nieradzik et al., 2025; Wang et al., 2024), our approach uses coordinate-based evidence localization to construct an inspectable inference trajectory. Corroborating recent findings that such semantic transparency need not compromise predictive power(Shang et al., 2024; Srivastava et al., 2024), our results indicate that LandAI enhances precision particularly in high-ambiguity scenarios, as evidenced by the consistent performance across heterogeneous categories like “Agriculture and Nature” (Fig. 2c). By enforcing a structural constraint where semantic conclusions must be preceded by bounding-box localization, we transform the opaque decision process into an auditable cognitive chain. For urban land-use analysis, this design makes model outputs easier to inspect, question, and potentially use in planning and management workflows(Mai et al., 2025; Wang et al., 2024).

Structural inductive biases can reduce data requirements in fine-grained geospatial analysis. While broader scaling laws posit that model performance generally scales logarithmically with dataset size (Shukor et al., 2025; Q. Yao et al., 2025), LandAI-L1 achieved competitive performance relative to state-of-the-art baselines under the evaluated land-use sample configuration, whose sample count corresponds to approximately 25% of that reported for LandGPT (Fig. 2a). This comparison concerns the approximately 20,000 parcel-level land-use samples in the evaluated LandAI-L1 configuration and the approximately 80,000 land-use training samples reported for LandGPT. The approximately 334,000 general and geographic reasoning samples used to initialize LandAI-Base serve to develop broad reasoning capabilities and are reported separately as data for the preceding reasoning-training stage. We attribute this efficiency to the specific role of geometric constraints: whereas standard end-to-end training often requires massive datasets to learn texture invariance(Burgert et al., 2026), explicit grounding appears to facilitate the more rapid acquisition of object-level rules. This mechanism also highlights a distinct advantage of GRPO-L1 in terms of annotation efficiency. By decoupling geometric grounding from dense supervision, our framework demonstrates that explicit bounding box annotations are not strictly necessary for reinforcement learning. Instead, coordinate-format constraints, combined with answer-correctness rewards, encourage the model to express its pretrained visual capabilities through explicit spatial references in land-use reasoning. This weakly supervised grounding strategy may lower the barrier to deploying interpretable GeoAI in settings where pixel-level annotations are scarce or unavailable(F. Li et al., 2024; Gavrikov et al., 2025). Accordingly, these results highlight LandAI-L1's efficiency in learning parcel-level land-use classification from labeled land-use data, while the composite reasoning corpus provides a separate foundation for general and geographic reasoning.

Coordinate-format supervision and visual grounding play distinct but connected roles in LandAI. Coordinate-format compliance provides a structural condition for spatially explicit reasoning, while its connection to meaningful visual evidence arises from the interaction of the pretrained vision-language backbone, the reasoning structure, and sequence-level optimization. Qwen2.5-VL already possesses pretrained visual localization capabilities. Within LandAI, the coordinates and their corresponding natural-language descriptions of visual evidence are generated jointly in the same CoT and conditioned on the same multimodal representation. The overall reward further combines coordinate-format compliance with answer correctness and structured multimodal reasoning, thereby favoring outputs in which explicit spatial references contribute to a semantically correct conclusion rather than rewarding coordinate strings in isolation. The independent expert audit reported in Section 3.3 provides direct evidence for this connection: three domain experts examined the remote-sensing images, the complete model-generated CoTs, and the overlaid bounding boxes, and evaluated whether the boxes corresponded to the visual evidence described in the CoTs. Overall, 64.5% of the evaluated boxes contained relevant or partially relevant spatial information. The complete protocol and

representative high-quality, partially relevant, and incorrect cases are provided in Supplementary Note 5 and Supplementary Fig. 2. These findings indicate that the coordinate constraint can elicit and organize the model's pretrained localization capabilities into meaningful spatial references. Meanwhile, the remaining incorrect localizations clarify that coordinate-format supervision does not ensure precise localization for every sample. Together, the reward design and expert evaluation explain how LandAI establishes geometrically guided reasoning without relying on dense box-level supervision.

A second advantage of the grounded framework is its improved resistance to misleading non-visual cues. Emerging empirical evidence attributes multimodal hallucinations to a critical modality imbalance, where LVLMs prioritize linguistic probability over visual presence (Ghosh et al., 2025; Huang et al., 2024; Leng et al., 2024; Wu, Zhang, et al., 2025; Zhu, Ji, et al., 2025). In our visual-linguistic conflict experiments, LandAI-L1 maintained a 37.0% vision-adherence rate against misleading textual priors, significantly outperforming the ungrounded baseline's 7.3% (Fig. 4a). The qualitative case in Fig. 4b further shows that contextual signals are not ignored, but are revised when they conflict with the image-grounded interpretation. For parcel-level land-use analysis, this behavior is valuable because POI distributions and temporal activity patterns can provide useful context in many cases, but may also bias the model when they are inconsistent with the observed land-cover structure. Additional analyses extend this evaluation to two highly confusable cross-category settings. The vision-following and distractor-following rates were 17.8% and 13.2%, respectively, for Public Services → Agriculture and Nature, and 14.7% and 18.2% for Industrial Land → Residential Districts. Together with the reciprocal Commercial → Agriculture analysis, these results show that visually guided conflict resolution can be observed across multiple category pairs, while its effectiveness varies with category-specific visual ambiguity and contextual informativeness. Complete results and further analysis are provided in Supplementary Note 8.

Grounded reasoning also has practical value beyond classification, particularly for dataset verification and maintenance. While label noise arising from human subjectivity persists in large-scale datasets(Cao et al., 2024; Pan et al., 2025; Schultz et al., 2025), our model exhibits systemic resistance to such errors by strictly adhering to visual logic. This capability supports a hierarchical verification pipeline in which a commercial foundation model audits LandAI-L1's reasoning chains to isolate high-confidence label errors for expert review (Fig. 5). Because the reasoning chain is explicit, the intermediate auditor can verify the logic rather than re-interpret the image from scratch. This substantially reduces manual review effort and suggests that explainable reasoning can support dataset maintenance and quality control in large geospatial knowledge bases(Gröger et al., 2024; Wu et al., 2024).

We acknowledge specific limitations regarding mechanism boundaries and deployment costs. First, while effective on standard parcels, the visual indexing mechanism may degrade in scenarios with severe visual degradation or complex mixed-use arrangements. Parcel delineation may also truncate relevant structures or combine different functions within a single parcel, complicating the identification of visual evidence even for experts. Such data conditions may compound the model's own recognition and localization errors. See Supplementary Note 2 for qualitative examples of failure cases and Supplementary Note 4 for details of the human verification protocol. Future work will build on the expert-rated high-quality localizations by further verifying and organizing the corresponding images, box coordinates, and evidence descriptions into a dataset for iterative training, with the aim of improving localization quality and the reliability of spatial explanations. Second, although the current evaluation covers more than 100,000 parcels across 81 Chinese cities with diverse urban forms, the parcel-level random split does not directly evaluate city-level or spatially separated generalization, and international transfer remains to be examined. A rigorous cross-regional comparison requires compatible land-use taxonomies and parcel definitions, as well as aligned parcel-level remote-sensing, POI, and temporal population data, which are not consistently available across regions. Constructing such a benchmark requires expert harmonization of labels and multimodal data, together with substantial computational resources. Future work will develop standardized cross-regional benchmarks with city-level and spatial holdouts to evaluate transfer across different cultural-geographic environments and urban forms. We additionally evaluated LandAI-L1 using remote-sensing imagery alone in Urumqi, Lhasa, and Kashgar; details and results are provided in Supplementary Note 11 and Supplementary Fig. 3. Third, the inference latency of generating long CoT sequences restricts real-time edge deployment. Street-view imagery also represents a promising extension because it provides human-interpretable ground-level evidence, including building façades, streetscape functions, and visible commercial activities, that cannot be fully captured from overhead imagery. Future work should therefore examine lighter-weight grounded reasoning models, the integration of street-view and remote-sensing observations, and extensions to related urban and environmental tasks such as change detection, parcel updating, and spatiotemporal monitoring (Che et al., 2025; Fang et al., 2022; Wu et al., 2023).

Geometrically guided reasoning is a useful design principle for building more interpretable multimodal systems for urban land-use analysis. Rather than treating explainability as a separate post-hoc layer, the LandAI framework incorporates evidence localization directly into the reasoning process. This integration appears to improve data efficiency, robustness to cross-modal conflict, and the practical usefulness of model outputs for analysis and verification.

## 5. Conclusion

This study systematically explored the application of chain-of-thought reasoning in GeoAI and introduced the LandAI family of models for explainable parcel-level land-use classification. By integrating structured geographic knowledge, multimodal learning, and grounded reasoning, the proposed framework moves GeoAI beyond conventional pattern matching toward a more explicit reasoning-based paradigm.

First, we constructed Geo-Base-Thinking-14 K, a large-scale Chinese chain-of-thought dataset for geographic reasoning that covers geographic knowledge, spatial relations, and land-use rules. Fine-tuning on this corpus produced LandAI-Base, which substantially strengthened the model's domain reasoning capacity and provided the foundation for subsequent multimodal adaptation. Building on this base, we introduced GRPO-L1 and the three-stage “visual indexing, coordinate-guided localization, and language reasoning” paradigm, which requires the model to generate not only a class label but also an inspectable reasoning trace linked to explicit visual evidence. This design connects abstract inference with localized spatial cues and makes the decision process more transparent.

Through continued reinforcement learning on remote-sensing parcels with chain-of-thought annotations, we further developed LandAI-L1 as a geometrically guided and explainable land-use reasoning model. On a test set of more than 13,000 parcels, LandAI-L1 achieved 86.41% overall accuracy, outperforming comparable models. More importantly, the explicit reasoning traces produced by LandAI-L1 allow model decisions to be tracked and examined, offering a practical basis for interpretable parcel-level land-use analysis.

Overall, this study presents a reusable methodological framework for geometrically guided and explainable GeoAI. The results show how multimodal large language models can be guided to combine “seeing” with “reasoning” in a way that supports both predictive performance and interpretability. In this sense, LandAI provides a promising foundation for future GeoAI systems that are not only accurate, but also transparent, verifiable, and useful for real-world urban land-use analysis and related applications.

## CRediT authorship contribution statement

**Geyuan Zhu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation. **Yao Yao:** Writing – review & editing, Resources, Funding acquisition, Conceptualization, Project administration, Writing – original draft, Supervision, Methodology. **Chenglong Yu:** Validation, Supervision, Methodology, Formal analysis, Conceptualization, Writing – original draft, Writing – review & editing. **Xiang Zhang:** Methodology, Investigation, Conceptualization, Writing – review & editing, Writing – original draft. **Liangyang Dai:** Supervision, Methodology, Data curation, Conceptualization, Writing – original draft. **Zhihui Hu:** Project administration, Data curation, Conceptualization, Writing – original draft. **Yueheng Ma:** Investigation, Data curation, Conceptualization, Writing – review & editing, Writing – original draft. **Huanjun Hu:** Supervision, Resources, Funding acquisition, Conceptualization, Writing – review & editing, Writing – original draft. **Qingfeng Guan:** Resources, Funding acquisition, Conceptualization, Writing – review & editing, Writing – original draft.

## Fundings

This work was supported by the National Natural Science Foundation of China [42471491, 42171466], the CCF-DiDi GAIA Collaborative Research Funds [GAIA202620].

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

The Supplementary Information provides additional materials supporting reproducibility and extended validation, including complete prompts and reasoning examples, detailed training configurations, human verification and bounding-box quality audits, modality and commercial-model stability analyses, visual–textual conflict evaluation, and controlled ablations of the reasoning data, reinforcement-learning stage, and grounding reward.

## Data availability

All data, code, and model weights are publicly available via Figshare. The LandAI-L1, LandAI-L1-Zero and LandAI-Base model weights, the GRPO-L1 training code, and the demo dataset are publicly available on Figshare at https://doi.org/10.6084/m9.figshare.31249966.

## References

Alayrac, J.-B., Donahue, J., Luc, P., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems, 35*, 23716–23736.

Bell, A., Aides, A., Helmy, A., et al. (2025). *Earth AI: Unlocking geospatial insights with foundation models and cross-modal reasoning*.

Bodnar, C., Bruinsma, W. P., Lucic, A., et al. (2025). A foundation model for the earth system. *Nature, 641*, 1180–1187.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., et al. (2022). Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data, 9*, Article 251.

Burgert, T., Stoll, O., Rota, P., et al. (2026). *ImageNet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression*.

Cao, S., Ruan, J., Dong, B., et al. (2024). Tackling instance-dependent label noise with class rebalance and geometric regularization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 211–221).

Chakraborti, T., Banerji, C. R. S., Marandon, A., et al. (2025). Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence, 7*, 522–530.

Che, L., et al. (2025). Unsupervised urban land use mapping with street view contrastive clustering and a geographical prior. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*.

Chen, B., Xu, Z., Kirmani, S., et al. (2024). SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14455–14465).

Chen, M., Qian, Z., Boers, N., et al. (2023). Iterative integration of deep learning in hybrid earth surface system modelling. *Nature Reviews Earth and Environment, 4*, 568–581.

Comanici, G., Bieber, E., Schaekermann, M., et al. (2025). *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*.

Fang, F., et al. (2022). Spatial context-aware method for urban land use classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing, 192*, 1–12.

Feng, J., Zeng, J., Long, Q., et al. (2025). *A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science*.

Gao, J., Li, Y., Cao, Z., et al. (2025). Interleaved-modal chain-of-thought. In *Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19520–19529).

Gavrikov, P., Lukasik, J., Jung, S., et al. (2025). Can we talk models into seeing the world differently?. In *The Thirteenth International Conference on Learning Representations*.

Ghosh, S., Evuru, C. K. R., Kumar, S., et al. (2025). Visual description grounding reduces hallucinations and boosts reasoning in LVLMs. In *The Thirteenth International Conference on Learning Representations. Presented at the The Thirteenth International Conference on Learning Representations*.

Gröger, F., Lionetti, S., Gottfrois, P., et al. (2024). Intrinsic self-supervision for data quality audits. *Advances in Neural Information Processing Systems, 37*, 92273–92316.

Guo, D., Yang, D., Zhang, H., et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature, 645*, 633–638.

Höhl, A., Obadic, I., Fernández-Torres, M.-Á., et al. (2024). Recent trends challenges and limitations of explainable AI in remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8199–8205).

Huang, Q., Dong, X., Zhang, P., et al. (2024). Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13418–13427).

Huang, W., Jia, B., Zhai, Z., et al. (2025). *Vision-R1: Incentivizing reasoning capability in multimodal large language models*.

Janowicz, K., Gao, S., McKenzie, G., et al. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science, 34*, 625–636.

Jin, W., Wang, N., Zhang, L., et al. (2025). A review of AI-driven automation technologies: Latest taxonomies, existing challenges, and future prospects. *Computers, Materials and Continua, 84*, 3961–4018.

Kuckreja, K., Danish, M. S., Naseer, M., et al. (2024). GeoChat: Grounded large vision-language model for remote sensing. In *Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 27831–27840).

Leng, S., Zhang, H., Chen, G., et al. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13872–13882).

Li, F., Jiang, Q., Zhang, H., et al. (2024). Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12861–12871).

Li, Z., Xu, J., Wang, S., et al. (2024). *StreetviewLLM: Extracting geographic information using a chain-of-thought multimodal large language model*.

Liu, P., & Biljecki, F. (2022). A review of spatially-explicit GeoAI applications in urban geography. *International Journal of Applied Earth Observation and Geoinformation, 112*, Article 102936.

Mai, G., Xie, Y., Jia, X., et al. (2025). Towards the next generation of geospatial artificial intelligence. *International Journal of Applied Earth Observation and Geoinformation, 136*, Article 104368.

Maslov, K. A., Persello, C., Schellenberger, T., et al. (2025). Globally scalable glacier mapping by deep learning matches expert delineation accuracy. *Nature Communications, 16*, 43.

Mehdiyev, N., Majlatow, M., & Fettke, P. (2025). Interpretable and explainable machine learning methods for predictive process monitoring: A systematic literature review. *Artificial Intelligence Review, 58*, Article 378.

Nieradzik, L., Stephani, H., & Keuper, J. (2025). Reliable evaluation of attribution maps in CNNs: A perturbation-based approach. *International Journal of Computer Vision, 133*, 2392–2409.

Obermeyer, Z., Powers, B., Vogeli, C., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*, 447–453.

OpenAI. (2024). Introducing OpenAI o1 [WWW Document]. https://openai.com/index/introducing-openai-o1-preview/ (accessed 7.20.25).

Pan, B., Li, Q., Tang, X., et al. (2025). NLPrompt: Noise-label prompt learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 19963–19973).

Peng, Z., Wang, W., Dong, L., et al. (2023). *Kosmos-2: Grounding multimodal large language models to the world*.

Reichstein, M., Camps-Valls, G., Stevens, B., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature, 566*, 195–204.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.
Schultz, M., Li, H., Wu, Z., et al. (2025). OSMlanduse a dataset of european union land use at 10 m resolution derived from OpenStreetMap and sentinel-2. *Scientific Data, 12*, 750.
Shang, C., Zhou, S., Zhang, H., et al. (2024). Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11030–11040).
Shukor, M., Bethune, L., Busbridge, D., et al. (2025). *Scaling laws for optimal data mixtures*.
Srivastava, D., Yan, G., & Weng, T.-W. (2024). VLG-CBM: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems, 37*, 79057–79094.
Suzgun, M., Gur, T., Bianchi, F., et al. (2025). Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, 1–11.
Wang, S., Han, W., Huang, X., et al. (2024). Trustworthy remote sensing interpretation: Concepts, technologies, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing, 209*, 150–172.
Wang, X., Cui, Y., Wang, J., et al. (2026). Multimodal learning with next-token prediction for large multimodal models. *Nature, 650*, 327–333.
Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.
Wu, K., Zhang, Y., Ru, L., et al. (2025). A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence, 7*, 1235–1249.
Wu, M., et al. (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.
Wu, T.-H., Lee, H., Ge, J., et al. (2025). *Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling*.
Wu, T.-H., Lian, L., Gonzalez, J. E., et al. (2024). Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6327–6336).
Wu, Z., Zhang, C., Gu, X., et al. (2023). Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature Communications, 14*, 3072.
Yang, A., Li, A., Yang, B., et al. (2025). *Qwen3 technical report*.
Yang, C., Li, H., Zhu, R., et al. (2026). Snow or rain? Hybrid AI deciphers surface precipitation phase from satellite observations. *Nature Communications, 17*, Article 2813.
Yao, Q., Yang, C.-H. H., Jiang, R., et al. (2025). *Towards neural scaling laws for time series foundation models*.
Yao, Y., Gao, R., Wu, H., 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.
Yerramilli, S., Pande, N., Grover, R., et al. (2025). *GeoChain: Multimodal chain-of-thought for geographic reasoning*.
Zhan, Y., Xiong, Z., & Yuan, Y. (2025). SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing, 221*, 64–77.
Zhang, W., et al. (2025). EarthGPT-X: A spatial MLLM for multilevel multisource remote sensing imagery understanding with visual prompting. *IEEE Transactions on Geoscience and Remote Sensing, 63*, 1–21.
Zhou, T., et al. (2025). MACityChat: Integrating remote sensing professional large model with general-purpose large model for multi-domain urban land use analysis. *Applied Soft Computing, 185*(Part B), Article 113929.
Zhu, G., Tang, M., Ma, Y., et al. (2025). LandGPT: A multimodal large language model for parcel-level land use classification with multi-source data. *International Journal of Geographical Information Science, 39*(12), 2862–2885.
Zhu, L., Ji, D., Chen, T., et al. (2025). Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 1624–1633).
Zhu, X. X., Xiong, Z., Wang, Y., et al. (2026). On the foundations of earth foundation models. *Communications Earth & Environment, 7*, Article 103.

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

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

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

**Mr. Liangyang Dai** is a Ph.D. candidate at Wuhan University. His research interests are data mining and its applications in building healthy cities.

**Mr. Zhihui Hu** is a Ph.D. candidate at East China Normal University. His research interests are geospatial big data mining and geospatial foundation modeling.

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

**Dr. 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.

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

**Dr. Yao Yao** is a Professor at China University of Geosciences (Wuhan). His research interests are geospatial big data mining, analysis, and computational urban science.