MVT: MASK-GROUNDED VISION-LANGUAGE MODELS FOR TAXONOMY-ALIGNED LAND-COVER TAGGING

Siyi Chen[1,2,*] Kai Wang[3,*] Weicong Pang[4,*] Ruiming Yang[4] Ziru Chen[1]
Renjun Gao[5] Alexis Kai Hon Lau[1] Dasa Gu[1,†] Chenchen Zhang[1,†] Cheng Li[1,†,♣]

[1] HKUST [2] JHU [3] CUHKSZ [4] NUS [5] MUST

clieo@connect.ust.hk dasagu@ust.hk czhangej@connect.ust.hk

* Equal contribution † Corresponding authors ♣ Project Leader

***Abstract*—Land-cover understanding in remote sensing increasingly demands taxonomy-agnostic systems that generalize across datasets while remaining spatially precise and interpretable. We study a geometry-first discovery-and-interpretation setting under cross-dataset domain shift, where candidate regions are delineated class-agnostically and supervision avoids lexical class names via anonymized identifiers. Complementary to open-set recognition and open-world learning, we focus on coupling class-agnostic mask evidence with taxonomy-grounded scene interpretation, rather than unknown rejection or continual class expansion. We propose MVT, a three-stage framework that (i) extracts boundary-faithful region masks using SAM2 with domain adaptation, (ii) performs mask-grounded semantic tagging and scene description generation via dual-step LoRA fine-tuning of multimodal LLMs, and (iii) evaluates outputs with LLM-as-judge scoring calibrated by stratified expert ratings. On cross-dataset segmentation transfer (train on OpenEarthMap, evaluate on LoveDA), domain-adapted SAM2 improves mask quality; meanwhile, dual-step MLLM fine-tuning yields more accurate taxonomy-aligned tags and more informative mask-grounded scene descriptions. The project is available at https://charlescsyyy.github.io/MVT**

***Index Terms*—remote sensing, class-agnostic region discovery, taxonomy-grounded interpretation, segmentation, MLLMs**

I. Introduction

Earth observation is entering a big-data regime, increasing the need for scalable land-cover understanding from remote sensing imagery [1, 2]. Yet most pipelines remain *closed-set*: they assume a fixed taxonomy and degrade under cross-dataset domain shift and emerging or rare land covers [3–5]. This motivates *open-world* remote sensing, where systems should generalize beyond a dataset-specific label set while remaining useful for mapping and monitoring [6].

For practical mapping, outputs must be both spatially precise and interpretable: pixel-accurate regions provide geometric evidence, while standardized taxonomy tags enable consistent reporting. However, existing open-set methods often decouple these requirements by (i) rejecting "unknown" without interpretable reporting [7, 8], (ii) localizing coarsely (e.g., boxes) [9], or (iii) relying on predefined vocabulary sets that constrain naming under true novelty [10].

We propose **MVT**, a geometry-first framework that couples class-agnostic mask discovery with taxonomy-aligned tile-level interpretation under domain shift (Fig. 1). MVT uses a promptable segmenter (SAM2) to extract boundary-faithful masks as structured evidence, then adapts MLLMs via a lexical-label-free two-step LoRA schedule with mask cues injected as grounding inputs. Finally, we evaluate generated descriptions with an LLM-as-judge protocol calibrated by stratified expert ratings [11].

Our contributions are: (1) a taxonomy-agnostic discovery-and-interpretation setting for remote sensing under domain shift; (2) a mask-grounded, lexical-label-free MLLM tuning strategy that produces taxonomy-aligned tags and grounded descriptions; and (3) a scalable evaluation protocol combining LLM judging with expert calibration.

II. Related Work

A. Taxonomy-agnostic Perception and Promptable Segmentation

Remote-sensing semantic segmentation has evolved from CNN encoder–decoder architectures to stronger context-aggregation and transformer-based models that better capture multi-scale cues and preserve boundaries [12–17]. Beyond architectural advances, Remote Sensing (RS) specific efforts integrate multi-sensor fusion, structured refinement, shape priors, and domain generalization to mitigate cross-sensor and seasonal shifts [18]. Despite progress on benchmarks such as OpenEarthMap and LoveDA [19, 20], most pipelines remain closed-set and rely on dense semantic labels, which limits robustness to emerging land covers and unseen domains [21].

Open-set and open-world perception address unseen categories via unknown rejection, proposal mining, and incremental learning [22–26]. In remote sensing, heterogeneous sensors and evolving taxonomies further complicate these settings; open-set domain adaptation and large-scale Earth Observation (EO) Out-of-Distribution (OOD) detection explicitly study such shifts [8, 27]. Open-vocabulary detection and segmentation broadens label spaces through language supervision, but still depends on pre-specified vocabularies and alignment

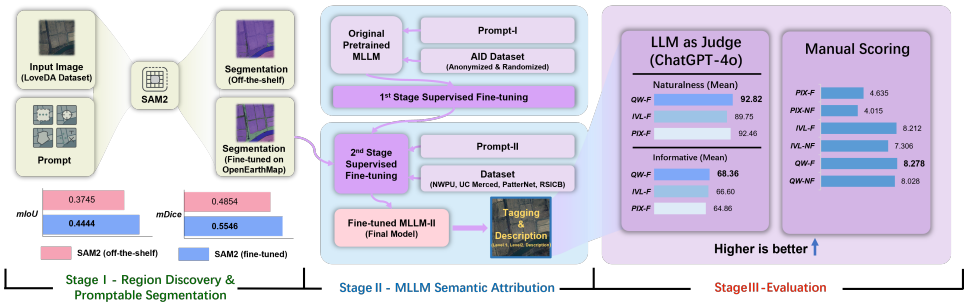

Fig. 1: Overview of the proposed MVT framework. The architecture includes the segmentation phase, the two-step fine-tuning of the MLLMs, and the evaluation phase.

quality [10]. In contrast, MVT targets taxonomy-agnostic, pixel-accurate region discovery under cross-dataset domain shift. We adopt SAM2 as a class-agnostic front end and apply adaptation to improve prompting robustness on high-resolution remote sensing imagery [28].

B. Taxonomy-Grounded Interpretation with MLLMs and Evaluation

Modern MLLMs (e.g., Flamingo, BLIP-2, LLaVA, and Qwen-VL/Qwen2-VL) enable open-ended naming and rich descriptions [29–32], yet can be brittle under open-set conditions: they may over-select from limited candidate sets or hallucinate plausible labels [33–35]. Remote-sensing MLLMs such as GeoChat and RS-LLaVA improve domain awareness [36, 37], but many studies remain scene-centric and do not explicitly support region-level grounding or standardized, taxonomy-aligned reporting [38]. Recent open-set RS works begin to leverage MLLMs for unknown discovery and naming by describing or labeling mined proposals [39, 40]. MVT advances this direction by using class-agnostic masks as structured grounding evidence. By replacing lexical labels with anonymized identifiers during training, our framework supports a standardized taxonomy interface independent of dataset-specific naming conventions.

For scalable assessment of language outputs, LLM-as-judge protocols provide rubric-based scoring that can correlate with human evaluation [41]. We adopt GPT-4o for automatic scoring and calibrate it with stratified expert ratings to improve robustness and reproducibility under remote-sensing domain shift [11].

III. Methodology

A. Data and Lexical-Label-Free Setup

This study evaluates MVT under cross-dataset domain shift with geometry discovery decoupled from semantic interpretation. Stage I adapts a promptable segmenter on OpenEarthMap and tests transfer on LoveDA [19, 20]. Stage II fine-tunes MLLMs in two steps: Step I uses AID; Step II adds NWPU-RESISC45, UC Merced, PatternNet, and RSI-CB, and injects SAM2 mask evidence [42–46, 28]. All models are evaluated on LoveDA.

To prevent lexical leakage, we remove class-name cues from file paths and metadata by mapping each label to an anonymized ID (e.g., *grassland→category01*) and shuffling all samples into a single directory. In total, 70,705 samples are used for MLLM tuning (Step I: 2,904; Step II: 67,801).

B. Stage I: Promptable Region Discovery with SAM2

To enable taxonomy-agnostic region discovery, we adopt SAM2 [28] as a class-agnostic, promptable segmentation front end. We compare: (i) off-the-shelf SAM2 pretrained on large-scale generic imagery, and (ii) a domain-adapted SAM2 fine-tuned on OpenEarthMap [19] using polygon annotations purely as *mask* supervision while discarding category names to avoid semantic leakage. Fine-tuning follows the SAM2 objective of producing high-quality prompt-conditioned masks [28].

The research evaluates cross-domain generalization on LoveDA [20], where each tile typically contains multiple co-occurring land-cover types with intricate boundaries. At inference, we apply point prompting to extract pixel-accurate, class-agnostic region masks. These masks serve as structured geometric evidence for Stage II. For segmentation metrics on LoveDA, we perform IoU matching between predictions and ground truth instances to compute instance mIoU and mDice.

C. Stage II: Two-Step MLLM Fine-Tuning for Taxonomy Tagging and Mask-Grounded Description

Stage II performs *tile-level* taxonomy tagging and description generation. Each LoveDA tile is assigned a single dominant land-cover label (the largest-area class in the LoveDA ground truth); SAM2 region masks are used only as grounding cues to support evidence-based reasoning, not as per-region semantic targets. We fine-tune three complementary MLLMs: Qwen2.5-VL-7B [32], Pixtral-12B [47], and InternVL3-8B-hf [48]. To keep supervision lexical-label-free, both steps use anonymized numeric pseudo-label IDs as training targets; semantic class names are not used as supervised outputs. We apply LoRA [49, 50] while freezing the vision backbone and multimodal projection layers to preserve pretrained multimodal representations.

a) Step I (lexical-label-free recognition).: Starting from the pretrained MLLM, Step I builds basic remote-sensing visual discrimination using the anonymized AID dataset [42]. Training prompts restrict the model to output *only* an anonymized numeric label ID (no free-form explanation) to prevent lexical leakage and force reliance on visual cues (layout, geometry, texture, contrast, and context). We train

LoRA with rank r=8 and α=16 for 120 steps using a cosine-decay schedule; batch size is 16 with gradient accumulation 8 and context length 8192.

b) Step II (mask-grounded refinement under a standardized taxonomy interface).: Step II continues from the Step I checkpoint and (i) expands training data with NWPU-RESISC45 [43], UC Merced [44], PatternNet [45], and RSI-CB [46], and (ii) injects SAM2-derived region cues as structured prompt inputs. Specifically, for each image/tile we append per-region annotations parsed from SAM2 JSON outputs, including bounding box $bbox(x, y, w, h)$, pixel area, and the pixel-level mask encoded as run-length encoding counts (with size) [28]. A standardized land-use taxonomy (Chinese Standard first-level categories) [51] is included in the prompt as a *reasoning scaffold*, while the supervised target remains strictly the anonymized ID. Step II uses the same LoRA configuration and frozen vision/projector layers, with optimization adjusted to emphasize subtle, evidence-sensitive distinctions (initial learning rate 1×10^{-5}, batch size 4, gradient accumulation 8, 220 steps). At inference and evaluation, we use a fixed prompt that constrains the model to (i) select exactly one Level-1 category from the Chinese-Standard Level-1 list provided in the prompt, (ii) output only the anonymized Level-2 ID, and (iii) generate a mask-grounded description; taxonomy terms are used only as a non-supervised reasoning scaffold.

D. Stage III: Evaluation of Tags and Descriptions

On LoveDA, we evaluate per-tile Level-1/Level-2 tags and generated descriptions using (i) expert manual scoring and (ii) an LLM-as-judge protocol [52]. Remote-sensing analysts assess tag correctness and description quality, while GPT-4o scores description *naturalness* and *informativeness*, with scores calibrated by stratified expert ratings for reliability [11].

IV. Experiments and Analysis

A. Environment Setup

For Stage I, we use the official SAM2 2.1429 Hiera-L checkpoint as the off-the-shelf baseline, finetuning uses AdamW on two 80 GB A100 GPUs (Python 3.10, PyTorch 2.8). For Stage II, we run experiments on two 96 GB NVIDIA H20 GPUs.

B. Segmentation Evaluation

This study evaluated segmentation quality on the LoveDA dataset, comparing the off-the-shelf SAM2 [28] with a fine-tuned SAM2 variant that is domain-adapted on OpenEarthMap. LoveDA contains multiple co-occurring land-cover types with intricate boundaries [20], posing higher requirements for geometric characterization and generalization ability. We report mean Intersection over Union (mIoU) and mean Dice score (mDice) [53].

For each annotated instance $i \in \mathcal{I}$, with predicted mask P_i and ground-truth mask G_i,

$$\mathrm{IoU}_i = \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \quad \mathrm{Dice}_i = \frac{2|P_i \cap G_i|}{|P_i| + |G_i|}. \tag{1}$$

We report dataset-level instance means:

$$\mathrm{mIoU} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathrm{IoU}_i, \quad \mathrm{mDice} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathrm{Dice}_i. \tag{2}$$

On LoveDA, SAM2 fine-tuned on OpenEarthMap improves instance-level mIoU/mDice to 0.4444/0.5546 vs. 0.3745/0.4854 off-the-shelf, and yields finer boundaries in complex urban–rural scenes (Fig. 2).

C. MLLM Evaluation

We evaluate MLLMs on LoveDA [20] to assess (i) tile-level Level-1/Level-2 tagging accuracy and (ii) description quality in our label-set-agnostic discovery-and-interpretation setting. We conduct a qualitative ablation by comparing Single-Step (S) baselines (directly fine-tuned with Step II only) against our Dual-Step (D) pipeline (full two-step schedule).

1) Visualization of Semantic Tagging: Table I compares Single-Step vs. Dual-Step outputs (Fig. 2). InternVL3 benefits most from Dual-Step tuning: IVL-D correctly predicts *Hydraulic Construction Land* and grounds it with man-made cues (e.g., dam/levee near water), whereas IVL-S tends to over-generalize to natural coastal patterns (e.g., tidal flats). Pixtral (PIX-S/D) is consistently reliable on *Paddy Field*. Qwen2.5VL shows clear refinement with Dual-Step tuning: QW-S misclassifies the scene as a public park, while QW-D shifts to *Water Bodies* (River Surface). Overall, Dual-Step tuning improves discrimination between subtle man-made hydraulic infrastructure and visually similar natural water/coastal scenes.

These qualitative gains match the intended step-wise roles: Step I builds coarse RS discrimination under lexical-label-free, ID-only supervision; Step II injects region-level geometric evidence (SAM2 masks, bounding boxes, and areas) to encourage explicit grounding and fine-grained subtype reasoning. Consistently, improvements are more pronounced for Level-2 tagging and description quality than for Level-1 categorization (Table II).

2) Manual Scoring: Manual scoring serves to rigorously assess semantic tagging accuracy at the scene (per-image/tile) level. We randomly select 25% (449 image tiles) of the LoveDA Dataset for evaluation. The evaluation metric incorporates three dimensions: first-level tagging, second-level tagging, and description quality. For each sample, Level-1 and Level-2 categories are scored binarily (0/1), while the description is graded on a three-level scale (0 for incorrect, 0.5 for partly correct, and 1 for fully correct), with a per-sample maximum of 3 points. The overall evaluation score (OES; full score is 9) is

$$\mathrm{OES} = \frac{\text{Level-1} + \text{Level-2} + \text{Description}}{\text{Number of Samples}} \times 3. \tag{3}$$

Results in Table II reveal a clear distinction between single and dual-step finetuned models. Dual-Step finetuned Qwen2.5VL and InternVL3 achieve the highest overall scores (8.278 and 8.212), significantly outperforming their single-finetuned counterparts. Pixtral shows the weakest tagging accuracy even after dual-step finetuning.

Model	Level-1	Level-2	Description
QW-S	PAPSL	Park and Green Space	The image shows a park-like area with a road running through it, surrounded by greenery and water bodies. This suggests a recreational or public space designed for the enjoyment and use of the public, which aligns with the characteristics of a park and green space.
QW-D	WBHFL	River Surface	The image shows a water body with a visible river surface, surrounded by land that appears to be developed with roads and some structures, which is consistent with the characteristics of the river surface.
IVL-S	WBHFL	Coastal Tidal Flats	The image shows a coastal area with tidal flats adjacent to a body of water. The presence of a road and some structures near the water indicates a combination of natural and managed land use typical of tidal flat environments.
IVL-D	WBHFL	Hydraulic Construction Land	The image shows a structure adjacent to a water body, likely a **dam or levee**, indicative of **hydraulic construction**. The surrounding areas include roads and possibly some vegetation, typical of locations requiring hydraulic management and infrastructure. This setting aligns with features associated with **water control and management** facilities.
PIX-S	Cultivated land	Paddy Field	The image shows a large expanse of water-filled fields indicative of rice cultivation, which is characteristic of paddy fields. The uniform and structured appearance of the fields, along with the presence of water, aligns with the typical features of paddy fields commonly found in agricultural regions. Additionally, the overall layout and the organized pattern of the fields further support this classification.
PIX-D	Cultivated land	Paddy Field	The image displays a distinct pattern of rectangular plots which are characteristic of organized agricultural land. The presence of water-filled paddies suggests that this is a rice cultivation area, commonly found in regions where paddy fields are prevalent. The uniform coloration and the arrangement further confirm that this is a structured farming system typically used for paddy field.

TABLE I: Comparison of different MLLMs Outputs in Single-Step (S) and Dual-Step (D) settings based on Figure 2. QW: Qwen2.5VL-7B, IVL: InternVL3-8B, PIX: Pixtral 12B. Level-1/2 denote hierarchical land-cover classification. WBHFL: Water Bodies and Hydraulic Facility Land, PAPSL: Public Administration and Public Service Land.

Method	Level-1	Level-1 Acc	Level-2	Level-2 Acc	Descr.	OES
QW-S	439	0.978	378	0.842	384.5	8.028
QW-D	433	0.964	**400**	**0.891**	406	**8.278**
IVL-S	405	0.902	331	0.737	357.5	7.306
IVL-D	**442**	**0.984**	379	0.844	**408**	8.212
PIX-S	204	0.454	174	0.388	185.5	3.765
PIX-D	247	0.550	204	0.454	218	4.470

TABLE II: Manual scoring results on the LoveDA evaluation subset.

Method	Mean	Min	Max	Q1	Q3	Med.	Var.	Std
QW-D	**92.82**	69	100	88.5	96.5	94.0	28.68	5.36
IVL-D	89.76	0	100	88.0	96.5	92.5	269.97	16.43
PIX-D	92.47	61	100	88.0	96.5	92.5	30.31	5.51

TABLE III: GPT-4o naturalness evaluation (scores in [0,100]).

Method	Mean	Min	Max	Q1	Q3	Med.	Var.	Std
QW-F	**68.36**	42.0	87.0	62.5	74.5	68.5	53.51	7.28
IVL-F	66.60	0.0	89.0	61.5	75.0	69.5	187.13	13.68
PIX-F	64.86	41.5	84.5	59.5	70.5	64.0	54.48	7.38

TABLE IV: GPT-4o informativeness evaluation (scores in [0,100]).

Fig. 2: (a) Original Image; (b) Off-the-shelf; (c) Fine-tuned.

Fig. 3: Hydraulic Construction Land example.

3) LLM-as-Judge with GPT-4o: To systematically assess the linguistic quality of generated descriptions, we adopt an LLM judge protocol [52] using GPT-4o [11]. Descriptions are evaluated along two axes: *naturalness* and *informativeness*. Naturalness is scored via five weighted sub-modules: grammar & syntax (0.25), discourse coherence & flow (0.25), lexical naturalness & idiomaticity (0.20), style & register (0.15), and human-likeness vs. "tells" (0.15). Informativeness is scored via coverage of key facets (0.25), specificity & quantification (0.25), concreteness & observability (0.20), context/constraints/relations (0.20), and relevance & non-redundancy (0.10). To ensure deterministic and objective scoring, we set the sampling temperature to 0 and utilize structured prompting to enforce single-line fixed numerical output format. The weighting follows established linguistic quality frameworks and manual-evaluation practice [54].

The per-sample score, scaled to $[0, 100]$, is computed as

$$\text{Score}_{\text{total}} = 100 \times \sum_{i=1}^{5} w_i \left(\frac{s_i}{5} \right), \tag{4}$$

Tables III and IV report descriptive statistics for naturalness and informativeness. For naturalness, Qwen obtains the highest mean (92.82), closely followed by Pixtral and InternVL3. For informativeness, Qwen again leads (68.36), with InternVL3 and Pixtral trailing. Combining manual accuracy and GPT-4o-based naturalness/informativeness, Qwen and InternVL3 emerge as the most effective models for taxonomy-grounded RS tagging and description in our discovery setting, while Pixtral shows limited semantic tagging capability despite reasonable linguistic fluency.

V. Conclusion

This work introduces MVT, an architecture-agnostic pipeline for class-agnostic region discovery and taxonomy-grounded interpretation of land cover under domain shift. A domain-adapted SAM2 provides boundary-faithful masks as pixel-level evidence, and two-step fine-tuning enables MLLMs to translate region cues into standardized tags and context-aware descriptions without relying on lexical label supervision. Overall, MVT offers a practical route to unify geometric fidelity and interpretable semantics for dynamic monitoring and cartographic updating. Future work will broaden comparisons, and extend evaluation to more classical withheld-class open-set protocols as well as more open-vocabulary naming.

References

[1] Y. Tang and etal, "Design of remote sensing image data analysis and processing platform based on environmental monitoring," in *Journal of Physics: Conference Series*, vol. 2136, no. 1. IOP Publishing, 2021, p. 012056.

[2] S. Gui and etal, "Remote sensing object detection in the deep learning era—a review," *Remote Sensing*, vol. 16, no. 2, p. 327, 2024.

[3] S. Song and etal, "Synthetic data matters: Re-training with geo-typical synthetic labels for building detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[4] X. Zhang and etal, "Remote sensing object detection meets deep learning: A meta-review of challenges and advances," 2023. [Online]. Available: https://arxiv.org/abs/2309.06751

[5] Z. Yu and etal, "Exploring foundation models in remote sensing image change detection: A comprehensive survey," 2024. [Online]. Available: https://arxiv.org/abs/2410.07824

[6] G. Wei and etal, "From word to sentence: A large-scale multi-instance dataset for open-set aerial detection," 2025. [Online]. Available: https://arxiv.org/abs/2505.03334

[7] M. Sodano and etal, "Open-world semantic segmentation including class similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 3184–3194. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Sodano_Open-World_Semantic_Segmentation_Including_Class_Similarity_CVPR_2024_paper.html

[8] J. Zheng and etal, "Open-set domain adaptation for scene classification using multi-adversarial learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 245–260, 2024.

[9] K. J. Joseph and etal, "Towards open world object detection," 2021. [Online]. Available: https://arxiv.org/abs/2103.02603

[10] C. Zhu and etal, "A survey on open-vocabulary detection and segmentation: Past, present, and future," 2024. [Online]. Available: https://arxiv.org/abs/2307.09220

[11] OpenAI and etal, "Gpt-4o system card," 2024. [Online]. Available: https://arxiv.org/abs/2410.21276

[12] J. Long and etal, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_for_Semantic_Segmentation_CVPR_2015_paper.html

[13] V. Badrinarayanan and etal, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," 2016. [Online]. Available: https://arxiv.org/abs/1511.00561

[14] Y. Wang and etal, "An improved semantic segmentation algorithm for high-resolution remote sensing images based on deeplabv3+," *Scientific Reports*, vol. 14, no. 1, p. 9716, 2024.

[15] Z. Cheng and etal, "Remote sensing image segmentation method based on hrnet," in *IGARSS*, 2020, pp. 6750–6753.

[16] L. Wang and etal, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," in *IEEE Geoscience and Remote Sensing Letters*, 2022.

[17] J. He and etal, "Shifted window segformer for remote sensing image segmentation," in *ICIIBMS*, 2023, pp. 117–121.

[18] J. Zhang and etal, "Rsam-seg: A sam-based model with prior knowledge integration for remote sensing image semantic segmentation," *Remote Sensing*, vol. 17, no. 4, 2025.

[19] J. Xia and etal, "Openearthmap: A benchmark dataset for global high-resolution land cover mapping," 2022. [Online]. Available: https://arxiv.org/abs/2210.10732

[20] J. Wang and etal, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2110.08733

[21] Q. Cao and etal, "Open-vocabulary remote sensing image semantic segmentation," *arXiv preprint arXiv:2409.07683*, 2024.

[22] A. Bendale and etal, "Towards open set deep networks," in *CVPR*, 2016.

[23] A. Dhamija and etal, "The overlooked elephant of object detection: Open set," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1021–1030.

[24] K. J. Joseph and etal, "Towards open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5830–5840. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Joseph_Towards_Open_World_Object_Detection_CVPR_2021_paper.html

[25] X. Du and etal, "Unknown-aware object detection: Learning what you don't know from videos in the wild," 2022. [Online]. Available: https://arxiv.org/abs/2203.03800

[26] Y. Li and etal, "Open world object detection: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 2, p. 988–1008, 2025.

[27] B. Ekim and etal, "Distribution shifts at scale: Out-of-distribution detection in earth observation," in *CVPR*, 2025, pp. 2265–2274.

[28] N. Ravi and etal, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[29] A. Awadalla and etal, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[30] J. Li and etal, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023, pp. 19 730–19 742.

[31] H. Liu and etal, "Visual instruction tuning," *NeurIPS*, vol. 36, pp. 34 892–34 916, 2023.

[32] S. Bai and etal, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[33] W. Kuo and etal, "F-vlm: Open-vocabulary object detection upon frozen vision and language models," *arXiv preprint arXiv:2209.15639*, 2022.

[34] Y. Zhou and etal, "Led: Llm enhanced open-vocabulary object detection without human curated data generation," *arXiv preprint arXiv:2503.13794*, 2025.

[35] Y. Tu and etal, "Ode: Open-set evaluation of hallucinations in multimodal large language models," in *CVPR*, 2025, pp. 19 836–19 845.

[36] K. Kuckreja and etal, "Geochat: Grounded large vision-language model for remote sensing," in *CVPR*, 2024, pp. 27 831–27 840.

[37] Y. Bazi and etal, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[38] Z. Ji and etal, "Step-wise hierarchical alignment network for image-text matching," 2021. [Online]. Available: https://arxiv.org/abs/2106.06509

[39] N. Saini and etal, "Advancing open-set object detection in remote sensing using multimodal large language model," in *WACVW*, 2025, pp. 451–458.

[40] J. Xie and etal, "Llama-unidetector: An llama-based universal framework for open-vocabulary object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[41] Y. Liu and etal, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2023, pp. 2511–2522.

[42] G.-S. Xia and etal, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[43] G. Cheng and etal, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[44] Y. Yang and etal, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[45] W. Zhou and etal, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, p.

197–209, 2018.
[46] H. Li and etal, "Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, p. 1594, 2020.
[47] P. Agrawal and etal, "Pixtral 12b," *arXiv preprint arXiv:2410.07073*, 2024.
[48] J. Zhu and etal, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.
[49] E. J. Hu and etal, "Lora: Low-rank adaptation of large language models," *ICLR*, 2022.
[50] Y. Zheng and etal, "Llamafactory: Unified efficient fine-tuning of 100+ language models," 2024. [Online]. Available: https://arxiv.org/abs/2403.13372
[51] B.-m. Chen and etal, "Explanation of current land use condition classification for national standard of the people's republic of china," *Journal of Natural Resources*, vol. 22, no. 6, pp. 994–1003, 2007.
[52] J. Gu and etal, "A survey on llm-as-a-judge," 2025. [Online]. Available: https://arxiv.org/abs/2411.15594
[53] K. Qiu and etal, "Noise-consistent siamese-diffusion for medical image synthesis and segmentation," 2025. [Online]. Available: https://arxiv.org/abs/2505.06068
[54] A. R. Lommel and etal, "Multidimensional quality metrics: a flexible system for assessing translation quality," in *Proceedings of Translating and the Computer 35*, 2013.