ComplexLU: An Evidence-Grounded Multi-Modal Dataset for Complex Land-Use Segmentation

Anonymous ACL submission

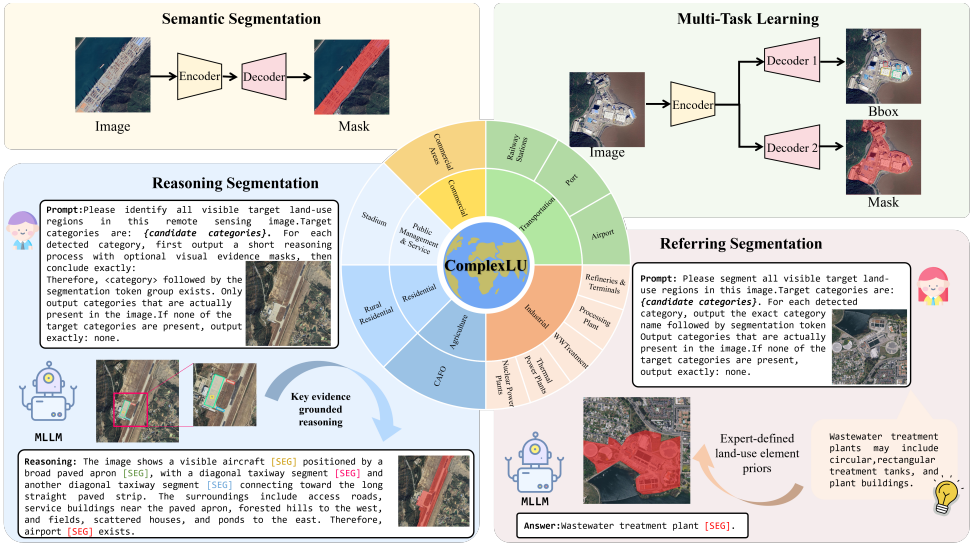

Figure 1: Overview of the ComplexLU dataset and the evaluated segmentation paradigms.

Abstract

Thematic mapping of complex land-use categories is crucial for regional sustainable development. However, intricate category compositions and strict domain expertise requirements lead to suboptimal generalization in large-scale applications. Existing methods either rely on implicit visual pattern matching or depend on predefined rules and exhaustive annotations. While vision-language models offer opportunities to integrate expert knowledge with visual features, the thematic mapping field lacks an evidence-grounded multi-modal dataset to validate explicit reasoning paradigms. To address this gap, we construct ComplexLU as a high-resolution dataset encompassing 12 functional land-use scenarios. By formalizing expert knowledge, this dataset unifies region-level semantic masks, element-level key evidence, and text-level reasoning descriptions. We benchmark representative setups including general semantic segmentation, multi-task segmentation, referring segmentation, and reasoning segmentation. Experiments reveal that pure visual approaches struggle to stably recognize complex geographic targets. Integrating language models with expert prior prompts effectively improves recognition stability. However, the limited performance gains of explicit reasoning over referring segmentation suggest possible architectural limitations of current vision-language models in capturing relationships between local parts and holistic targets. By explicitly bridging textual reasoning with geospatial perception, this work highlights the value of high-quality logic-rich data and provides a concrete empirical foundation for optimizing grounded visual-language reasoning. Our dataset has been anonymously released at https://anonymous.4open.science/r/ComplexLU.

1 Introduction

The accurate thematic mapping of land-use categories from remote sensing imagery is crucial for regional sustainable development and intelligent spatial decision-making (Dong et al., 2019; Vallet et al., 2024). While some land-use types exhibit obvious features and simple structural compositions, complex land-use categories, including large-scale industrial facilities, commercial complexes, and transportation hubs, are characterized by intricate internal compositions, vast spatial extents, and highly diverse morphological layouts. These complex scenarios typically necessitate dedicated thematic mapping methods. However, due to their structural complexity, the high demand for domain expertise, and the sparse geographic distribution of specialized types, existing large-scale mapping efforts frequently suffer from generalization bottle-

necks and unstable recognition accuracy. When applied to downstream regional spatial analysis, this lack of reliability fails to meet the precision requirements of real-world applications, creating an urgent need for more robust interpretation paradigms (Tong et al., 2020).

Existing automated approaches to land-use interpretation generally fall into three paradigms, all of which exhibit certain limitations. The first paradigm encompasses conventional scene classification and semantic segmentation networks, which rely primarily on purely data-driven visual pattern matching. Operating as black boxes, they capture statistical correlations rather than underlying functional logic, leading to performance degradation under domain shifts (Castelluccio et al., 2015; Cheng et al., 2017; Ghosh and Kaabouch, 2014). The second paradigm introduces structural logic through knowledge graphs or hierarchical classification. However, these methods often depend on fixed predefined rules and require comprehensive full-element recognition, making the annotation process highly costly and difficult to scale (Cheng et al., 2022; Mohammadimanesh et al., 2019). The third paradigm leverages multi-source auxiliary data, such as point-of-interest or social sensing data, to assist in functional inference. While effective in data-rich urban centers, such auxiliary data is typically scarce and difficult to acquire at national or global scales (Du et al., 2020). Consequently, methods heavily reliant on these inputs suffer from severe performance drops or even complete category failure in data-sparse regions.

The continuous evolution of large language models and vision-language models presents an opportunity to overcome these bottlenecks. By integrating expert knowledge with deep visual features, these models offer a pathway to reduce reliance on dense auxiliary data and shift toward robust visual reasoning. Recent works have begun exploring vision-language reasoning in geospatial contexts. For instance, Geo-R1 (Zhang et al., 2026) leverages reinforcement fine-tuning to improve few-shot geospatial referring expression understanding, while recent studies on urban socio-semantic segmentation incorporate vision-language reasoning to extract social attributes (Wang et al., 2026, 2024). However, these works primarily focus on localizing specific entities defined by particular social attributes, rather than conducting generalized thematic mapping of standardized land-use categories. Consequently, their data construction methodologies fundamentally differ from the requirements of complex functional scene interpretation, which relies on the physical-functional interplay of multiple spatial elements. Crucially, the remote sensing community still lacks a multi-modal dataset endowed with expert knowledge to effectively train and validate explicit reasoning frameworks tailored for complex functional scenes.

Theoretically, integrating expert knowledge and core visual evidence provides a pathway to unlock robust generalization in complex land-use mapping. However, the actual value and architectural feasibility of these evidence-grounded paradigms have yet to be comprehensively assessed. To address this, we systematically evaluate the role of explicit visual evidence and textual reasoning using 12 typical complex land-use categories. We construct ComplexLU as a high-resolution multi-modal dataset that strictly pairs remote sensing imagery with region-level semantic masks, element-level key evidence, and text-level reasoning descriptions. We test representative setups across general semantic segmentation, multi-task segmentation, referring expression segmentation, and reasoning segmentation (Figure 1). The evaluation results explicitly demonstrate that pure visual approaches struggle with complex geographic morphology. While incorporating expert prior prompts and element-level cues improves recognition stability, the marginal performance gains of explicit reasoning expose inherent architectural bottlenecks. Specifically, current auto-regressive text generation processes interfere with dense spatial recovery, indicating a failure to genuinely establish relationships between local components and holistic targets within the language model. Our benchmark thus provides critical insights into the capabilities and current architectural limitations of vision-language models. Our main contributions are summarized as follows:

- We introduce ComplexLU as an evidence-grounded multi-modal dataset for complex land-use segmentation that unifies high-resolution imagery, region-level semantic masks, element-level key evidence, and text-level reasoning descriptions.
- We establish comprehensive benchmarks across representative segmentation paradigms to systematically evaluate the impact of visual evidence, expert knowledge, and textual reasoning on complex spatial interpretation.

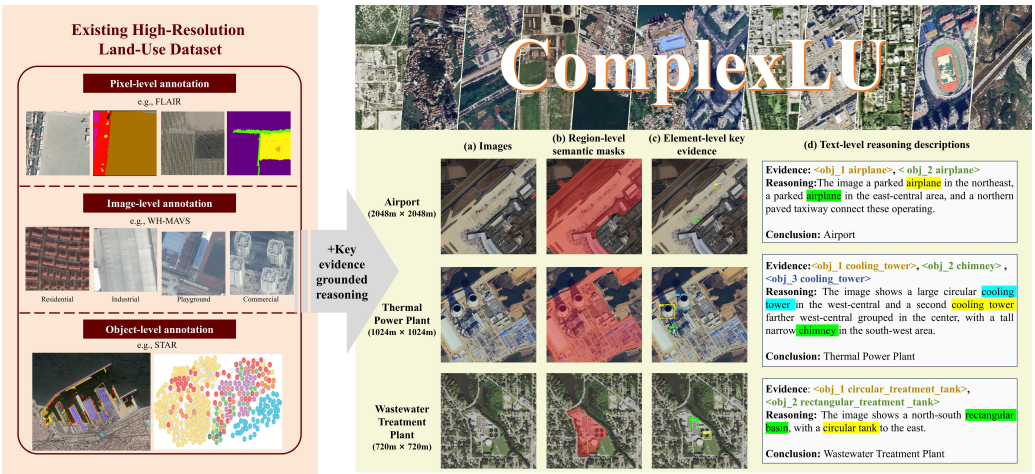

Figure 2: Comparison between prior annotation paradigms and ComplexLU. Existing datasets mainly provide pixel, image, or object-level supervision, whereas ComplexLU additionally introduces element-level key evidence and grounded text-level reasoning.

- Extensive evaluations reveal that while language-guided prompts improve recognition stability, existing vision-language models lack specific architectural designs to capture part-to-whole relationships. This exposes inherent limitations in balancing logical reasoning with fine-grained spatial detail recovery and provides concrete directions for future model optimization.

2 Related Work

2.1 High-Resolution Land-Use Datasets

High-resolution remote sensing datasets have advanced semantic segmentation, land-use change analysis, and functional understanding. For example, ISPRS Vaihingen/Potsdam (Cramer, 2010; Rottensteiner et al., 2012) and LoveDA (Wang et al., 2021) provide pixel-level annotations for buildings, roads, vegetation, and water bodies, enabling fine-grained land-cover recognition. As shown in Table 1, datasets such as WH-MAVS (Yuan et al., 2022), GID (Tong et al., 2020), and FLAIR (Garioud et al., 2023) extend coverage to large-scale mapping, enabling the extraction of both land-cover and general land-use categories. METER-ML (Zhu et al., 2022) targets methane-related infrastructure with image-level labels, while FUSU (Yuan et al., 2024a) further considers functional semantics and temporal dynamics. STAR (Li et al., 2025a) provides oriented bounding boxes and scene graphs that capture object relationships. However, constructing such graphs requires full-instance object detection, which is challenging to scale to large functional areas.

Despite these advancements, existing land-use datasets predominantly provide only pixel-level semantic segmentation masks. They treat functional zone extraction as a black-box pattern matching problem, lacking the visual markers of key functional evidence and the explicit text-level reasoning processes necessary for interpretable understanding. To address this limitation, we construct ComplexLU, an evidence-grounded multi-modal dataset covering 12 complex land-use scenarios. Each sample unifies region-level semantic masks, element-level key evidence bounding boxes, and text-level reasoning descriptions, specifically designed to evaluate evidence-grounded interpretation methods for complex thematic mapping.

2.2 Visual Understanding and Reasoning Methods

Semantic Segmentation assigns a class label for each pixel in an image and serves as a core task for land-cover mapping and region extraction. Early encoder-decoder CNNs such as U-Net (Ronneberger et al., 2015), DeepLabV3+ (Chen et al., 2018), and PSPNet (Zhao et al., 2017) recover spatial detail via skip connections, atrous convolutions, and pyramid pooling strategies, while Transformer-based frameworks including SegFormer (Xie et al., 2021) and Mask2Former (Cheng et al., 2022) leverage self-attention to capture long-range dependencies. Mamba-based architectures such as RS3Mamba (Ma et al., 2024) further explore efficient sequential modeling for high-resolution remote sensing imagery. In this work, we deploy these semantic segmentation models as our primary baselines. By evaluating their performance under

Dataset	Year	Image Size	Count	Resolution	Source	Classes	Complex	Supported Tasks
WH-MAVS (Yuan et al., 2022)	2022	200×200	47,137	1.2 m	GEI	14	×	Scene Classification
METER-ML (Zhu et al., 2022)	2022	720×720	86,599	1 m / 10 m	Aerial + Sentinel-1/2	7	×	Scene Classification
ISPRS Vaihingen (Cramer, 2010)	2013	1887–3816	33	0.09 m	Aerial	6	×	General Semantic Segmentation
ISPRS Potsdam (Rottensteiner et al., 2012)	2013	6000×6000	38	0.05 m	Aerial	6	×	General Semantic Segmentation
GID (Tong et al., 2020)	2020	6800–7200	150	1 m / 4 m	GF-2	5 / 15	×	General Semantic Segmentation
LoveDA (Wang et al., 2021)	2021	1024×1024	5,987	0.3 m	GEI	7	×	General Semantic Segmentation Domain Adaptation
FLAIR (Garioud et al., 2023)	2023	512 / 40	77,762	0.2 m / 10 m	Aerial + Sentinel-2	19	×	General Semantic Segmentation
FUSU (Yuan et al., 2024a)	2024	512 / 128	62,752	0.2–0.5m	GEI + Sentinel-1/2	17	✓	General Semantic Segmentation Change Detection
STAR (Li et al., 2025a)	2025	512–31096	1,273	0.15–1.0 m	GEI	11	✓	Scene Graph Generation
ComplexLU (ours)	2026	512–6144	4,489	0.3–1.0 m	GEI + Aerial	13	✓	Scene Classification General Semantic Segmentation Multi-Task Segmentation Referring Segmentation Reasoning Segmentation

Table 1: Comparison of high-resolution land-use datasets in image size, sample count, resolution, data sources, class count, scene complexity, and supported tasks.

pure mask supervision, we assess the capability and limits of directly learning implicit visual features, providing a foundational reference to measure the performance gains brought by explicit evidence and reasoning.

Multi-Task Learning jointly models region segmentation and key element detection to improve scene understanding. YOLOP (Wu et al., 2022) and YOLOPv3 (Zhan et al., 2024) demonstrate that object detection, drivable area segmentation, and lane detection can benefit from shared backbone features, while relational part-aware learning in remote sensing (Yuan et al., 2024b) shows that recognizing complex objects requires attention to constituent parts and their spatial relationships. However, existing multi-task methods predominantly focus on traffic scenes, single composite object detection, or conventional land-cover tasks, and have not explored how key element identification can guide the segmentation of complex land-use functional regions. Our dataset, with its paired region-level masks and element-level annotations, naturally supports this paradigm, allowing models to learn the spatial correspondence between core elements and functional region boundaries.

Referring Segmentation requires models to associate natural language expressions with pixel-level region prediction. In general vision, InstructSeg (Wei et al., 2025) and PSALM (Zhang et al., 2024b) unify different instruction-based segmentation paradigms, while PixelLM (Xu et al., 2024) emphasizes pixel-level reasoning with large multimodal models. GSVA (Xia et al., 2024) further generalizes referring segmentation with multi-modal large language models by supporting both multiple target objects and no-target cases, making it closely related to segmentation settings that require category-conditioned localization and rejection. In remote sensing domains, SegEarth-OV3 (Li et al., 2026) explores open-vocabulary semantic segmentation, enabling models to generate region masks from class names or simple text prompts. In this work, we utilize this setup to evaluate the upper-bound region-extraction capability of models when the target category is explicitly provided as a prompt. This allows us to assess whether large language models can accurately identify functional zone boundaries by relying on prior category knowledge, while also noting that practical applications would necessitate coupling with a scene classification module.

Reasoning Segmentation naturally extends standard referring segmentation by incorporating explicit reasoning or visual evidence into the prediction process. Models such as OMG-LLaVA (Zhang et al., 2024a), LIRA (Li et al., 2025b), Geo-R1 (Zhang et al., 2026), TerraScope (Shu et al., 2026), and SegEarth-R2 (Xin et al., 2026) explicitly perform reasoning to infer relevant evidence before executing segmentation actions. Recent multimodal benchmarks also connect reasoning with explicit grounding; for example, Seeing Culture

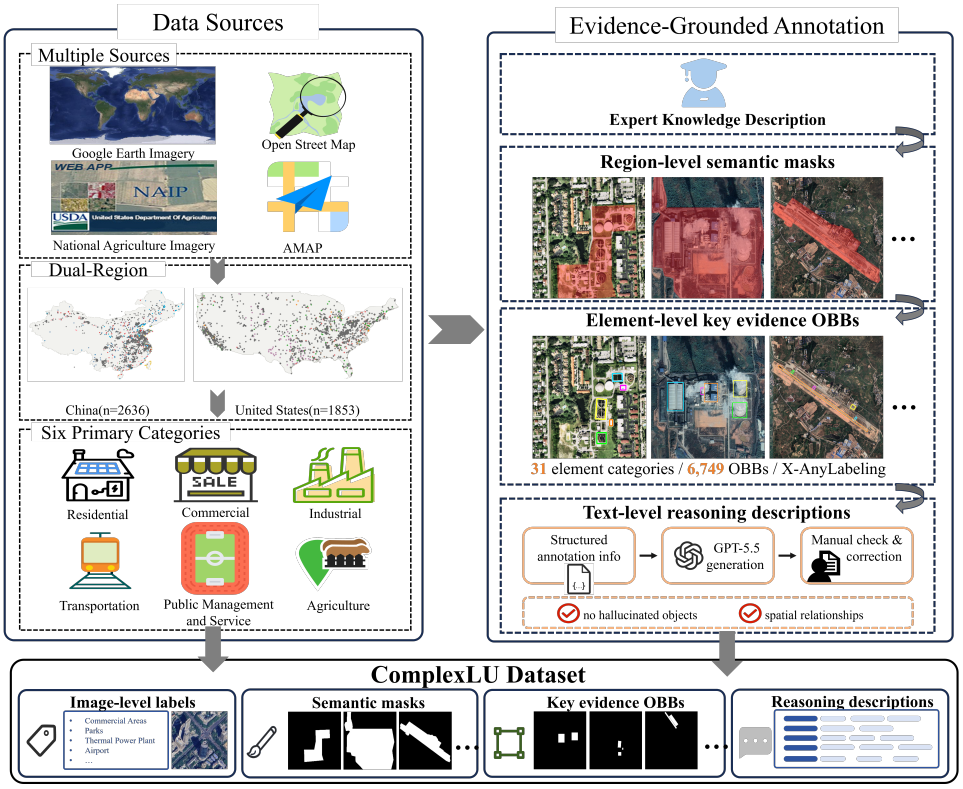

Figure 3: ComplexLU dataset construction pipeline: data sources, evidence-grounded annotation, and multi-modal components.

(Satar et al., 2025) requires models to answer visual reasoning questions and then segment the relevant visual artifact as supporting evidence. While this class of methods is inherently suited for utilizing expert knowledge and visual evidence, they are typically designed for open-ended, complex reasoning tasks. Therefore, we benchmark reasoning segmentation on ComplexLU to systematically evaluate the effectiveness of simultaneously leveraging textual reasoning and explicit visual evidence for complex land-use interpretation.

3 Dataset

We introduce ComplexLU, an evidence-grounded multi-modal dataset for complex land-use understanding in high-resolution remote-sensing imagery. ComplexLU contains 4,489 images from China and the United States, including 1,099 positive samples and 3,390 negative samples, with spatial resolutions of approximately 0.3–1 m. The positive samples cover 12 complex land-use scenarios grouped into six functional categories, as summarized in Appendix 4. Each positive sample is annotated with region-level semantic masks, element-level key evidence, and text-level reasoning descriptions. By explicitly linking visible evidence to functional scene interpretation, ComplexLU supports region localization, key evidence modeling, and explainable complex land-use reasoning, while negative samples enable evaluation under realistic non-target scenes.

3.1 Data Collection

We construct ComplexLU by integrating public high-resolution remote sensing imagery with open geographic data. Candidate regions are sourced from OpenStreetMap, Google Earth, and NAIP imagery from METER-ML. To guarantee dataset quality, all candidates are manually labeled and verified through visual interpretation and cross-checked with OSM data for boundary consistency. This strict screening process effectively filters out any ambiguous or low-quality samples.

We further collect negative samples from hard background scenes from Amap and Google Earth. To account for the heterogeneous spatial extent of functional land-use regions, we construct image crops with category-dependent coverage rather than enforcing a single fixed footprint. The crop extent is selected to retain the complete target region and its surrounding contextual cues while preserving sufficient spatial detail for element-level annotation. The resulting category distribution, number of annotated elements, and element categories are

summarized in Appendix Table 5.

3.2 Evidence-Grounded Annotation

The central design of ComplexLU is to explicitly connect functional land-use regions with their supporting visual evidence and reasoning descriptions. For each positive sample, we provide three annotation types: region-level semantic masks, element-level key evidence annotations, and text-level reasoning descriptions. These annotations form an evidence-grounded supervision structure that supports region localization, key evidence modeling, and explainable semantic reasoning.

Region-level semantic masks. Region-level semantic masks delineate the spatial extent and boundary of the target functional land-use scene. For samples with available OSM polygon priors, we use the polygon as the initial mask and refine it according to the visible remote sensing image content. Samples with unreliable boundaries, missing scene bodies, severe mixed land use, or ambiguous dominant functions are discarded during manual inspection. Negative samples are also checked to ensure that they do not contain any target complex land-use category. Through this process, ComplexLU provides pixel-level supervision for extracting complete functional regions rather than isolated land-cover objects.

Element-level key evidence annotations. Element-level annotations identify the key visual evidence that supports complex land-use interpretation. For each of the 12 target land-use categories, we define category-specific core elements, as summarized in Table 5 (see supplementary material). In total, 31 element categories are defined, yielding 6,749 oriented bounding box annotations across all categories. For example, Airports are characterized by runways, aircraft, taxiways, and terminal aprons; Wastewater Treatment Plants by circular/rectangular treatment tanks and plant buildings. Annotators use X-AnyLabeling (Wang, 2023) to annotate these key elements with rotated bounding boxes, which better represent arbitrarily oriented objects in remote sensing imagery.

Different from conventional object detection datasets that pursue exhaustive instance-level annotation, our element-level annotations focus on core elements directly related to functional interpretation. Therefore, they are not intended as complete object detection supervision. Instead, they serve as key evidence for element-guided region extraction, multi-task segmentation, and interpretable reasoning.

Text-level reasoning descriptions. Text-level reasoning descriptions explain how the observed key elements, spatial relationships, and scene scale jointly support the corresponding land-use category. For each positive sample, we generate one reasoning description from structured annotation information rather than directly from the binary mask. The structured input includes the image category, image size, spatial resolution, element categories, and element locations. Element information describes the type and spatial distribution of supporting visual evidence, while image size and spatial resolution help estimate the real- world ground coverage of the scene.

We use GPT-5.5 (OpenAI, 2026) as the expert large language model to generate initial reasoning descriptions from a structured prompt template, as summarized in Appendix B. The prompt requires the model to select key evidence elements supporting the specified category, describe their spatial relationships with the surrounding scene, and explain why these cues indicate the corresponding land-use function. To reduce hallucination, the prompt explicitly prohibits the model from introducing objects, facilities, or environmental information absent from the structured input.

The generated descriptions, illustrated by the example responses in Appendix Figure 5, are manually checked and corrected. Descriptions with hallucinated objects, incorrect spatial relationships, insufficient category explanations, or inconsistencies with the annotations are revised or regenerated.

4 Experiment

4.1 Experimental Setup and Metrics

The dataset is partitioned into spatially disjoint training and test sets, comprising 3,596 and 893 images respectively. To preserve class distributions while avoiding spatial leakage, we adopt a stratified split that groups geographically proximate images into the same set and maintains consistent category proportions across both subsets. All quantitative evaluations are conducted on this test set.

We categorize the evaluated methods into five comparison settings, each utilizing a different level of ComplexLU supervision. General semantic segmentation methods use only region-level semantic masks. Multi-task segmentation architectures jointly learn semantic masks and element-

level bounding boxes. Referring segmentation uses candidate-category prompts and is evaluated under two prompt variants: *category-only*, where the prompt provides only candidate land-use categories, and *expert-prior*, where the prompt additionally lists possible evidence element types for each category. Reason-seg first generates an explanatory response and then predicts the final land-use mask. Seg-reason uses the same reasoning content but places the final category and mask before the explanatory text, isolating the effect of output order on mask binding. For reasoning-based settings, both variants supervise the reasoning text and final land-use mask. The difference is that *target-only* supervision applies mask loss only to the final land-use region, whereas *evidence-grounded* supervision additionally inserts evidence-mask tokens into the reasoning text and supervises the corresponding intermediate evidence masks; both settings are evaluated on the final land-use mask.

All models use RGB inputs resized to 1024×1024 and are fine-tuned from publicly available pretrained weights with an effective batch size of 8. General semantic segmentation, multi-task segmentation, and referring-segmentation models are trained for 10k steps, while the main reasoning-segmentation comparison reports 20k-step checkpoints. We keep other hyperparameters at their model-default settings. Detailed backbone and fine-tuning settings are provided in Appendix C.

We report mean Intersection over Union (mIoU), mean Accuracy (mAcc), and Overall Accuracy (OA) for quantitative assessment. Mean Intersection over Union measures the average spatial overlap between predictions and ground-truth regions across all categories. Mean Accuracy evaluates the average proportion of correctly classified pixels per class. Overall Accuracy computes the global fraction of correctly assigned pixels across the entire test set.

We evaluate representative models across the five established comparison settings. All models are fine-tuned on the ComplexLU training set. For general semantic segmentation, we benchmark Mask2Former (Cheng et al., 2022), VMamba (Liu et al., 2024), and DeepLabV3 (Chen et al., 2018). For multi-task learning, we consider YOLOP (Wu et al., 2022) and YOLOPv3 (Zhan et al., 2024), which simultaneously predict semantic masks and element-level bounding boxes. For referring segmentation, we benchmark GSVA (Xia et al., 2024), PSALM (Zhang et al., 2024b), and PixelLM (Xu et al., 2024) under both category-only and expert-prior prompts. Unlike the standard setting that assumes a single known target category, we provide a list of candidate categories and require the model to output masks only for categories that are present in the image. For reasoning-based segmentation, we benchmark PixelLM and SegEarth-R2 (Xin et al., 2026) under both Reason-seg and Seg-reason output orders. The corresponding prompt–answer templates are provided in Appendix F.

Paradigm	Model	Setting	mAcc	mIoU	OA
General	Mask2Former	–	52.19	42.27	79.52
	VMamba	–	48.62	38.05	70.27
	DeepLabV3	–	51.74	39.92	75.36
Multi-task	YOLOP	–	60.85	50.17	85.72
	YOLOPv3	–	65.52	53.49	88.29
Referring	GSVA	Category-only	73.96	59.09	96.59
	PSALM	Category-only	79.08	60.92	96.47
	PixelLM	Category-only	78.52	65.51	97.11
	GSVA	Expert-prior	73.26	59.59	96.62
	PSALM	Expert-prior	79.74	62.21	96.50
	PixelLM	Expert-prior	**82.11**	65.78	97.08
Reason-seg	PixelLM	Target-only	75.93	63.80	97.02
	SegEarth-R2	Target-only	75.34	64.83	97.13
	PixelLM	Evidence-grounded	75.01	61.66	96.96
	SegEarth-R2	Evidence-grounded	77.78	63.55	97.10
Seg-reason	PixelLM	Target-only	78.72	**66.02**	97.16
	SegEarth-R2	Target-only	78.95	65.92	**97.31**
	PixelLM	Evidence-grounded	77.29	64.26	97.12
	SegEarth-R2	Evidence-grounded	77.65	65.22	97.19

Table 2: Experimental results on ComplexLU. Red bold marks the best result in each metric column, and blue marks the second-best result.

4.2 Quantitative Results

Table 2 summarizes the unified evaluation results. All methods are fine tuned on the same training set and evaluated on the same test set. We can derive three main findings. First, pure visual approaches exhibit poor performance and low precision in scenarios characterized by complex morphology and limited training samples. Utilizing multi-task learning to introduce structural cues notably improves performance. However these results remain inferior to language-guided settings. Integrating large language models effectively mitigates these limitations and improves recognition stability.

Second, the referring segmentation paradigm demonstrates that providing sample-agnostic expert knowledge yields notable improvements without requiring instance-specific textual supervision. For architectures optimized for fine-grained instructions such as PixelLM, providing expert prior prompts significantly improves metrics over category only prompts. This highlights the critical

Train set	Model	Setting	Steps	mAcc	mIoU	OA
Full	Mask2Former	Mask only	10k	52.19	42.27	79.52
2-shot	PixelLM	Expert-prior	10k	43.15	34.67	94.38
	SegEarth-R2	Evidence-grounded	10k	50.40	37.42	94.00
	SegEarth-R2	Evidence-grounded	20k	52.22	38.20	93.89
∼8-shot	PixelLM	Expert-prior	10k	66.13	55.98	96.26
	SegEarth-R2	Evidence-grounded	10k	65.84	50.83	95.54
	SegEarth-R2	Evidence-grounded	20k	66.57	51.88	95.71

Table 3: Few shot evaluation on the same test set. Mask2Former serves as the full data baseline utilizing mask only supervision. PixelLM and SegEarth-R2 utilize limited annotated samples enriched with visual evidence.

role of prompt engineering in activating the spatial awareness of language models for specific candidate categories.

Finally, while explicit reasoning connects visual evidence with functional interpretation, the performance gains over referring segmentation remain limited. Current vision language models lack specific architectural designs to capture the relationship between complex holistic targets and their local constituents. Even though explicit reasoning links visual evidence with functional semantics, this part-to-whole relationship is not genuinely established within the language model. Our work indicates that current vision language models have substantial room for optimization in method design for thematic mapping tasks. This provides concrete optimization directions for advancing multi-modal geographic applications.

4.3 Discussion

We further study sample efficiency using the two strongest vision-language models from the main benchmark. These are PixelLM under the expert-prior referring-segmentation setting and SegEarth-R2 under the evidence-grounded Reason-seg setting. We construct a 2-shot subset and an approximately 8-shot subset while keeping the same 893-image test set. The 2-shot subset contains only two positive samples per target category plus negative samples, yielding 100 training images in total. The approximately 8-shot subset contains 366 training images. Detailed sampling statistics are provided in Appendix Table 6. Zero-shot inference performs poorly across the evaluated vision-language models, often failing to produce valid target masks for complex functional land-use categories, so we focus the quantitative analysis on few-shot adaptation. In this low-data study, all results are obtained from independent few-shot fine-tuning runs initialized from the corresponding pretrained models. They are used to compare training budgets under the same few-shot data regime rather than as intermediate checkpoints from the full-data fine-tuning runs.

Table 3 demonstrates the sample efficiency of vision-language models when fine-tuned on high-quality evidence-grounded data. Adapting these architectures with a minimal number of positive samples enables them to match or exceed the performance of pure-vision baselines trained on the entire dataset. These findings explicitly confirm that complex land-use interpretation benefits more from compact evidence-rich supervision than from large volumes of mask-only annotations.

5 Conclusion

We introduce ComplexLU as an evidence-grounded multi-modal dataset to benchmark complex land use segmentation. Evaluations reveal that pure visual approaches struggle to achieve stable and effective recognition of complex geographic targets. Incorporating visual evidence provides essential structural cues, and further integrating language models with expert prior prompts improves recognition stability. However, the limited performance gains of existing explicit reasoning over referring segmentation suggests possible architectural limitation of current vision language models in capturing relationships between local parts and holistic targets. Furthermore, ComplexLU demonstrates the value of high-quality data enriched with logical reasoning, providing a solid empirical foundation to guide the targeted optimization of vision language models in future thematic mapping applications.

Limitations

While our dataset enhances spatial interpretation, its uniform textual reasoning and exclusive focus on core evidence may cause models to underrepresent secondary elements. Future research must develop dynamic reasoning mechanisms that adaptively integrate diverse spatial cues. Additionally, current vision-language models face poor zero-shot performance and prohibitive computational overhead, limiting large-scale deployment. However, because fine-tuning on minimal high-quality samples ensures mapping stability, future efforts should prioritize activating intrinsic expert knowledge through compact datasets and exploring knowledge distillation for efficient large-scale interpretation.

References

Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. 2015. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

Michael Cramer. 2010. The dgpf-test on digital airborne camera evaluation overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*, pages 73–82.

Jinwei Dong, Graciela Metternicht, Patrick Hostert, Rasmus Fensholt, and Rinku Roy Chowdhury. 2019. Remote sensing and geospatial technologies in support of a normative land system science: Status and prospects. *Current Opinion in Environmental Sustainability*, 38:44–52.

Shouji Du, Shihong Du, Bo Liu, Xiuyuan Zhang, and Zhijia Zheng. 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. *GIScience & Remote Sensing*, 57(3):411–430.

Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, and Boris Wattrelos. 2023. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36:16456–16482.

Debabrata Ghosh and NAIMA Kaabouch. 2014. A survey on remote sensing scene classification algorithms. *WSEAS Transactions on Signal Processing*, 10(1):504–519.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. 2026. Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. *Preprint*, arXiv:2512.08730.

Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, Bo Dang, Yongjun Zhang, Yi Yu, and Junchi Yan. 2025a. Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1832–1849.

Zhang Li, Biao Yang, Qiang Liu, Shuo Zhang, Zhiyin Ma, Liang Yin, Linger Deng, Yabo Sun, Yuliang Liu, and Xiang Bai. 2025b. Lira: Inferring segmentation in large multi-modal models with local interleaved region assistance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24056–24067.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063.

Xianping Ma, Xiaokang Zhang, and Man-On Pun. 2024. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5.

Fariba Mohammadimanesh, Bahram Salehi, Masoud Mahdianpari, Eric Gill, and Matthieu Molinier. 2019. A new fully convolutional neural network for semantic segmentation of polarimetric sar imagery in complex land cover ecosystem. *ISPRS journal of photogrammetry and remote sensing*, 151:223–236.

OpenAI. 2026. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-06.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. 2012. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3:293–298.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243. Association for Computational Linguistics.

Yan Shu, Bin Ren, Zhitong Xiong, Xiao Xiang Zhu, Begüm Demir, Nicu Sebe, and Paolo Rota. 2026. Terrascope: Pixel-grounded visual reasoning for earth observation. *arXiv preprint arXiv:2603.19039*.

Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322.

Améline Vallet, Stéphane Dupuy, Matthieu Verlynde, and Raffaele Gaetano. 2024. Generating high-resolution land use and land cover maps for the greater mariño watershed in 2019 with machine learning. *Scientific Data*, 11(1):915.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. 2021. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc.

Libo Wang, Sijun Dong, Ying Chen, Xiaoliang Meng, Shenghui Fang, and Songlin Fei. 2024. Metasegnet: Metadata-collaborative vision-language representation learning for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11.

Wei Wang. 2023. Advanced auto labeling solution with added features. https://github.com/CVHub520/X-AnyLabeling.

Yu Wang, Yi Wang, Rui Dai, Yujie Wang, Kaikui Liu, Xiangxiang Chu, and Yansheng Li. 2026. Urban socio-semantic segmentation with vision-language reasoning. In *International Conference on Learning Representations*.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. 2025. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20193–20203.

Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and Wen-Yu Liu. 2022. Yolop: You only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6):550–562.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Yuchen Xiao, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. 2026. Segearth-r2: Towards comprehensive language-guided segmentation for remote sensing images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026*.

Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. 2024. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039.

Jingwen Yuan, Lixiang Ru, Shugen Wang, and Chen Wu. 2022. Wh-mavs: A novel dataset and deep learning benchmark for multiple land use and land cover applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1575–1590.

Shuai Yuan, Guancong Lin, Lixian Zhang, Runmin Dong, Jinxiao Zhang, Shuang Chen, Juepeng Zheng, Jie Wang, and Haohuan Fu. 2024a. Fusu: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. *Advances in Neural Information Processing Systems*, 37:132417–132439.

Shuai Yuan, Lixian Zhang, Runmin Dong, Jie Xiong, Juepeng Zheng, Haohuan Fu, and Peng Gong. 2024b. Relational part-aware learning for complex composite object detection in high-resolution remote sensing images. *IEEE Transactions on Cybernetics*, 54(10):6118–6131.

Jiao Zhan, Jingnan Liu, Yejun Wu, and Chi Guo. 2024. Multi-task visual perception for object detection and semantic segmentation in intelligent driving. *Remote Sensing*, 16(10):1774.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. 2024a. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. 2024b. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer.

Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Yuxiang Cai, Zhonggen Su, Yongheng Shang, Zhaojun Liu, Jianwei Yin, and Xiang Li. 2026. Georl: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:113–129.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Bryan Zhu, Nicholas Lui, Jeremy Irvin, Jimmy Le, Sahil Tadwalkar, Chenghao Wang, Zutao Ouyang, Frankie Y Liu, Andrew Y Ng, and Robert B Jackson. 2022. Meter-ml: A multi-sensor earth observation benchmark for automated methane source mapping. *arXiv preprint arXiv:2207.11166*.

A Additional Dataset Statistics

This section provides supplementary statistics for the ComplexLU category system, image sources, evidence elements, and few-shot splits. These tables and figures complement the dataset description and sample-efficiency analysis by making the category hierarchy, category-wise evidence statistics, low-data training subsets, image-resolution variation, and element-frequency distribution explicit.

Level I	Level II
01 Residential	01 Rural Residential
02 Commercial	02 Commercial Areas
03 Industrial	03 Processing Plant 04 Refineries & Terminals 05 Nuclear Power Plant 06 Thermal Power Plant 07 Wastewater Treatment Plant
04 Transportation	08 Airport 09 Port 10 Railway Station
05 Public Management & Service	11 Stadium
06 Agriculture	12 Concentrated Animal Feeding Operation

Table 4: Land-use category system for complex land-use scenes. Level II categories correspond to the current 12 positive target classes in ComplexLU.

Expert knowledge descriptions for Table 4.

- **Rural Residential** scenes contain clustered or scattered low-rise *house* elements with surrounding farmland or open space, indicating non-urban residential functions.
- **Commercial Areas** are characterized by dense *large building* elements and extensive *parking lot* areas, indicating concentrated retail, catering, accommodation, and entertainment activities.
- **Processing Plant** scenes contain industrial facilities with *pipe rack* and *horizontal tank* evidence, indicating natural gas processing and storage operations.
- **Refineries & Terminals** are identified by large *round tank* structures and dense *pipe rack* layouts near waterfronts or pipeline corridors, indicating petroleum refining and oil storage.
- **Nuclear Power Plant** scenes include distinctive *reactor* structures together with *cooling tower* and *plant building* evidence, indicating nuclear energy generation with strict safety zoning.
- **Thermal Power Plant** scenes contain hyperbolic *cooling tower* structures, tall *chimney* elements, and *coal yard* areas, indicating fossil-fuel combustion power generation.
- **Wastewater Treatment Plant** scenes are represented by *circular treatment tank* and *rectangular treatment tank* structures connected with *plant building* evidence, indicating municipal or industrial sewage treatment.
- **Airport** scenes contain long linear paved *runway* structures with *taxiway*, *apron*, and sometimes *airplane* evidence, indicating civil or military air transportation hubs.
- **Port** scenes are characterized by berthing *dock* structures with *crane*, *vessel*, and *yard* evidence along coastlines or major rivers, indicating maritime freight and logistics operations.
- **Railway Station** scenes contain station halls, *platform*, *railway track*, *station building*, and *train* evidence, indicating passenger and freight railway hubs.
- **Stadium** scenes contain large open-air sports venues with *athletic field*, *basketball court*, *tennis court*, and *stadium* evidence, indicating major public sports and event spaces.
- **Concentrated Animal Feeding Operation** scenes contain dense rows of long rectangular *shed* structures adjacent to large *manure lagoon* areas in rural settings, indicating intensive livestock breeding operations.

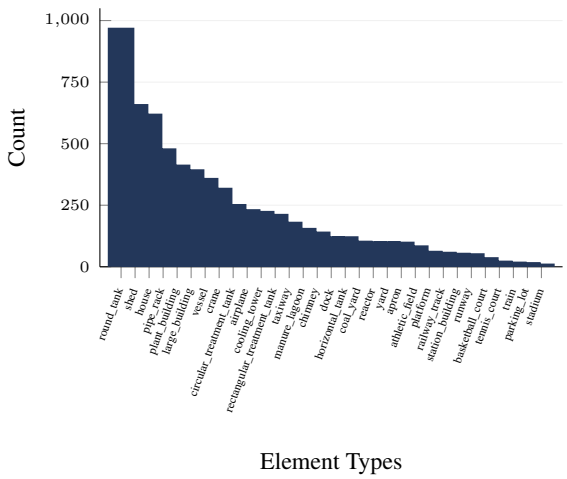

Figure 4: Frequency distribution of 31 element categories in ComplexLU (6,749 element annotations in total).

Class Name	Image Count	Element Count	Element Types
Airport	100	441	*runway*, *taxiway*, *apron*, *airplane*
Concentrated Animal Feeding Operation	95	662	*shed*, *manure lagoon*
Commercial Area	92	339	*large building*, *parking lot*
Nuclear Power Plant	63	225	*reactor*, *cooling tower*, *plant building*
Port	91	722	*dock*, *vessel*, *crane*, *yard*
Processing Plant	79	442	*pipe rack*, *horizontal tank*
Railway Station	82	156	*railway track*, *platform*, *station building*, *train*
Refinery and Terminal	99	856	*round tank*, *pipe rack*
Rural Residential	102	507	*house*
Stadium	95	125	*athletic field*, *basketball court*, *tennis court*, *stadium*
Thermal Power Plant	99	426	*cooling tower*, *chimney*, *coal yard*, *plant building*
Wastewater Treatment Plant	102	536	*circular treatment tank*, *rectangular treatment tank*, *plant building*
Negative	3,390	—	—
Total	**4,489**	**6,749**	**31 element categories**

Table 5: Category-wise statistics of ComplexLU, including image counts, element annotation counts, and key evidence types.

Class	Full train	2-shot	∼8-shot
Airport	80	2	8
Concentrated Animal Feeding Operation	76	2	8
Commercial Area	74	2	8
Nuclear Power Plant	51	2	6
Port	73	2	8
Processing Plant	64	2	7
Railway Station	66	2	7
Refinery and Terminal	80	2	8
Rural Residential	82	2	9
Stadium	76	2	8
Thermal Power Plant	80	2	8
Wastewater Treatment Plant	82	2	9
Positive total	884	24	94
Negative total	2,712	76	272
Training images	**3,596**	**100**	**366**

Table 6: Sampling statistics for the few-shot training subsets. The 2-shot subset uses 100 training images, while the approximately 8-shot subset uses 366 training images.

Class Name	Image Size (px)	Source
CAFOs	720×720	NAIP
Processing Plant	720×720	NAIP
RefineriesAndTerminals	720×720	NAIP
WWTreatment	720×720	NAIP
Airports	2048×2048; 6144×6144	GEI
CommercialAreas	512×512; 1024×1024	GEI
NuclearPowerPlants	1024×1024; 6144×6144	GEI/NAIP
Ports	2048×2048	GEI
RailwayStations	512×512; 1440×1440	GEI
RuralResidential	1024×1024	GEI
Stadiums	512×512; 1024×1024	GEI
ThermalPowerPlants	1024×1024; 2048×2048	GEI
Negative	720×720; 1024×1024; 1440×1440; 2048×2048; 4096×4096	GEI/NAIP

Table 7: Image size specifications of the multi-source dataset, including image sizes and sources for each category.

B GPT Prompt and Response Examples

This section reports the prompt template used to generate text-level reasoning annotations and gives a representative response format. The examples are intended to clarify how visible evidence, spatial layout, and category-level conclusions are connected during annotation.

Text-level reasoning generation.

- **Prompt:** You are a remote-sensing land-use interpretation expert. Given a remote-sensing image, the target land-use class *[target category]*, image-scale information, and visible evidence elements, write concise English reasoning for the land-use interpretation. Use only visible image evidence. Select the most important evidence elements from *[visible evidence list]*. Describe their spatial layout, surrounding context, and whether the observed scene scale is consistent with *[target category]*. Mention uncertainty when the visible evidence is incomplete or ambiguous. End with an explicit conclusion for *[target category]*.
- **Input:** *[remote-sensing image]*; target category: *[target category]*; image scale: *[image scale and approximate footprint]*; visible evidence: *[visible elements with approximate locations and spatial relations]*.
- **Answer:** The image shows *[key evidence elements]* arranged *[spatial layout]*. These elements appear within *[surrounding context]* and occupy a scene scale consistent with *[target category]*. *[uncertainty sentence if needed]* Therefore, the visible evidence supports the interpretation of this area as *[target category]*.

Example.

- **Prompt:** You are a remote-sensing land-use interpretation expert. Given the target class *airport*, image-scale information, and visible evidence including *runway* and *taxiway*, write concise English reasoning for the land-use interpretation.
- **Answer:** The image shows a long paved runway connected to taxiway-like paved strips, forming

a large transportation layout. The surrounding open land and access roads are consistent with the spatial scale of an airport facility. Therefore, the visible evidence supports the interpretation of this area as *airport*.

Additional response visualizations across the 12 ComplexLU categories are shown in Figure 5.

C Model Backbones and Fine-Tuning Settings

This section details the model backbones and optimization settings used in the experiments. The main paper reports unified metrics, while the descriptions below specify how each model family was adapted to ComplexLU.

All vision-language segmentation models are fine-tuned on the ComplexLU training split and evaluated on the held-out validation split. The split contains 3,596 training images and 893 validation images, with 12 foreground land-use categories and one background category. Positive samples provide target-region masks, while negative samples train the models to reject images that do not contain any target category. Across all settings, the input images are resized to 1024×1024 for segmentation supervision, and evaluation is conducted on the final land-use region mask.

General semantic segmentation. The general semantic-segmentation setting uses the same ComplexLU image split but only the region-level land-use masks. We formulate it as a closed-set semantic segmentation task with 13 classes, including 12 foreground land-use categories and background. Positive samples supervise the corresponding land-use category mask, while negative samples are treated as all-background images. Element-level bounding boxes and reasoning texts are not used in this setting, so the models can only learn from image appearance and region-mask supervision.

We benchmark Mask2Former, VMamba, and DeepLabV3 as representative Transformer, state-space, and convolutional segmentation backbones. Each model is initialized from its publicly released pre-trained weights and its final segmentation head is adapted to the 13 ComplexLU classes. All three models are fine-tuned for 10,000 steps with an effective batch size of 8 on the unified 1024×1024 inputs. Unless required by the class-number modification, the optimizer, learning-rate schedule, data augmentation, and other training hyperparameters follow the default recipe of each implementation. During inference, the predicted 13-class semantic map is directly evaluated against the final land-use region mask.

Multi-task segmentation. The multi-task setting uses the same images and region-level masks as general semantic segmentation, and additionally introduces the element-level annotations as detection supervision. Each positive sample provides a land-use region mask and a set of key-evidence bounding boxes drawn from the 31 element categories defined in ComplexLU. For architectures that require horizontal detection boxes, the oriented evidence annotations are converted to their minimum enclosing horizontal rectangles. Negative samples provide all-background segmentation masks and empty detection targets. This setting therefore tests whether learning to localize core functional elements can improve final land-use region segmentation under the same train/test split.

We evaluate YOLOP and YOLOPv3 using their original shared-backbone multi-task architectures and publicly available pre-trained weights. Their segmentation branches are adapted to predict the ComplexLU land-use region mask, and their detection branches are adapted to the 31 evidence-element categories. Both models are fine-tuned for 10,000 steps with an effective batch size of 8 on 1024×1024 RGB images. The segmentation and detection losses are optimized jointly following the default task-balancing strategy of each model. At evaluation time, only the final land-use segmentation output is used for the reported mIoU, mAcc, and OA, while the detection branch serves as auxiliary evidence supervision during training.

Referring segmentation. GSVA follows a generalized referring-segmentation formulation with a LLaVA-v1.1-7B backbone, CLIP ViT-L/14 visual tower, SAM ViT-H segmentation module, and released GSVA-7B GRES weights as initialization. PSALM uses its original vision-language segmentation architecture and is adapted through a candidate-category formulation, where the model identifies whether one of the candidate land-use categories is present and segments the target region. PixelLM uses a LLaVA-Lightning-7B language backbone with a CLIP ViT-L/14 visual encoder, and is adapted to ComplexLU by taking candidate land-use categories as input and predicting the corresponding land-use region mask.

We adapt GSVA, PSALM, and PixelLM under

Figure 5: **Example GPT-generated response visualization across 12 ComplexLU categories.** Each panel shows one representative remote-sensing image, selected evidence elements outlined with colored boxes, and a concise reasoning response for the same sample. Matching colors connect natural-language evidence mentions with their spatial regions in the image.

the same Category-only and Expert-prior prompt settings. All three models are trained for 10,000 steps with four GPUs, a per-device batch size of 2, gradient accumulation of 1, and an effective batch size of 8. GSVA uses a learning rate of 3×10^{-4}, PSALM uses 6×10^{-5}, and PixelLM uses 3×10^{-4}; other optimization settings follow the default recipes of the corresponding models unless otherwise specified.

Reasoning segmentation. The reasoning-segmentation setting extends the output space from a category-conditioned mask to a short reasoning response and the final land-use mask. PixelLM uses the same LLaVA-Lightning-7B language backbone and CLIP ViT-L/14 visual encoder as in the referring-segmentation setting. SegEarth-R2 uses a Mipha-3B language model, a SigLIP visual encoder, and a Mask2Former segmentation decoder. We adapt both models to ComplexLU with two output-order settings: Reason-seg, where the model first generates the reasoning response and then outputs the final land-use segmentation token, and Seg-reason, where the final category and segmentation token are placed before the same reasoning response.

PixelLM and SegEarth-R2 use the same ComplexLU split under both output-order settings. All reasoning-segmentation models are trained with four GPUs, a per-device batch size of 2, gradient accumulation of 1, and an effective batch size of 8. We evaluate both 10,000-step and 20,000-step checkpoints in the appendix, while the main comparison reports the 20,000-step checkpoints.

Setting	Order	Steps	mAcc	mIoU	OA
Evidence-anchored	Reason-seg	10k	75.63	62.88	97.05
		20k	74.75	62.05	97.02
	Seg-reason	10k	78.45	66.04	97.27
		20k	79.39	66.30	97.38

Table 8: SegEarth-R2 results for the additional Evidence-anchored setting. Local evidence phrases are supervised through attention maps, while evaluation is performed only on the final complete land-use mask.

PixelLM uses a learning rate of 3×10^{-4}, and SegEarth-R2 uses a learning rate of 5×10^{-5}; other optimization settings follow the default recipes of the corresponding models unless otherwise specified. During evaluation, both Reason-seg and Seg-reason are evaluated with free generation under the same model-specific parsing protocol.

D Additional SegEarth-R2 Evidence Analysis

This section reports the additional Evidence-anchored SegEarth-R2 setting. The purpose of this diagnostic setting is to isolate whether attention-level supervision over local evidence phrases can improve the final complete land-use mask. Unlike the Evidence-grounded setting in the main table, local evidence in this setting is not supervised as an intermediate segmentation target. Instead, the evidence phrases are connected to SegEarth-R2's attention maps, while the mask loss remains defined only on the final land-use region.

Table 8 shows two patterns. First, Evidence-anchored Reason-seg does not consistently outperform the Target-only and Evidence-grounded

Model	Order	Setting	Steps	mAcc	mIoU	OA
SegEarth-R2	Reason-seg	Target-only	10k	72.75	60.43	96.84
			20k	75.34	64.83	97.13
		Evidence-grounded	10k	73.10	61.03	96.99
			20k	77.78	63.55	97.10
	Seg-reason	Target-only	10k	77.65	64.38	97.15
			20k	78.95	65.92	97.31
		Evidence-grounded	10k	75.80	63.73	97.12
			20k	77.65	65.22	97.19
PixelLM	Reason-seg	Target-only	10k	76.09	62.84	97.01
			20k	75.93	63.80	97.02
		Evidence-grounded	10k	76.23	63.21	96.99
			20k	75.01	61.66	96.96
	Seg-reason	Target-only	10k	78.12	65.83	97.16
			20k	78.72	66.02	97.16
		Evidence-grounded	10k	76.63	64.42	97.17
			20k	77.29	64.26	97.12

Table 9: Training-length comparison for reasoning-based free-generation settings.

Reason-seg settings in Table 2. It improves over Target-only Reason-seg at 10,000 steps, but drops at 20,000 steps, suggesting that attention-level evidence grounding can help early training but may not always translate into better final-mask recovery under the reasoning-first order. Second, Evidence-anchored Seg-reason is the strongest SegEarth-R2 variant, reaching 66.30 mIoU at 20,000 steps. This supports the main finding that placing the final category and mask token before long reasoning text stabilizes mask binding. Under this output order, attention-level evidence anchoring becomes more useful because the final land-use target has already been established before the model generates detailed evidence reasoning.

E Training-Length Analysis

We further compare 10,000-step and 20,000-step checkpoints to assess whether longer fine-tuning consistently improves reasoning-based segmentation. The 10,000-step checkpoint is saved from the same training trajectory as the corresponding 20,000-step checkpoint, so the comparison isolates continued training rather than variation across independent runs. The comparison is limited to the free-generation settings used in the main experiments, because these settings reflect the actual segmentation behavior evaluated in the paper.

Table 9 indicates that longer training is helpful but not uniformly so. SegEarth-R2 benefits from 20,000 steps in most settings, especially Target-only Reason-seg and Seg-reason, suggesting that its decoder and language-guided mask binding require more adaptation to ComplexLU. PixelLM shows smaller gains: Target-only improves slightly, whereas evidence-grounded Reason-seg decreases from 63.21 to 61.66 mIoU. This pattern suggests that longer training can overfit the reasoning template or local evidence mentions when the output sequence contains multiple grounded phrases. Overall, training length interacts with output order and evidence grounding; simply doubling the number of steps does not guarantee better final land-use masks.

F Task-Specific Prompt–Answer Examples

This section gives representative prompt–answer templates for the task settings used in ComplexLU. Across model implementations, the exact segmentation-token syntax can vary, but the text order, semantic content, and final land-use-mask evaluation target are kept consistent.

Referring segmentation: category-only.

- **Prompt:** Given the image and candidate land-use categories *[candidate categories]*, identify the present category and segment its land-use region. If the image does not contain any candidate category, answer *none*.
- **Answer:** *[detected category]* <seg>; or *none*.

Referring segmentation: expert-prior.

- **Prompt:** Given the image, candidate land-use categories *[candidate categories]*, and possible evidence element types *[element type list]*, identify the present category and segment its land-use region. If the image does not contain any candidate category, answer *none*.
- **Answer:** *[detected category]* <seg>; or *none*.

Reason-seg: Target-only.

- **Prompt:** Given the image and target category *[target category]*, generate a reasoning response and then output the final category with its land-use region.
- **Answer:** The image shows *[scene-level reasoning]*. Therefore, *[target category]* <seg> exists.

Reason-seg: evidence-grounded.

- **Prompt:** Given the image and target category *[target category]*, generate an evidence-grounded reasoning response and then output the final category with its land-use region.
- **Answer:** The image shows *[evidence phrase 1]* <seg> and *[evidence phrase 2]* <seg>, *[scene-level reasoning]*. Therefore, *[target category]* <seg> exists.

Seg-reason: Target-only.

- **Prompt:** Given the image and target category *[target category]*, first output the final category with its land-use region, and then generate a short reasoning response.
- **Answer:** *[target category]* <seg>: The image shows *[scene-level reasoning]*.

Seg-reason: evidence-grounded.

- **Prompt:** Given the image and target category *[target category]*, first output the final category with its land-use region, and then generate an evidence-grounded reasoning response.
- **Answer:** *[target category]* <seg>: The image shows *[evidence phrase 1]* <seg> and *[evidence phrase 2]* <seg>, *[scene-level reasoning]*.

G Ethics, Risks, and Artifact License

ComplexLU does not contain human-subject data, personal attributes, faces, or individual-level behavioral records. The benchmark focuses on remote-sensing interpretation of land-use regions and associated visual evidence. Potential risks mainly come from geospatial dual use: high-resolution mapping methods may support both benign planning applications and sensitive monitoring. We therefore position ComplexLU as a research benchmark for evaluating complex land-use segmentation, evidence grounding, and model diagnostics, rather than as an operational surveillance or enforcement system.

ComplexLU is intended for non-commercial research use, including method development, reproducibility studies, and diagnostic evaluation of vision-language models for thematic mapping. The released artifacts will include task definitions, annotations, prompts, evaluation code, and model outputs where permitted by the original access conditions. These artifacts do not grant additional rights over any third-party imagery, map services, or geospatial resources used to construct the benchmark. Users of the dataset are responsible for complying with the licenses and terms of the original data providers, and derivative datasets or models should be used consistently with those research-oriented access conditions.