EarthGPT-X: A Spatial MLLM for Multi-level Multi-Source Remote Sensing Imagery Understanding with Visual Prompting

Wei Zhang, Miaoxin Cai, *Graduate Student Member, IEEE,* Yaqian Ning, Tong Zhang, *Graduate Student Member, IEEE,* Yin Zhuang†, *Member, IEEE,* Shijian Lu†, *Member, IEEE,* He Chen, *Member, IEEE,* Jun Li, *Fellow, IEEE*, and Xuerui Mao†

arXiv:2504.12795v3 [cs.CV] 6 Nov 2025

***Abstract*—Recent advances in natural-domain multi-modal large language models (MLLMs) have demonstrated effective spatial reasoning through visual and textual prompting. However, their direct transfer to remote sensing (RS) is hindered by heterogeneous sensing physics, diverse modalities, and unique spatial scales. Existing RS MLLMs are mainly limited to optical imagery and plain language interaction, preventing flexible and scalable real-world applications. In this article, EarthGPT-X is proposed, the first flexible spatial MLLM that unifies multi-source RS imagery comprehension and accomplishes both coarse-grained and fine-grained visual tasks under diverse visual prompts in a single framework. Distinct from prior models, EarthGPT-X introduces: 1) a dual-prompt mechanism combining text instructions with various visual prompts (i.e., point, box, and free-form) to mimic the versatility of referring in human life; 2) a comprehensive multi-source multi-level prompting dataset, the model advances beyond holistic image understanding to support hierarchical spatial reasoning, including scene-level understanding and fine-grained object attributes and relational analysis; 3) a cross-domain one-stage fusion training strategy, enabling efficient and consistent alignment across modalities and tasks. Extensive experiments demonstrate that EarthGPT-X substantially outperforms prior natural and RS MLLMs, establishing the first framework capable of multi-source, multi-task, and multi-level interpretation using visual prompting in RS scenarios. The code and dataset are available at *https://github.com/wivizhang/EarthGPT-X*.**

***Index Terms*—Multi-source, spatial reasoning, remote sensing (RS), multi-modal large language models (MLLMs).**

This work was supported by the National Natural Science Foundation of China under Grant 92152109, in part by the General Program of the National Natural Science Foundation of China under Grant 62371048, and in part by the National Natural Science Foundation of China under Grant T2225019, and in part by the MOE Tier-2 project of Singapore under Grant MOE-T2EP20123-0003. (*† Co-corresponding authors: Shijian Lu, Yin Zhuang, and Xuerui Mao*)

Wei Zhang is with the School of Interdisciplinary Science, Beijing Institute of Technology, Beijing 100081, China, and with the College of Computing and Data Science, Nanyang Technological University, Singapore (e-mail: w.w.zhanger@gmail.com).

Xuerui Mao is with Beijing Institute of Technology (Zhuhai), Zhuhai, 519088, China. State Key Laboratory of Explosion Science and Safety Protection, Beijing, 100081, China. (e-mail: maoxuerui@sina.com).

He Chen, Yin Zhuang, Miaoxin Cai, and Tong Zhang are with the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, Beijing Institute of Technology, Beijing 100081, China. (e-mails: chenhe@bit.edu.cn, yzhuang@bit.edu.cn, 3120220667@bit.edu.cn, bit_zhangtong@163.com).

Yaqian Ning is with the School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China.

Shijian Lu is with the College of Computing and Data Science, Nanyang Technological University, Singapore. (e-mail: Shijian.Lu@ntu.edu.sg).

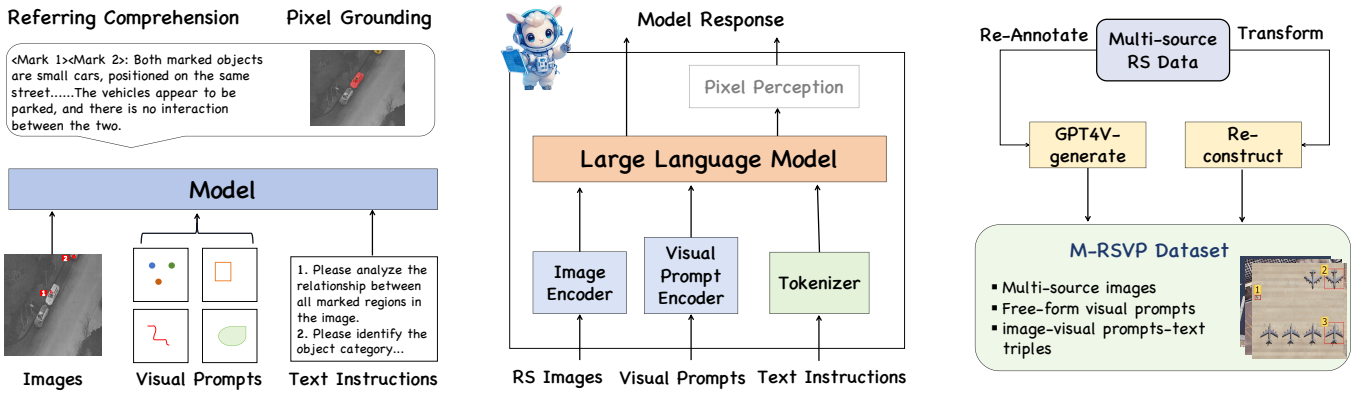
Jun Li is with the School of Computer Science and Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan, 430078, China (e-mail: lijuncug@cug.edu.cn).

I. Introduction

TYPICAL spatial understanding tasks in multi-modal learning require aligning image regions with text semantics [1]. Beyond holistic analysis, spatial tasks involve accurately interpreting specific regions or objects based on user instructions [2], [3]. Most recently, enabling multi-modal large language models (MLLMs) with comprehensive spatial understanding has become the focus of research in natural scenarios [4]–[6]. However, these models struggle with the remote sensing (RS) imagery, which features multi-source, multi-viewpoint, and scale variations due to different imaging conditions [2]. Although existing RS MLLMs can handle multiple RS visual tasks [7], [8], they primarily focus on holistic image understanding, lacking spatial perception. They solely rely on text instructions, which can lead to ambiguity and limited context. It is indispensable to develop a spatial-aware MLLM to interpret RS imagery under flexible interactions.

Notably, in the human world, people can achieve an all-around understanding of visual information, allowing for both zoom-in fine-grained analyses and zoom-out coarse-grained overviews [9]. Inspired by this, visual prompting models advance the field by injecting visual prompts beyond text instructions to enable multi-grained visual reasoning and enhance spatial perception [10]. Representative works include the promptable segmentation model Segment Anything (SAM) [11], which enables class-agnostic delineation from visual or text prompts and shows strong generalization as a versatile backbone. The region-of-interest model GPT4RoI [4] integrates large language models (LLMs) for localized reasoning, while Ferret [1] focuses on fine-grained grounding, which can accept diverse region inputs, such as points and bounding boxes. Additionally, Wu et al. [12] provide a review of visual prompting approaches, including prompt generation and prompt learning. However, designed for natural images, these models cannot be directly adapted to the heterogeneous RS domain due to fundamental differences in sensing physics, modalities, and spatial resolutions.

Limited efforts in the RS domain have been made to enhance spatial understanding. A key example is RSPrompter [13], which is based on SAM and develops an automated prompts generation approach to facilitate interactive segmentation. However, RSPrompter is only designed for instance segmentation. RSVG [2] defines the visual grounding



(a) **Task:** Multi-level RS Image Understanding Tasks (b) **Model:** EarthGPT-X (c) **Data:** Multi-source Training Data

Fig. 1. Overview of the proposed framework for flexible and comprehensive spatial understanding in RS imagery: (a) **Tasks**: a wide range of multi-level visual reasoning tasks; (b) **Model**: EarthGPT-X integrates an image encoder, a visual prompt encoder, a text tokenizer, an LLM, and a pixel perception module; (c) **Data**: the training dataset M-RSVP is constructed from multi-source RS data (i.e., optical, synthetic aperture radar (SAR), infrared) and further enriched with GPT-4V–based annotations.

tasks in the RS domain and can locate the referred objects using boxes under natural language expressions. Recent notable works, such as EarthGPT [7] and Geochat [8], have realized RS holistic and regional imagery comprehension under the unified visual-language learning framework, facilitating the development of MLLMs in the RS domain. Nevertheless, these models solely support language interaction without visual prompts to realize referring expression comprehension. Notably, EarthMarker [14] is a visual prompting model in the RS domain, however, it lacks support for multi-source imagery and free-form visual prompts, hampering flexible human-AI interaction and comprehensive understanding of RS data. Thus, enabling MLLMs to interpret multi-source RS imagery at any granularity remains in its infancy.

To overcome these limitations, we propose EarthGPT-X, a spatial MLLM built upon visual prompt learning and tailored for multi-source RS imagery. It supports arbitrary visual prompt types (e.g., points, boxes, circles, and scribbles) and enables comprehensive interpretation from overall scene understanding to fine-grained regional analysis, thereby providing more flexible and scalable visual–language interaction than prior approaches. As depicted in Fig. 1, the EarthGPT-X architecture consists of three interconnected components: multi-level RS image understanding *tasks*, a visual prompting *model*, and a construction pipeline of training *dataset*. Specifically, EarthGPT-X not only allows for zoom-in fine-grained reasoning (e.g., free-form region captioning, referring object classification, and object inter-relationship analysis), but also supports zoom-out coarse-grained tasks (e.g., image-level captioning, scene classification). Secondly, the key of the model is to realize multi-modal content integration, including images, visual prompts, and text instructions. The multi-modal signals are then fed into the LLMs for a one-stage all-in-one training. Additionally, a pixel perception module is integrated to extend the model with object segmentation masks. Most importantly, the training dataset named M-RSVP is constructed, transforming from existing multi-sensor datasets, which are uniformly reorganized and relabeled as *image-visual prompts-text* triples. In addition, a portion of the data re-annotated using GPT-4V [15] to generate longer and more detailed descriptions. A side-by-side comparison of EarthGPT-X with recent RS MLLMs is presented in Tab. I, highlighting its advancement in modality coverage, dataset quality, prompting flexibility, and fine-grained multi-task reasoning.

To further validate EarthGPT-X's comprehensive interpretation ability on multi-source imagery, we compare it with recent state-of-the-art (SOTA) natural and RS MLLMs. The results indicate that EarthGPT-X achieves accurate multi-source reasoning and provides more detailed RS image understanding. Overall, EarthGPT-X not only attains competitive performance on visual–language benchmarks, but also demonstrates remarkable advantages in complex object-level reasoning tasks.

Our contributions can be summarized as follows.

- ***EarthGPT-X: First Multi-source Spatial MLLM.*** We propose EarthGPT-X, a spatial MLLM designed for multi-source (i.e., optical, SAR, and infrared) RS imagery understanding. It advances existing RS MLLMs by supporting free-form visual prompts (e.g., points, boxes, scribbles) to enable flexible multi-granularity interaction. EarthGPT-X unifies diverse spatial reasoning tasks into a visual prompting framework, providing a generalizable interpretation across heterogeneous RS data.
- ***M-RSVP: First Multi-source Visual Prompting Dataset.*** The first multi-source visual prompting dataset for RS, titled M-RSVP, consisting of over 650k *image-visual prompt-text* triples, is constructed. The constructed M-RSVP bridges the gap of multi-source visual prompting data in the RS domain, enabling multi-granularity image interpretation across diverse scenes and sensors.
- ***Multi-domain Multi-modal Multi-level Interaction***. A hybrid signal mutual understanding mechanism is designed to strengthen the interplay between dense image features, sparse visual prompts, and text instructions, thereby improving the interpretation of referring regions within holistic images. Moreover, our cross-domain one-stage fusion training significantly streamlines the training

TABLE I
COMPARISONS OF EARTHGPT-X *v.s.* RECENT MLLMS.

Model	Textual	Modalities		RS Dataset		Tasks		Visual Prompts			Pix.
	Prompt	Opt.	Multi.	Pub.	Refi.	Regu.	Comp.	Point	Box	Free-form	Grd.
Shikra (*arXiv 2023*) [6]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗
GPT4RoI (*ECCV 2024*) [16]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗
ViP-LLaVA (*CVPR 2024*) [17]	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗
RS-LLaVA (*RS 2024*) [18]	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
GeoChat (*CVPR 2024*) [8]	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
EarthGPT (*TGRS 2024*) [7]	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
Popeye (*JSTAR 2024*) [19]	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓
LHRS-Bot (*ECCV 2024*) [20]	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
VHM (*AAAI 2025*) [21]	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗
EarthDial (*CVPR 2025*) [22]	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
EarthMarker (*TGRS 2025*) [14]	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗
EarthGPT-X (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: Comparisons in terms of modalities (***opt****ical or* ***multi****-source*), dataset construction (*converted directly from* ***pub****lic data or* ***Refi****ned using powerful MLLMs such as GPT-4V or Gemini*), supported tasks (***Regu****lar visual tasks or* ***comp****lex reasoning such as inter-relationship analysis*), visual prompting capabilities, and extended pixel-level grounding.

procedure of MLLMs, while ensuring seamless integration of multi-modal knowledge.

- ***Superior Performance.*** Extensive experiments benchmark EarthGPT-X against SOTA natural and RS MLLMs across diverse multi-level visual tasks and multi-source imagery, outperforming existing models. Furthermore, EarthGPT-X also excels in complex reasoning tasks, such as object inter-relationship analysis. These advances collectively represent a significant milestone toward developing versatile and flexible MLLMs for RS scenarios.

II. RELATED WORK

A. MLLMs in Natural Scenarios

The rapid development of LLMs has significantly advanced natural language processing (NLP), demonstrating remarkable capabilities in diverse contexts [23]–[25]. Then, an essential step toward achieving general intelligence is integrating multi-modal perception into LLMs, allowing them to handle multi-modal data beyond text. Several works [26]–[29] focus on incorporating multi-modal inputs to support task-specific objectives across various modalities. Models like VisualGPT [26], Flamingo [29], and SPHINX-X [30] have shown strong multi-modal reasoning abilities by aligning LLMs with visual data. This alignment is achieved through projection layers, zero-shot attention, and intermediate networks, as in MiniGPT-4 [31], LLAMA-Adapter V2 [32], and mPLUG-Owl [33]. More recent advances include Qwen2.5-VL [34], MiniCPM-V2.6 [35], InternVL2.5 [36], InternLM-XComposer2.5 [37], and DeepSeek-VL2 [38]. These enhanced versions advance prior frameworks by broadening input modalities, leveraging higher-quality large-scale datasets, and extending the scope of multi-modal reasoning tasks.

B. Specialist and MLLMs in RS Scenarios

RS imagery encompasses diverse modalities such as SAR, optical, and infrared, each capturing complementary physical characteristics of the observed scene. Recent advanced SAR detection adopts generative or predictive paradigms, e.g., MaDiNet [39] with Gamma Diffusion–based bounding-box generation and SAR-JEPA [40] with joint-embedding predictive recognition. In optical understanding, specialist models such as CSDS [41] are designed for scene classification, while Qonly [42] targets VQA tasks. For infrared detection, IRSAM [43] enhances SAM to achieve more discriminative representations. Although effective for their respective tasks, these specialist models lack the ability to perform multiple visual tasks in a unified architecture, thus limiting scalability and hindering broader real-world deployment.

Inspired by the success of MLLMs in natural image domains, RS research has also begun to adopt MLLMs with notable progress. RSGPT [44], built on Instruct-BLIP [45], demonstrates strong performance in captioning and VQA but struggles with classification and detection. Cross-modal transfer learning in the RS MLLM field has recently gained traction, with representative works exemplifying different stages of progress [46], [47]. GeoChat [8] represents the first versatile RS MLLM for multiple optical tasks, while SkyEyeGPT [48] further extends broader task coverage, EarthGPT [7] bridging MLLMs with optical, SAR, and infrared domains. LHRS-Bot [20] additionally contributes by introducing LHRS-Bench for standardized evaluation. These efforts are collectively paving the way for more unified and versatile frameworks. However, they remain limited to text-only prompts, lacking flexible interactive comprehension under visual instructions.

C. Prompting Learning

Prompt learning has become a key research direction in NLP [23], with works such as AutoPrompt [49] and CoOp [50] automating prompt design for language and visual–language models. Subsequently, the idea evolved into visual prompting [12], marking a broader shift from language-based to

multi-modal prompts. A milestone in this line is the SAM [11], which enables class-agnostic segmentation via point, box, or text prompts and demonstrates strong generalization across diverse domains. GPT4RoI [4] and RegionBlip [5] leverage spatial boxes and combine language with region-of-interest input for regional recognition. Colorful Prompting Tuning (CPT) [51] uses color-based prompts to enhance MLLMs' performance. Osprey [3] goes further by incorporating fine-grained mask regions into language instructions for pixel-level understanding. ViP-LLava [46] supports arbitrary visual cues like scribbles and arrows to interpret images flexibly. Wu et al. [12] provide a comprehensive survey of visual prompting approaches, covering visual prompting, prompt generation, and prompt learning. These models are primarily designed for the natural image domain and fail to generalize to the RS multi-source data. Notably, EarthMarker [14] is a visual prompting framework tailored for the RS data, but it remains confined to optical imagery and does not support flexible prompt types. These gaps highlight the necessity of a specialized model that can support multi-source imagery and flexible interaction in RS scenarios, motivating the development of the proposed model.

III. Methodology

The EarthGPT-X framework is detailed in this section. We first elaborate on the multi-modal content integration mechanism (Section III-A). Then, the cross-domain one-stage fusion training strategy is presented (Section III-B). The extended pixel perception module is introduced in Section III-C.

A. Multi-modal Content Integration

The multi-modal input stream of the EarthGPT-X is characterized by its tripartite structure. The multi-modal content encoding and integration are detailed as follows.

1) Visual Global and Local Perception: As shown at the bottom left of Fig. 2, to enhance visual comprehension, the high-resolution input images are divided into sub-images for local visual perception. Concurrently, the images are downsampled into lower resolutions and combined with the sub-images for subsequent mixed encoding. The Mixed Visual Encoders (MVE) are designed to encode input images and their variants I^i into visual representations V_{img}. To leverage the advantages of different network architectures (CNN and Vision Transformer (ViT) [52]) and integrate the local information and contextual dependencies, the semantically complementary backbones CLIP-ConvNeXt [53] and ViT self-supervised by DINOv2 [54] are employed as MVE. This visual encoding process can be formulated as

$$V_{\text{img}} = \text{Concat } (\text{MVE } (I^i)),\ i = 1, 2, ..., N. \tag{1}$$

2) Free-form Spatial-aware: Free-form visual prompts are expected to be flexibly used to guide the model in generating appropriate responses. The shapes of visual prompts include points, boxes, circles, scribbles, etc. To meet the need for flexible interaction at any granularity, a noise augmentation method is developed. Specifically, Gaussian distribution based on the dimensions of the box prompts is sampled to simulate the free-formed shapes. The introduced noise encompasses only a portion of the ground truth or expands it. Subsequently, considering that the visual prompts exhibit varying degrees of sparsity across different shapes, as illustrated in the bottom right of Fig. 2, a randomly initialized ViT is employed as a spatial-aware encoder (SAE) specifically for feature extraction, without any pre-trained weights suitable for image features. The various types of visual prompts V_{p} are encoded using the designed SAE. The implementation equation can be expressed as follows

$$E_{\text{prompt}} = \text{SAE } (V_{\text{p}}), \tag{2}$$

where E_{prompt} denotes the encoded visual prompts features with free-form shapes.

3) Text Instructions Tokenizer: As displayed in left of Fig. 2, to facilitate linguistic comprehension, the text instructions T_{orig} are encoded by the tokenizer [55] module, and projected into language embeddings L_{instruct}, preparing for the subsequent LLM to capture the nuanced interdependence. The detailed process after disassembly is first to use the subword tokenization [55] method to divide text instructions into subwords, enabling understanding of the structure of the linguistic signals. Subsequently, the Word2Vec [56] is adopted to map each subword to a one-dimensional vector embedding. The tokenizer process can be simply described as

$$L_{\text{instruct}} = \text{Tokenizer}(\ T_{\text{orig}}). \tag{3}$$

4) Hybrid Signals Mutual Understanding: To bridge the gap and enhance the interplay between dense visual representations and sparse visual prompts features, a hybrid signal mutual understanding mechanism is proposed. Specifically, as displayed in the middle of Fig. 2, the image representations V_{img} and visual prompts features E_{prompt} are processed by a self-attention layer and feed-forward layer to obtain V_{sa} and E_{ff}, respectively. The subsequent cross-attention and second self-attention blocks are used to compute the correlation H_{ca} between hybrid referring information and multi-scale holistic signals. In essence, this designed mechanism enhances detailed and contextual visual perception, leading to a deeper comprehension of the referring areas in holistic images. The whole procedure can be mathematically represented as

$$V_{\text{sa}} = \text{Self-attention } (V_{\text{img}}), \tag{4}$$

$$E_{\text{ff}} = \text{Feed-forward } (E_{\text{prompt}}), \tag{5}$$

$$H_{\text{ca}} = \text{Cross-attention } (V_{\text{sa}},\ E_{\text{ff}}), \tag{6}$$

$$[V\text{-}P]_{\text{hybrid}} = \text{Self-attention } (H_{\text{ca}} + E_{\text{ff}}), \tag{7}$$

where $[V\text{-}P]_{\text{hybrid}}$ is the output of this designed module, containing rich visual information and spatial-aware knowledge from the free-form visual prompts. Then, the hybrid visual features are projected to the same dimension as text embeddings through the vision-to-language modality-align projection layer $\mathcal{F}_{\text{vl-p}}$. This process can be formulated as

$$[V\text{-}P]_{\text{proj}} = \mathcal{F}_{\text{vl-p}}\ ([V\text{-}P]_{\text{hybrid}}). \tag{8}$$

where $[V\text{-}P]_{\text{proj}}$ involve both multi-scale visual signals and free-form spatial-aware information. Notably, the cross-attention layer is updatable during training.

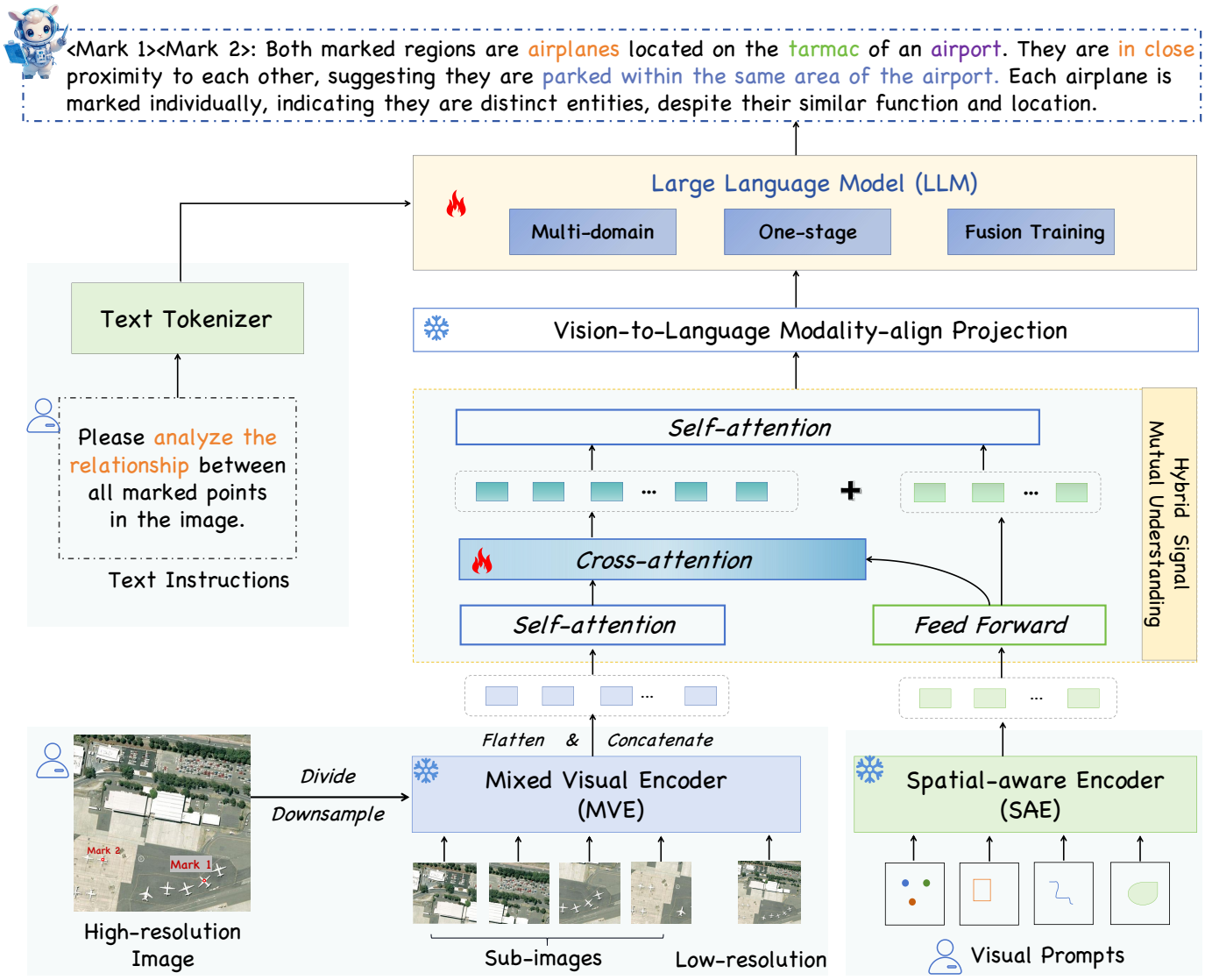


Fig. 2. Overview of the proposed EarthGPT-X architecture. It integrates images, visual prompts, and text instructions through a hybrid signal interaction mechanism, and employs a unified one-stage fusion training strategy to enable multi-source knowledge integration and multi-level multi-task reasoning.

B. Cross-domain One-Stage Fusion Training

The conventional training pipeline of MLLMs typically follows a two-stage approach [7], [14], which requires extensive manual intervention. In particular, researchers must carefully configure trainable parameters and dataset combinations across different stages, making the process complex and labor-intensive. Different from the prior models, we adopt a cross-domain one-stage fusion training strategy [30], which streamlines the process by uniformly treating all collected datasets while preserving strong performance. More importantly, this unified paradigm facilitates mutual understanding across heterogeneous domains, including natural, optical, SAR, and infrared imagery.

In this stage, the projected hybrid tokens are concatenated with the text instruction embeddings L_{instruct} to develop the multi-modal LLM input $\mathcal{X}$. Subsequently, the integrated multi-modal input is injected into LLM for alignment and deeply interactive understanding (see Fig. 2 LLM part). The process can be expressed as

$$Y = \text{LLM}\ (\ [V\text{-}P]_{\text{proj}},\ L_{\text{instruct}}). \tag{9}$$

Unlike other models that introduce additional learnable parameters, to simplify the model structure, the Transformer's attention layers are unfrozen during the training. Specifically, the self-attention head consists of key K, query Q, and value V, which are implemented by linear layers. The implementations can be expressed as follows

$$Q(\mathcal{X}) = W_{\text{q}}\ \mathcal{X} + b_{\text{q}}, \tag{10}$$

$$K(\mathcal{X}) = W_{\text{k}}\ \mathcal{X} + b_{\text{k}}, \tag{11}$$

$$V(\mathcal{X}) = W_{\text{v}}\ \mathcal{X} + b_{\text{v}}, \tag{12}$$

where $\mathcal{X}$ denotes multi-modal input. The parameters W_{q}, W_{k}, W_{v}, b_{q}, b_{k}, and b_{v} are updatable during the training.

To summarize, EarthGPT-X leverages multi-domain knowledge to generalize versatile visual reasoning into multi-source

RS scenarios. By tuning on abundant unified dual-prompting data, it achieves flexible referring comprehension with free-form visual prompts and text instructions.

C. Integrated with Pixel Perception Module

To further enhance spatial perception, we integrate a pixel perception module into EarthGPT-X, enabling pixel-level grounding under flexible visual prompts. In particular, the special token pix is adopted to represent the pixel grounding masks as embeddings [57]. The embeddings L_{pix} are generated from the LLM output corresponding to the pix token. Then, the language-to-pixel projection layer $\mathcal{R}_{\text{lp-p}}$ transforms the embeddings L_{pix} into the pixel decoder's feature space. In addition, the grounding encoder is denoted as $\mathcal{G}$, the pixel decoder $\mathcal{P}$. The binary segmentation mask M generated process can be represented as follows

$$M = \mathcal{P}\Big(\mathcal{R}_{\text{lp-p}}(L_{\text{pix}}), \mathcal{G}(I)\Big). \tag{13}$$

In summary, EarthGPT-X, built on a tripartite architecture and an all-in-one training strategy, enables effective cross-modal and multi-source knowledge comprehension. It further supports multi-grained RS visual tasks with flexible integration of referring and grounding in a unified framework.

TABLE II
STATISTICS OF TRAINING DATA.

Tasks	Raw Data	#Samples	Domain
Image Captioning	RSICD [58]	24.3k	Optical
	UCM-Captions [59]	10k	
Scene Classification	RSSCN7 [60]	0.56k	
	EuroSAT [61]	5.4k	
ROC (point)	SIOR [62]	200k	
Region Captioning	DIOR-RSVG [2]	31.5k	
ROC (box)	AIR-SARShip [63]	1.43k	SAR
	SSDD [64]	1.86k	
	SARDet [65]	106k	
	SAR Aircraft [66]	1.9k	
	SAR-Ship [67]	33.7k	
	MSAR [68]	32.5k	
ROC (box)	Sea-shipping [69]	8.4k	Infrared
	Infrared-security [70]	9.6k	
	Aerial-mancar [71]	31.6k	
	Double-light-vehicle [72]	5.1k	
Captioning, Relationship Analysis *(GPT-assisted)*	AIR-SARShip [63]	0.5k	SAR
	SARDet [65]	106k	
	SSDD [64]	1.0k	
	SAR Aircraft [66]	1.9k	
	Infrared-security [70]	9.6k	Infrared
	Aerial-mancar [71]	31.6k	

Notes: 'ROC' denotes Referring Object Classification. The 'box' and 'point' refer to box-level and point-level visual prompts, respectively.

IV. Dataset Construction

To support comprehensive multi-source RS understanding, we construct a large-scale dataset called M-RSVP, which comprises over 650k samples spanning various tasks, modalities, and annotation formats. The detailed introduction is as follows.

A. Multi-Source Multi-Domain Data Construction

For developing multi-source and versatile image understanding abilities of EarthGPT-X, it is indispensable to utilize data from different scenes and sensors, including outdoor natural images, RS optical images, SAR, and infrared images, covering different visual task types. Thus, as shown in Tab. II, we construct a multi-domain collection of raw data covering diverse vision-language tasks. For the RS optical modality, we adopt RSVP [14], the first visual prompting dataset in the RS domain, which enables the model to learn region- and point-level understanding in complex geographic environments. Meanwhile, we incorporate more optical datasets [2], [62] to enrich the diversity of tasks and enhance the model's generalization in RS scenarios. For SAR and infrared modalities, visual prompting data are mainly derived from publicly available object detection datasets, which are reorganized and annotated in a unified instruction format. Additionally, the natural scene dataset we selected is OpenImages [73] and Visual Genome [74], which meets this characteristic and is ideal for training the model's visual perception capabilities in general environments. The multi-domain data integration ensures comprehensive and flexible image understanding across multiple domains and sensor types.

B. Multi-level Prompting Data Annotation

A meticulously designed multi-strategy annotation method is employed to label diverse multi-domain data, which is uniformly reorganized and formatted into *image–visual prompt-text* triples. Notably, visual prompts include bounding boxes, points, and free-form types. The annotation process is detailed as follows.

1) Box prompts: The regional data are based on detection datasets, and ground truth bounding boxes are used as visual prompts to recognize object-level categories. As shown in Fig. 3 (a), an example of a SAR image is provided to illustrate the process. The traditional object detection data are transformed into multi-modal conversational samples, each comprising an image, visual prompts, and text (question-answer pairs). The image is directly sourced from the detection dataset, while the visual prompt is generated by applying slight perturbations to the ground-truth boxes, and the textual prompts are drawn from predefined templates. Specifically, image-level tasks use the boxes $[0, 0, \text{width}, \text{height}]$, which can cover the entire image as the visual prompt.

2) Point prompts: To enable point-based interaction during training, we construct point prompt annotations by randomly sampling pixels within each ground-truth object mask from semantic segmentation datasets. As illustrated in Fig. 3 (b), each image is divided into 32×32 patches, and the points are randomly sampled within each patch as the visual prompts, where each point corresponds to a pixel-level label annotation.

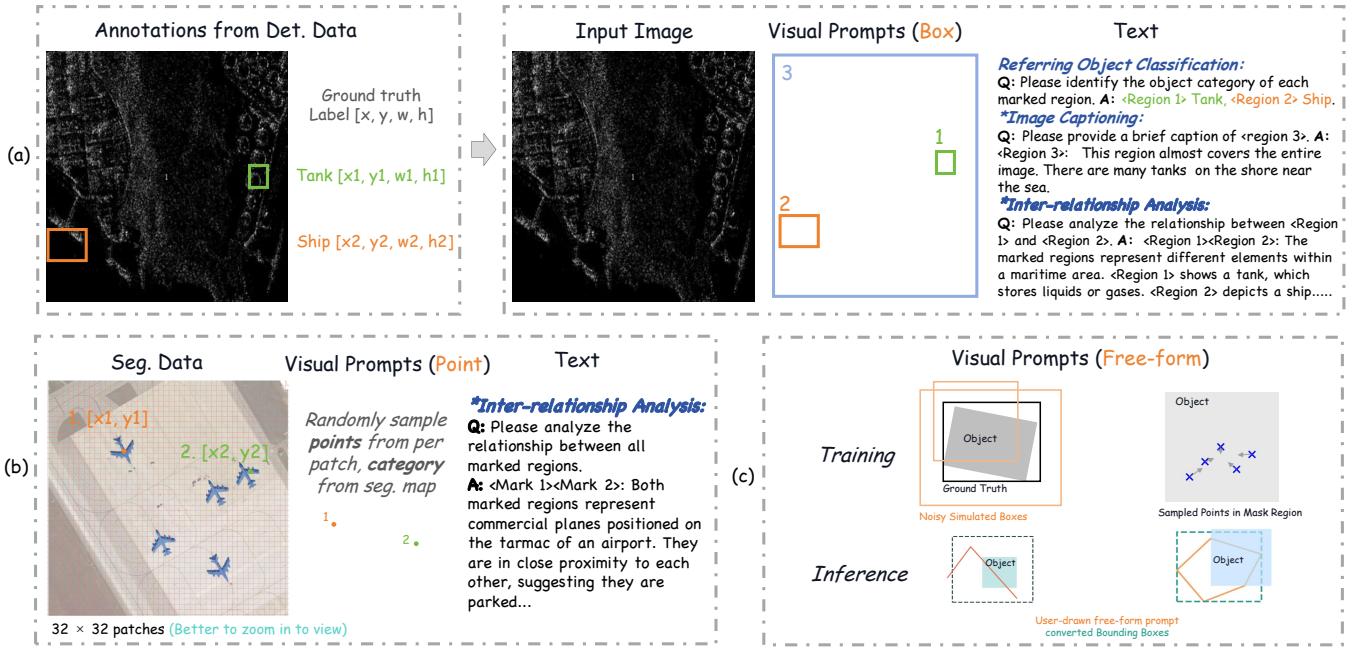


Fig. 3. (a) **Box prompts**: An example of converting SAR object detection data into multi-modal conversation training data. The training data consists of three parts: the image, the visual prompt, and the text (question and answer). The image is taken directly from the object detection dataset, while the visual prompts are from the ground-truth bounding boxes. The text prompts can be selected from fixed prompt templates (*note: enhanced by GPT-4V). (b) **Point prompts**: each image in the semantic segmentation dataset is divided into 32 × 32 patches, randomly sampled within each patch as the point prompts, with the category retrieved from the corresponding segmentation map. (c) **Free-form prompts**: Simulated by applying random noise-based augmentations to regions of interest. These free-form cues emulate flexible user interactions, enabling the model to generalize beyond structured visual prompts.

3) Free-form Prompts: To enable the model to handle diverse and flexible visual inputs at inference time, we simulate free-form prompts through noise-based augmentation methods. Specifically, we consider both box and point prompts and augment them to reflect the imprecision and looseness of user-defined inputs, as illustrated in Fig. 3 (c). First, for box prompt augmentation, to mimic diverse user-drawn free-form inputs, we apply Gaussian noise to the original ground-truth bounding boxes. Let $\mathcal{B} = (x, y, w, h)$ denote a bounding box parameterized by its top-left corner and dimensions. We generate an augmented box $\mathcal{B}'$ via:

$$\mathcal{B}' = \mathcal{B} + \mathcal{N}(0, \Sigma), \quad \text{where } \Sigma \propto \text{diag}(w^2, h^2, w^2, h^2). \quad (14)$$

This noise injection perturbs both the location and scale of the box in a size-aware manner, leading to boxes that may loosely cover, shift from, or extend beyond the target object, thus statistically approximating free-form visual prompts. For point prompts, we simulate user interactions by randomly sampling multiple pixels within the ground-truth object mask and associating them with the same instance label. Given a binary mask $\mathcal{M} \subset \Omega$ representing a target object, we uniformly select K points from the interior of $\mathcal{M}$. These sparse prompts guide the model to localize the target object via point-based reference. Additionally, at inference (see bottom part of Fig. 3 (c)), free-form inputs are pre-processed into bounding boxes, enabling flexible, intuitive user-drawn prompts.

4) Text Prompts: To support instruction tuning on multi-source RS data, we provide a set of carefully designed natural language instructions that correspond to different visual prompts (e.g., box or point). Each instruction guides the model to perform a specific visual understanding task, such as classification, captioning, or relational reasoning. Designed task instructions are as follows:

- *Please identify the object category of each marked region in the image.*
- *Please identify the labels of each marked point in the image.*
- *Please provide a brief caption of each marked region in the image.*
- *Please provide a detailed caption of each marked point in the image.*
- *Please analyze the relationship between all marked regions in the image.*
- *Please provide a summarized caption based on all the marked regions in the image.*
- *Please analyze how the marked objects interact with each other in the given scene.*

In addition, users can employ the <Mark n> or <Region n> identifiers to specify pixels or regions of interest and freely ask questions about them. These task-specific textual instructions guide the model to perform multiple visual tasks, promoting a more fine-grained understanding of visual content under different granularities.

C. GPT4V-assisted Data Generation

To enable high-quality annotations with GPT-4V [75], we follows the design principles of [76] and developed a set of rigorous and standardized prompt templates, as illustrated in Fig. 4. These templates are tailored to distinct tasks: brief captioning, detailed captioning, and inter-relationship analysis.

(a) Public Prompt Template

Image =



Messages = {"role":"system", "content": In the image, I have marked each visual object with a red point, and each is identified by a numeric ID against a black background.

The categories of each object are as follows:
<Mark 1>: Plane, <Mark 2>: Plane, <Mark 3>: Car, <Mark 4>: Tree

Your analysis will revolve 3 main functions:
<Role>
I will now provide predefined output templates corresponding to 3 specific roles in the context of remote sensing imagery. Please follow these templates exactly as given, without introducing any modifications, including extra spaces, punctuation, or line breaks. Strict adherence to the format is essential for maintaining consistency and ensuring accurate alignment between multi-modal remote sensing inputs and structured outputs.

Format:
<Format>
Proceed with your analysis, keeping the language natural and clear.}

(b) The <Role> and <Format> Template

Role
<Role 1 (Brief Captioning)>
Provide a concise and standalone caption for each marked region in the remote sensing image. Avoid referring to or comparing with other regions. Each caption should clearly reflect what is observed within the specific region.
<Role 2 (Detailed Captioning)>
Describe each marked region in detail using natural language. Include the object's category, structure, spatial location, surrounding context, and any visible features. Treat each region independently, and ensure the caption is no less than 30 words. Envision interpreting the satellite view directly and convey your observations as thoroughly and accurately as possible.
<Role 3 (Inter-Relationship Analysis)>
In your analysis, refer to specific regions using identifiers such as <Mark 1>, <Mark 2>, and so on. Clearly explain the spatial or functional relationships between these regions. Your discussion may include relative positioning, land use patterns, co-occurrence, differences, or environmental associations. If certain relationships are weak or not meaningful, you may omit them. For isolated or uncorrelated regions, emphasize their distinctiveness, provide a detailed, self-contained interpretation.
Format
Role 1 :
<Mark 1> Your Brief Captioning, ...,<Mark N> Your Brief Captioning
Role 2 :
<Mark 1> Your Detailed Captioning,.....,<Mark N> Your Detailed Captioning
Role 3:
<Mark 1><Mark 2>...<Mark N>: Your Detailed Analysis,.....,<Mark 1><Mark 2>...<Mark N>: Your Detailed Analysis

Fig. 4. Illustration of the standardized prompt templates designed for GPT-4V annotation generation. The (a) part shows the public templates for generic data construction, while the (b) part demonstrates the customized task templates, $<$ Role $>$ and $<$ Format $>$, tailored to multi-source RS imagery. These templates embed object categories explicitly, ensuring the correctness and consistency of generated annotations.

Notably, category names and ground-truth bounding boxes are explicitly embedded in the prompts, ensuring the correctness of the generated annotations by guiding GPT-4V to accurately associate text with visual regions. Beyond correctness, the structured templates also guarantee annotation consistency across large-scale data, avoiding ambiguity and hallucination. To further strengthen the coupling between visual prompts and textual outputs, we adopt Set-of-Marks (SoM) prompting [77], which highlights multiple target objects simultaneously and reinforces spatial-textual grounding. This design enables GPT-4V to produce coherent and complete descriptions even in cluttered scenes. In conclusion, the integration of diverse multi-domain datasets through this rigorous pipeline allows joint training of EarthGPT-X on standardized supervision, enhancing its robustness across RS modalities and improving its generalization to complex real-world scenarios. Additionally, to further ensure reliability, we conduct RS expert inspection on sampled annotations, validating semantic accuracy, coherence, and completeness. This human-in-the-loop verification step complements automatic generation and safeguards the overall quality of the constructed dataset.

V. Experiment

A. Experimental Setups

1) Implementation Details: The one-stage all-in-one training is conducted in this paper. We use transformer-based decoder-only Llama 2-13B [78] as the basic LLM. The pre-trained DINOv2-ViT L/14 [54] and CLIP-ConvNeXt [53] are adopted as a hybrid image encoder MVE. The training image resolution is 448×448. Subsequently, the randomly initialized ViT model is seated as the spatial-aware encoder

to refine sparse features. Then, we employ a pre-trained SAM encoder and decoder [11] as the grounding encoder and pixel decoder. The image resolution required to be reshaped for the segmentation model is 1024×1024. Furthermore, we utilize the simple MLPs as the projection layer to realize multi-modal content dimension alignment.

Throughout the training, to keep training efficiency and preserve the generalization ability of backbone models, we only unfreeze the self-attention layers of LLM and the added cross-attention module for tuning. Other modules, including visual encoders, projection layers, etc., are kept frozen. The AdamW optimizer [79] with a weight decay of 0.01 and a learning rate of 2e-5 is utilized. The total training is conducted on 8 NVIDIA A100 GPUs with 80 GB memory, and the training time is around 310 hours.

2) Evaluation Metrics: We adopt three categories of evaluation metrics tailored to different visual tasks in our benchmark. *(i) Region and Image Captioning Tasks:* We employ standard language generation metrics, including BLEU [80], METEOR [81], ROUGE [82], CIDEr [83], and SPICE [84]. *(ii) Scene Classification:* We report classification accuracy, which directly reflects the model's ability to recognize high-level scene semantics in RS images. *(iii) Referring Object Classification:* We employ two semantic-aware metrics: Semantic Intersection over Union (S-IOU) and Semantic Similarity (SS) [3], [76], [85]. SS measures the similarity of predicted and ground-truth labels in a semantic embedding space, while S-IOU captures the semantic overlap between them by computing the intersection-over-union in the embedding space. These metrics provide a comprehensive assessment of the quality of the identified object information and generated captions. Consider the referring expression *"An airplane is on the right of a small vehicle"*, where the ground-truth target labels are *airplane* and *vehicle*. Suppose the model predicts *aircraft* and *car*, both outputs achieve high SS since they are semantically close to the labels in the embedding space, and also high S-IoU because of strong token-level overlap. In contrast, if the predictions are *ship* and *building*, SS drops due to weaker semantic similarity, while S-IoU becomes very low as the key object words differ significantly from the ground-truth labels.

B. Scene Classification

We evaluate EarthGPT-X on two RS benchmarks: AID [86] and UCMerced [87]. Following GeoChat [8], we use 20% of AID for testing, while the entire UCMerced serves as a zero-shot test set. To fit the visual prompting setting, we define a full-image box $[0, 0, \text{width}, \text{height}]$ and use the instruction: "Please identify the object category of each marked region in the image." Classification accuracy is computed under zero-shot settings. As shown in Tab. III, EarthGPT-X achieves 78.09% on AID and 87.89% on UCMerced, outperforming all baselines. This further confirms that the proposed visual prompting paradigm achieves stronger task alignment than directly applying vanilla MLLMs (e.g., Sphinx and GeoChat) without prompt adaptation.

TABLE III
ZERO-SHOT SCENE CLASSIFICATION (UCMERCED AND AID).

Models	Publications	UCMerced Acc	AID Acc
Qwen-VL [88]	arXiv 2023	62.90	52.60
MiniGPTV2 [89]	arXiv 2023	4.76	12.90
LLaVa-1.5 [28]	NeurIPS 2023	68.00	51.00
Sphinx [90]	ECCV 2024	62.76	58.20
InternVL [91]	CVPR 2024	58.23	60.40
GeoChat [8]	CVPR 2024	84.43	72.03
EarthMarker [14]	TGRS 2025	86.52	77.97
EarthGPT-X (Ours)		**87.89**	**78.09**

C. Referring Object Classification

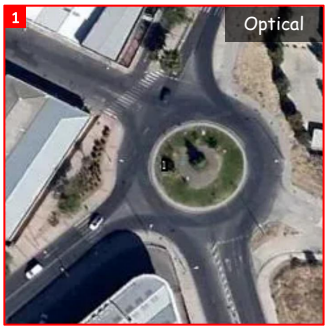
1) Evaluation on Optical Data: Referring object classification aims to identify the category labels of objects referred to by visual prompts and textual instructions. To assess the merits of the proposed EarthGPT-X, we report its performance and compare it with the recent vanilla and visual prompting MLLMs on DIOR-RSVG [2]. We adapt DIOR-RSVG by converting referring expressions and their bounding boxes into classification samples, where each target region serves as input and its category label as output. The text instruction is "Please identify the object category of the marked region in the image". The vanilla MLLMs do not support drawing visual prompts on the images, therefore, the coordinates contained in the text instruction serve as the referring regions. The vanilla MLLMs used for comparison include natural-domain models such as Sphinx [90], InternVL2.5 [37], InternVL3 [92], as well as RS-domain models including GeoChat [88] and EarthGPT [7]. For visual prompting MLLMs, we select the most recent representatives, namely the natural-domain models Sphinx-V [76] and Vip-LLava [17], together with the RS-domain model EarthMarker [14]. We conduct experiments under three levels of visual prompting (point, box, and free-form) to provide a fair and comprehensive comparison across models. Following [1], [76], for point prompts, we randomly select a point within the ground-truth object that is also close to its boundary. For box prompts, we use the ground-truth bounding box provided by DIOR-RSVG. For free-form prompts, we randomly perturb the center positions and zoom the box prompts to simulate. As shown in Tab. IV, EarthGPT-X achieves optimal results in S-IoU under box prompts. We also observe that the box-level results are better than the point and free-form prompt ones.

2) Evaluation on SAR Data: We evaluate EarthGPT-X and recent InternVL series [37], [92], Sphinx-V [76] on the MSAR [68] test set (Tab. V). EarthGPT-X, trained with SAR-specific samples, achieves superior performance (SS: 99.85%, S-IoU: 97.75%), showing strong capacity in SAR-oriented understanding. In contrast, natural MLLMs InternVL2.5 and InternVL3, without SAR training, perform significantly lower. The superior results of EarthGPT-X illustrate its stronger modality adaptation and specialization capabilities, which are crucial for accurate SAR imagery understanding in the RS applications.

TABLE IV
REFERRING OBJECT CLASSIFICATION (DIOR-RSVG).

Models	Publications	Point		Box *(Coor)*		Free-form	
		SS	S-IoU	SS	S-IoU	SS	S-IoU
Vanilla MLLMs							
InternVL2.5 [37]	arXiv 2024	—	—	54.54	37.91	—	—
InternVL3 [92]	arXiv 2025	—	—	50.84	26.99	—	—
GeoChat [8]	CVPR 2024	—	—	79.59	68.80	—	—
Sphinx [90]	ECCV 2024	—	—	93.72	89.37	—	—
EarthGPT [7]	TGRS 2024	—	—	94.64	90.16	—	—
Visual Prompting MLLMs							
Vip-LLava-7b [17]	CVPR 2024	68.32	51.98	72.56	55.94	71.24	53.99
Vip-LLava-13b [17]	CVPR 2024	70.02	57.61	74.51	60.53	74.05	59.83
Sphinx -V [76]	ICLR 2025	86.21	78.78	89.07	81.62	88.46	79.97
EarthMarker [14]	TGRS 2025	95.96	93.49	98.37	97.24	—	—
EarthGPT-X (Ours)		**96.33**	**94.42**	**98.79**	**98.03**	**98.71**	**96.72**

Notes: Since vanilla MLLMs lack visual prompt support, textual coordinates (abbr. as Coor) are treated as the referring regions. '—' means no such capability. The underline denotes the best performance achieved under box prompts.



Brief Captioning

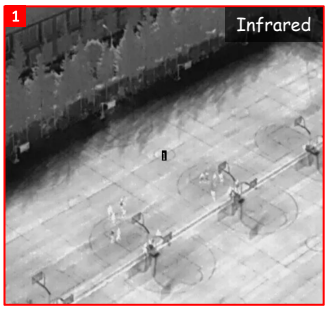
User: Please provide a brief caption of each marked region in the image.

EarthGPT-X:
<Region 1>: It seems that this region almost covers the entire image. There are several moving vehicles on the road at the roundabout, with some trees and a large lawn around the roundabout.

Detailed Captioning

User: Please provide a detailed caption of each marked region in the image.

EarthGPT-X: <Region 1>: It seems that this region almost covers the entire image. This is an image of roundabout. The image shows a large roundabout in the middle of a city intersection. The roundabout is surrounded by a gray overpass, and there are several small vehicles on the road, including a car on the top side of the overpass, and two other cars on the bottom side. The vehicles are positioned at various distances from the overpass, indicating that they are either approaching or leaving the intersection.



Brief Captioning

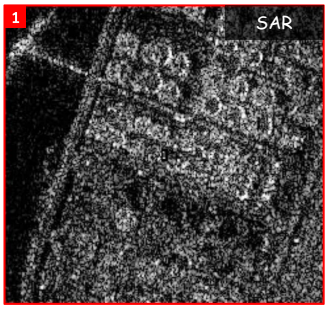
User: Please provide a brief caption of each marked region in the image.

EarthGPT-X:
<Region 1>: It seems that this region almost covers the entire image. This is an image of Basketball court. The image is a black and white photo of a basketball court.

Detailed Captioning

User: Please provide a detailed caption of each marked region in the image.

EarthGPT-X:
<Region 1>: It seems that this region almost covers the entire image. This is an image of Basketball court. The court is large and has a rectangular shape. There are several basketball hoops visible on the court, with some located closer to the top and others near the bottom. The hoops are spread out across the court, providing a variety of playing options for the players.



Brief Captioning

User: Please provide a brief caption of each marked region in the image.

EarthGPT-X:
<Region 1>: It seems that this region almost covers the entire image. This is an image of tank.

Detailed Captioning

User: Please provide a detailed caption of each marked region in the image.

EarthGPT-X:
<Region 1>: It seems that this region almost covers the entire image. This is an image of tank. This area is densely packed with large storage tanks, likely for oil or chemicals, given its proximity to the shore.

Fig. 5. Image brief and detailed captioning results on optical, infrared, and SAR modalities, demonstrating the superior scene-level visual understanding ability of EarthGPT-X.

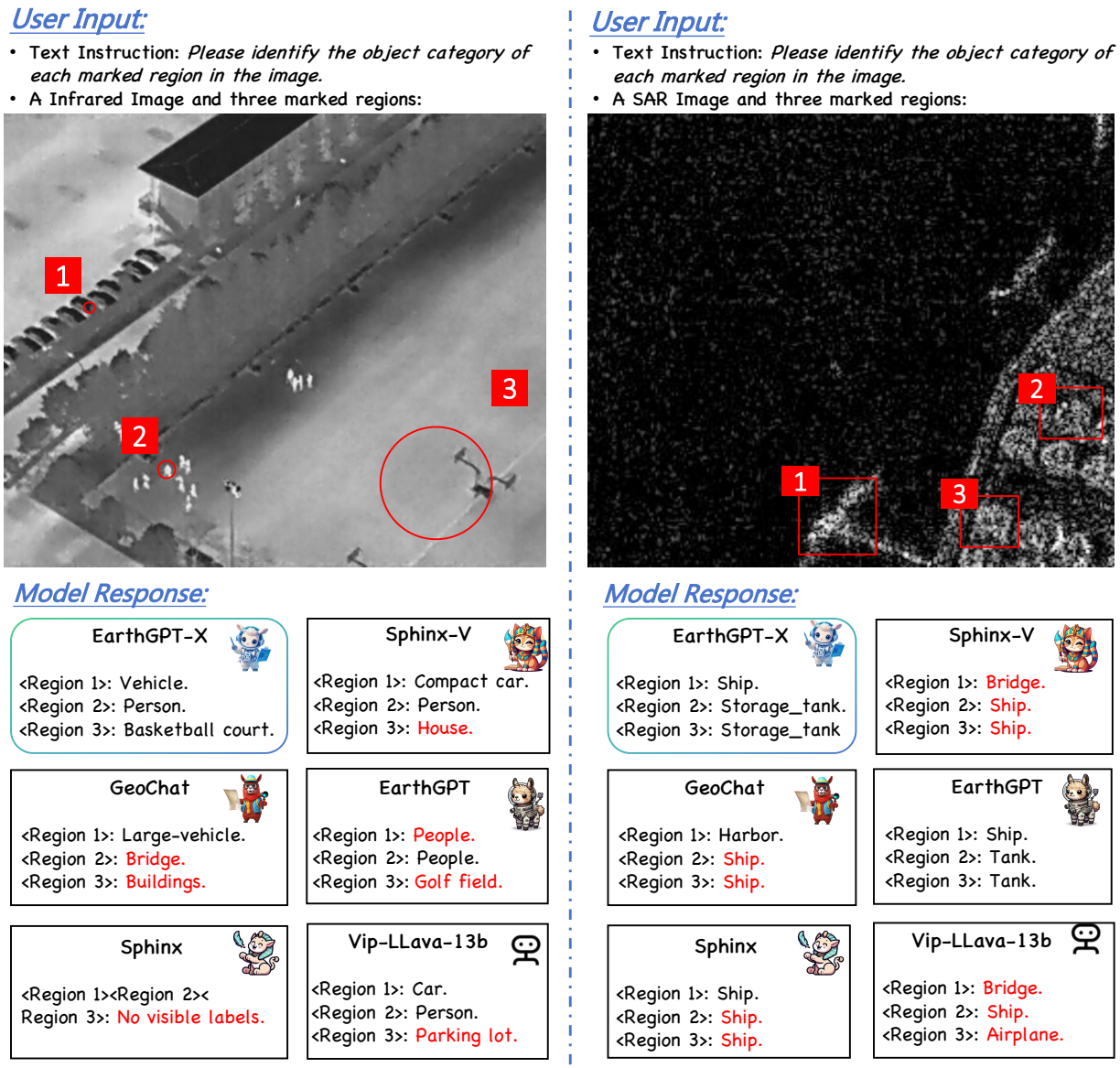


Fig. 6. The referring object classification results on SAR and infrared RS images demonstrate the superior object-level RS visual understanding capability of EarthGPT-X, compared to other vanilla and visual prompting MLLMs (*The black text indicates consistency with the ground truth, the red text represents incorrect answers*).

D. Image Captioning

To evaluate the image captioning capabilities, we adopt the NWPU-Captions [95] dataset to benchmark EarthGPT-X against other expert models under supervised settings. During evaluation, we use the instruction "Please provide a brief caption of each marked region in the image" alongside a full-image box $[0, 0, \text{width}, \text{height}]$ as the visual prompt to guide the captioning process. As shown in Table VI, EarthGPT-X outperforms most expert-designed models but remains inferior to task-specific architectures such as Aware-Transformer [97]. We note that the performance gap in image captioning is a common limitation of MLLMs, which aim to balance multiple capabilities rather than optimizing for a single task.

TABLE V
REFERRING OBJECT CLASSIFICATION (MSAR).

Models	Publications	Formats	SS	S-IoU
InternVL2.5 [37]	arXiv 2024	Coor	37.05	7.69
InternVL3 [92]	arXiv 2025	Coor	32.72	2.89
Sphinx-V [76]	ICLR 2025	Box	34.76	2.03
EarthGPT-X (Ours)		Box	**99.85**	**97.75**

E. Region Captioning

1) Evaluation on Optical Data: For the brief region captioning on optical data, the test set of OPT-RSVG [101] is employed for zero-shot evaluation. We elaborately choose both natural and RS-specific MLLMs for performance comparison. Natural scene MLLMs contain Qwen-VL [8], InternVL2.5 [37], InternVL3 [92], Sphinx [90], Sphinx-V [76], Vip-LLava [17] and GLaMM [17]. Popular RS-specific MLLMs

GeoChat [88] and EarthGPT [7] are selected. The text instruction for the region captioning task is designed as "Please provide a brief caption of each marked region in the image". As shown in Tab. VII, the proposed EarthGPT-X consistently outperforms previous works across all evaluation metrics. Notably, EarthGPT-X achieves significant gains over Sphinx-V, despite the latter being trained on large-scale RS datasets. This highlights the superior performance and stronger generalization ability of EarthGPT-X in multi-source RS data understanding. Notably, using boxes as visual prompts can yield the best performance on region captioning ability.

2) Evaluation on SAR and Infrared Data: As displayed in Tab. VIII, for the multi-source brief captioning task, we evaluate EarthGPT-X on two representative datasets: SAR-Ship [67] for supervised evaluation and HIT UAV [102] under a zero-shot setting. These datasets enable us to validate the generalization ability of EarthGPT-X across different modalities. For comparison, we include two representative general-purpose MLLMs, GPT-4o from OpenAI and InternVL2 [91] from OpenGVLab. We further reference the first RS MLLM, GeoChat [8], and the latest multi-source understanding model, EarthDial [22], as strong domain-related baselines. Experimental results demonstrate that EarthGPT-X achieves superior overall performance on multi-source datasets.

F. Ablation Study

1) Different Visual Encoders: To show the effectiveness of the proposed hybrid MVE module, we ablate on the performance of our EarthGPT-X when using different visual encoders. Specifically, this ablation uses CLIP-ConvNeXt-L, CLIP RN50×16, and DINOv2 ViT-L/14 for comparison. The results are shown in Tab. IX, it is clear that the setting using the CLIP-ConvNeXt-L and DINOv2 ViT-L/14 as the mixed visual experts achieves the most superior performance on scene classification, region captioning, and referring object classification tasks. These results demonstrate that utilizing a

TABLE VI
SUPERVISED IMAGE CAPTIONING (NWPU-CAPTIONS).

Models	Publications	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE -L	CIDEr(0-5)	SPICE
RS Expert Models									
CSMLF [93]	GRSL 2019	77.0	64.9	53.2	47.1	32.0	57.8	106.5	26.5
Qu *et al.* [59]	CITS 2016	72.5	60.3	51.8	45.5	33.6	59.1	117.9	27.6
Attention (soft) [58]	TGRS 2017	73.1	60.9	52.5	46.2	33.9	59.9	113.6	28.5
Attention (hard) [58]	TGRS 2017	73.3	61.0	52.7	46.4	34.0	60.0	110.3	28.4
FC-Att [94]	TGRS 2019	73.6	61.5	53.2	46.9	33.8	60.0	123.1	28.3
SM-Att [94]	TGRS 2019	73.9	61.7	53.2	46.8	33.0	59.3	123.6	27.6
MLCA-Net [95]	TGRS 2022	74.5	62.4	54.1	47.8	33.7	60.1	126.4	28.5
Capformer [96]	IGARSS 2022	89.0	81.1	74.5	69.2	45.4	78.4	201.9	–
Aware-Transformer [97]	CGI 2023	**91.5**	**85.2**	**79.6**	**75.0**	**47.7**	**81.1**	**214.7**	–
RS MLLMs									
RS-CapRet [98]	arXiv 2024	87.1	78.7	71.7	65.6	43.6	77.6	192.9	31.1
EarthGPT [7]	TGRS 2024	87.1	78.7	71.6	65.5	44.5	78.2	192.6	**32.2**
EarthMarker [14]	TGRS 2025	84.4	73.1	62.9	54.3	37.5	70.0	162.9	26.8
EarthGPT-X (Ours)		85.2	74.2	64.0	55.9	39.5	71.5	163.7	27.3

TABLE VII
ZERO-SHOT REGION CAPTIONING (OPT-RSVG).

Models	Publications	Format	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Vanilla MLLMs										
Qwen-VL [88]	arXiv 2023	Coor	12.78	6.47	2.77	1.24	5.97	17.06	13.15	6.60
InternVL2.5 [37]	arXiv 2024	Coor	9.87	4.18	1.69	0.52	5.71	10.55	9.22	8.28
InternVL3 [92]	arXiv 2025	Coor	15.42	8.91	6.34	4.38	18.64	14.28	13.27	14.87
GeoChat [8]	CVPR 2024	Coor	15.02	8.49	5.21	3.29	9.24	25.65	42.86	16.10
Sphinx [90]	ECCV 2024	Box	22.79	15.27	10.97	7.84	12.95	30.28	105.85	20.18
EarthGPT [7]	TGRS 2024	Box	30.28	19.99	13.58	8.73	15.29	31.52	88.92	21.05
Visual Prompting MLLMs										
Sphinx-V [76]	ICLR 2025	Box	23.35	14.84	10.37	7.16	14.10	32.08	81.79	22.46
Vip-LLava-7b [17]	CVPR 2024	Box	14.67	7.29	4.06	2.25	8.20	20.17	23.47	12.79
Vip-LLava-13b [17]	CVPR 2024	Box	16.39	8.40	4.77	2.80	8.47	21.36	25.85	12.41
GLaMM [99]	CVPR 2024	Box	12.50	6.06	3.25	1.74	7.98	22.01	23.85	12.40
EarthGPT-X (Ours)		Point	**52.57**	**48.14**	**45.72**	**43.38**	**35.28**	**64.63**	**479.92**	**63.47**
EarthGPT-X (Ours)		Box	**64.83**	**59.91**	**56.28**	**53.64**	**39.61**	**66.87**	**488.76**	**64.32**
EarthGPT-X (Ours)		Free-form	**63.38**	**58.62**	**55.12**	**52.41**	**38.72**	**65.43**	**483.63**	**63.58**

Note: The underline denotes the best zero-shot performance achieved under box prompts.

TABLE VIII
REGION CAPTIONING ON SAR AND INFRARED DATASETS.

Models	Publications	SAR-Ship (Supervised)			HIT UAV (Zero-Shot)		
		ROUGE-1	ROUGE-L	METEOR	ROUGE-1	ROUGE-L	METEOR
Generic MLLMs							
GPT-4o [100]	OpenAI 2024	7.49	7.24	7.07	10.96	9.02	8.23
InternVL2 [91]	OpenGV 2024	9.67	8.67	8.19	11.00	9.53	8.40
Specialized RS MLLMs							
GeoChat [8]	CVPR 2024	57.15	57.15	52.20	59.85	59.85	51.31
EarthDial [22]	CVPR 2025	63.10	63.10	54.83	61.83	61.83	**52.80**
EarthGPT-X (Ours)		**65.25**	**65.25**	**55.90**	**64.40**	**64.40**	51.56

TABLE IX
ABLATION RESULTS ABOUT THE DESIGN OF THE VISUAL ENCODERS.

CNN Model		ViT Model	Classification	Region Captioning	Ref. Obj. Classification
CLIP ConvNeXt-L	CLIP RN50×16	DINOv2 ViT-L/14	AID (Acc)	OPT-RSVG (CIDEr)	DIOR-RSVG (SIOU)
✓	×	×	73.65	471.02	94.11
×	✓	×	74.26	471.35	94.68
×	×	✓	73.05	479.12	96.07
×	✓	✓	75.51	480.63	96.92
✓	×	✓	**78.09**	**488.76**	**98.03**

Notes: Results for region-level tasks are reported under box-prompt settings.

network architecture with complementary functions has deepened the understanding of local information and contextual dependencies on RS imagery.

2) Sub-images Decomposition and Multi-scale: In the visual feature extraction stage, we strengthen feature integration by leveraging multi-scale features and high-resolution sub-images for richer RS image comprehension. Ablation results in Tab. X show that this design significantly outperforms single-scale settings under the referring object classification task, confirming that capturing finer details through multiple scales and sub-images is crucial for accurate interpretation.

G. Computation Analysis and Comparisons

In this section, we compare the GPU memory usage (MiB) and inference time (seconds) of EarthGPT-X with other MLLMs and visual prompting models. All evaluations are conducted on the region captioning task using an NVIDIA RTX A6000. For both MLLMs and visual prompting models, GPU memory usage is primarily attributed to model weight storage. As illustrated in Tab. XI, EarthGPT-X consumes more GPU memory than Sphinx [76], GeoChat [8], EarthGPT [7], Qwen-VL [88], and Vip-LLava-13B [17], but less than Sphinx-V [76]. Similarly, its inference time is longer than all other models except Sphinx-V, due to its larger parameter scale. However, the time difference remains within one second, indicating that the performance gap is marginal and does not affect deployment in practical applications, as EarthGPT-X remains well within the capacity of commonly used GPUs. In future work, we plan to optimize EarthGPT-X's inference efficiency to better balance speed and interpretability.

TABLE X
ABLATION RESULTS ON IMAGE PRE-PROCESSING.

Settings	Visual Encoder	SS	S-IoU
Single-scale	MVE	94.93	95.05
Multi-scale	MVE	96.18	95.29
Multi-scale + Sub-images	MVE	**98.79**	**98.03**

Notes: Results are reported under box-prompt settings.

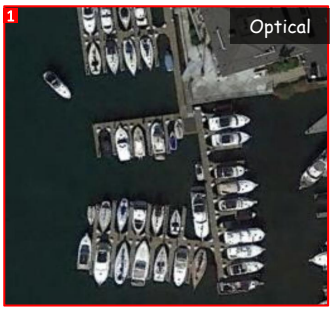
TABLE XI
COMPUTATION ANALYSIS AND COMPARISONS

Models	Inference Time(s)	GPU Memory(MiB)
MLLMs		
Sphinx	1.08	41853
EarthGPT	0.79	41766
GeoChat	0.74	15544
Qwen-VL	0.35	38145
Visual Prompting Models		
Sphinx-V	1.95	41794
Vip-LLava-13B	0.84	27748
EarthGPT-X(Ours)	1.60	41006

VI. VISUALIZATION

A. Multi-source Image Captioning

We present visualization results of EarthGPT-X on optical, SAR, and infrared images to demonstrate its capability in understanding diverse RS scenes. Given an entire



Brief Captioning (Image)

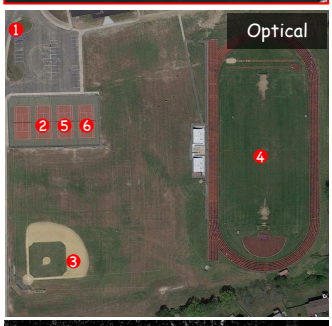
User: Please provide a brief caption of each marked region in the image.

EarthGPT-X: It seems that this region almost covers the entire image. There are many ships in the harbor.

Detailed Captioning (Image)

User: Please provide a detailed caption of each marked region in the image.

EarthGPT-X: It seems that this region almost covers the entire image. This is an image of harbor. The image shows a harbor with roughly 34 ships docked. The ships are of various sizes and are arranged in multiple rows, filling the harbor. Some ships are positioned closer to the top, while others are situated towards the bottom.....



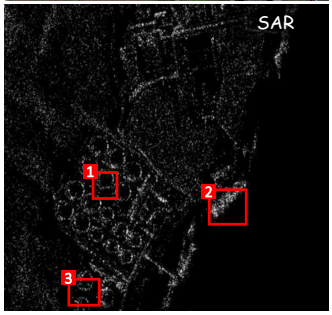
Detailed Captioning (Point)

User: Please provide a detailed caption of each marked point in the image.
EarthGPT-X: There are 1 ground-trackfield, 1 baseballfield, 4 tennis-courts in this remote sensing image. The ground track field is on the right of the large baseball field on the lower left.

Inter-relationship Analysis (Point)

User: Please analyze the relationship between all marked points in the image.

EarthGPT-X: <Region 2><Region 3><Region 4><Region5><Region 6>: The marked regions represent different sports facilities, each serving a specific sport. <Region 1> is a vehicle that may be used for transportation or maintenance, while <Region 2>, <Region 5>, <Region 6> are tennis courts, each with a similar design. <Region 3> is a baseball field, which is a different sport entirely...



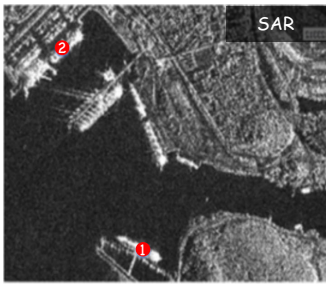
Referring Classification (Box)

User: Please identify the object category of each marked region in the image.

EarthGPT-X:
<Region 1>: Tank.
<Region 2>: Ship.
<Region 3>: Tank.

Inter-relationship Analysis (Box)

User: Please analyze the relationship between all marked points in the image.
EarthGPT-X: <Region 1><Region 2><Region 3>: The marked regions represent different elements within a larger scene, likely a maritime or industrial area. <Region 1> shows a tank, which is a common feature in such environments for storing liquids or gases. <Region 2> depicts a ship, which is a common mode of transportation for goods or people, and its proximity to the tank suggests it may be involved in the operations or logistics related to the tank. <Region 3> shows another tank, which, like <Region 1>...



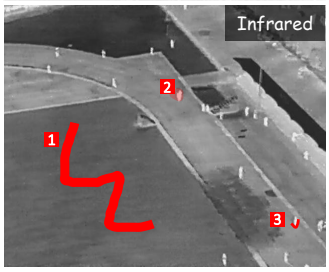
Referring Classification (Point)

User: Please identify the object category of each marked point in the image.

EarthGPT-X:
<Mark 1>: Ship.
<Mark 2>: Ship.

Inter-relationship Analysis (Point)

User: Please analyze the relationship between all marked points in the image.
EarthGPT-X:
<Mark 1><Mark 2>: Both marked areas are part of a larger maritime scene. <Mark 1> is a ship that is likely to be the main vessel, while <Mark 2> is a smaller boat that could be a tender or a service vessel. They are both situated in a harbor area, indicating they are part of the same maritime environment. The proximity of the two suggests they....



Summarized Captioning (Free-form)

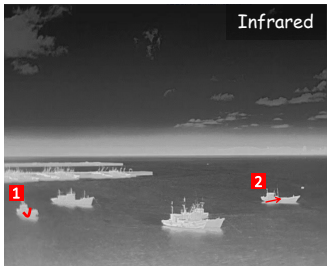
User: Please provide a summarized caption based on all the marked regions in the image.

EarthGPT-X:
In the image, there are many people including <Region 2> and <Region 3> positioned close to each other in the ground track field <Region 1>.

Region Captioning (Free-form)

User: Please provide a brief caption of each marked region in the image.

EarthGPT-X:
<Region 1>: The large ground track field is on the left of many people on the right.
<Region 2>: The person on the right.
<Region 3>: The person on the lower right.



Referring Classification (Free-form)

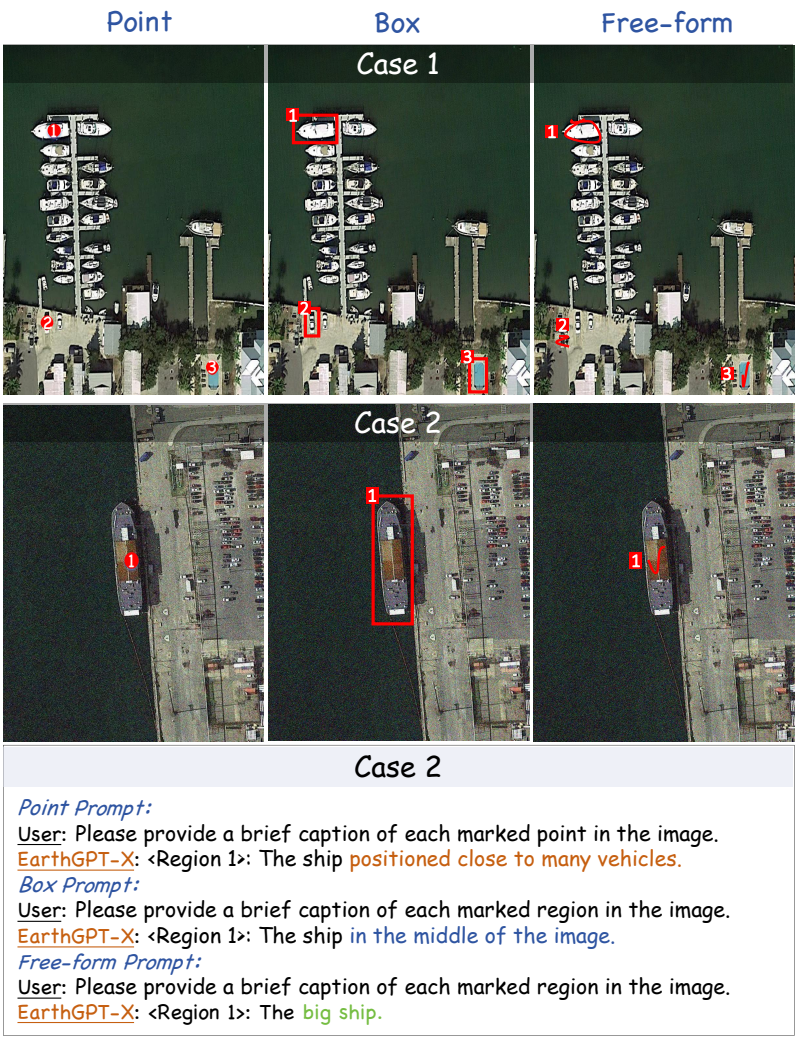
User: Please identify the object category of each marked region in the image.
EarthGPT-X:

<Region 1>: Fishing_boat.
<Region 2>: Fishing_boat.

Inter-relationship Analysis (Free-form)

User: Please analyze the relationship between all marked points in the image.
EarthGPT-X: <Mark 1><Mark 2>: Both marked regions represent fishing boats, likely part of the same fleet given their similar design. They are situated in close proximity to each other, which suggests they may be in the same fishing area or operation. The boats are positioned in a way that suggests they are in a calm part of the water, possibly near the shore or in a harbor.

Fig. 7. Qualitative examples of EarthGPT-X performing multi-task reasoning and comprehending multi-source RS imagery (optical, SAR, and infrared) under various visual prompts.



Case 2

Point Prompt:
User: Please provide a brief caption of each marked point in the image.
EarthGPT-X: <Region 1>: The ship positioned close to many vehicles.
Box Prompt:
User: Please provide a brief caption of each marked region in the image.
EarthGPT-X: <Region 1>: The ship in the middle of the image.
Free-form Prompt:
User: Please provide a brief caption of each marked region in the image.
EarthGPT-X: <Region 1>: The big ship.

Case 1

Point Prompt:
User: Please analyze the relationship between all marked points in the image.
EarthGPT-X:
<Region 1><Region 2><Region 3>: The three marked regions, while individually described as a marina, a vehicle, and a swimming pool, all contribute to a waterfront setting. The marina is a functional structure for boats, the vehicle is likely used for land transport, and the swimming pool is a recreational feature for the area. They share a common environment, being near the water, which is a typical setting for such facilities.

Box Prompt:
User: Please analyze the relationship between all marked regions in the image.
EarthGPT-X: <Region 1><Region 2><Region 3>: The three marked regions, while individually described as a marina, a vehicle, and a swimming pool, all contribute to a common theme of a waterfront or waterfront-related environment. They share a common setting, with the marina being on the water, the vehicle being on the land, and the pool being adjacent to the water. Each region represents a different aspect of the waterfront lifestyle: boating, land transportation, and leisure activities.

Free-form Prompt:
User: Please analyze the relationship between all marked region in the image.
EarthGPT-X: <Region 1><Region 2><Region 3>: The three marked regions are all part of the marina's environment. <Region 1> shows the docking structure, <Region 2> shows the vehicles near the dock, and <Region 3> shows the surrounding landscape that the marina is built near.

Fig. 8. Illustration of multiple visual prompting strategies. Specifically, *Point Prompt* highlights local functional roles of individual regions, *Box Prompt* emphasizes global thematic coherence and spatial relations, and *Free-form Prompt* provides concise contextual abstraction. Despite these differences, all three prompting types guide the model to establish precise region–text alignment.

image as the region of interest, we use the visual prompt $[0, 0, \text{width}, \text{height}]$ and provide instructions such as "Please provide a brief caption of each marked region in the image. " or "Please provide a detailed caption of each marked region in the image." As shown in Fig. 5, the model generates accurate and semantically rich captions across modalities, effectively handling varied visual characteristics and noise patterns.

B. Referring Object Classification

We conduct qualitative analysis to provide an intuitive performance display of EarthGPT-X. As shown in Fig. 6, the infrared image showcases a residential leisure area, and for the three marked regions, all other MLLMs respond incorrectly. Conversely, EarthGPT-X successfully recognizes the human, vehicle, and basketball court targets. Following this is a fuzzy port SAR image, where it is difficult to visually distinguish between ships and storage tanks. EarthGPT-X delivers precise interpretations for these targets, whereas other models perform poorly, except for EarthGPT. These results prove the significance of our work that empowers the MLLMs with fine-grained reasoning ability for object-level comprehension.

C. Multi-level Multi-task Reasoning

Traditional vision models in the RS field primarily focus on a single visual modality and single tasks, exhibiting poor generalization [103]–[105]. Thanks to our M-RSVP's rich multi-sensor knowledge, EarthGPT-X has robust generalization capabilities across optical, SAR, and infrared visual modalities. To verify and assess the comprehensive capabilities of the proposed model, we conduct a thorough qualitative evaluation. Our assessment also covers multi-source images and multiple levels of image interpretation granularity, including image, point, box, and free-form prompted levels. Moreover, the evaluated visual tasks are extensive, such as brief captioning, detailed captioning, ground captioning, region captioning, referring object classification, and inter-relationship analyses. As displayed in Fig. 7, our model demonstrates remarkable visual reasoning and analytical abilities. EarthGPT-X provides detailed descriptions for multi-source images, as evidenced by

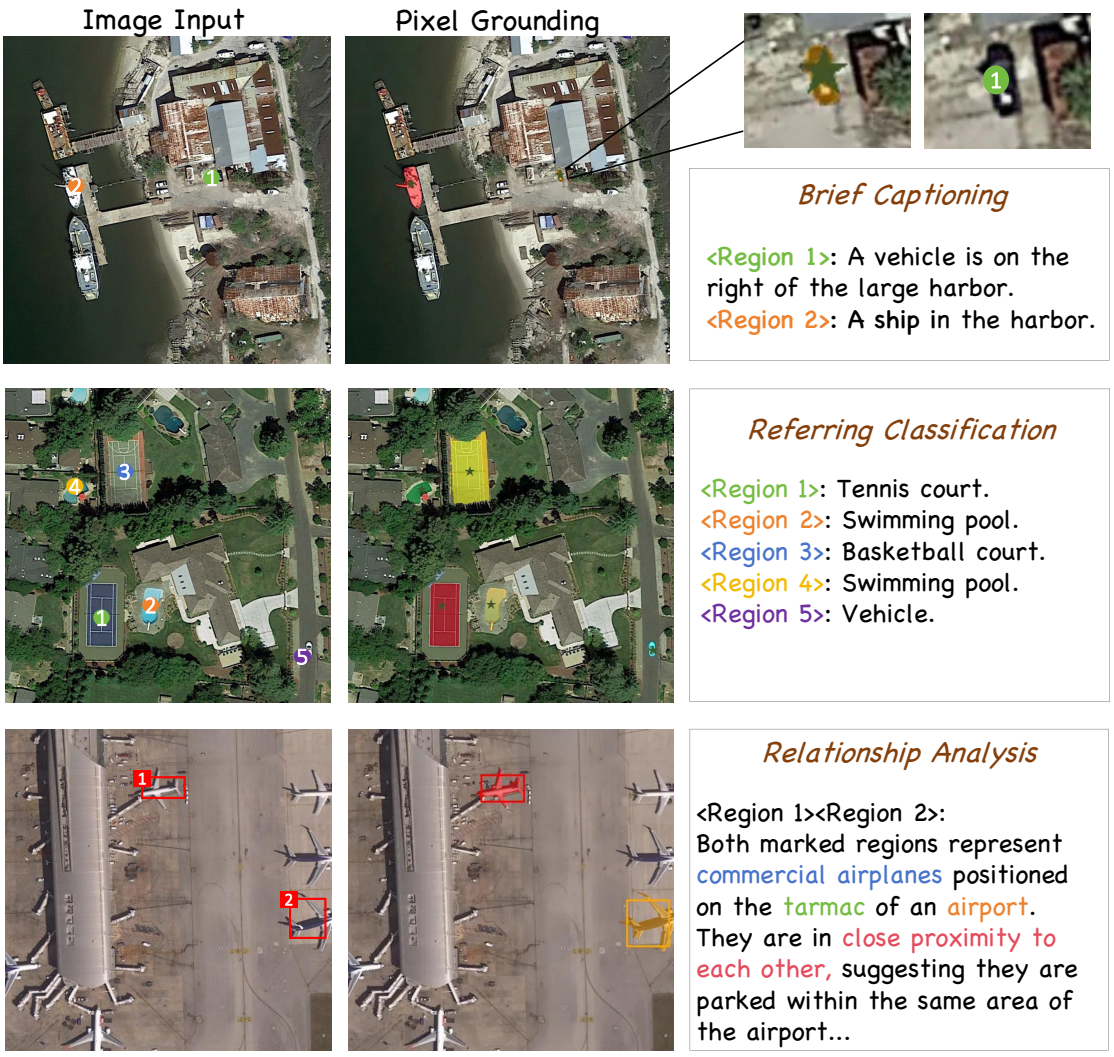


Fig. 9. Examples of EarthGPT-X performing visual reasoning and pixel-level grounding on optical imagery across diverse scenes (harbor, urban stadium, and airport). ***Better to zoom in to view clearly***.

the interpretation of the first image, which includes overall scene depiction, target counts, orientations, and higher-level reasoning. In the relationship analysis of the second and third examples, the model generates comprehensive and structured responses, effectively capturing both local and global semantics. Notably, even for low-quality SAR and infrared maritime images, EarthGPT-X can infer relative spatial relations, target sizes, vessel functionalities, and operational states (e.g., docking at a port). In summary, EarthGPT-X demonstrates a profound ability to dissect and understand complex visual information in the multi-source RS MLLMs.

D. Multiple Visual Prompts

To evaluate the adaptability of EarthGPT-X under different prompting shapes, we investigate point prompt, box prompt, and free-form prompt. These prompts vary in the way visual regions are highlighted to the model, thereby influencing the style and granularity of responses. ***i) Similarities*:** Across all prompt types, EarthGPT-X consistently grounds the responses to the marked regions, ensuring semantic coherence and correct alignment with the visual input. ***ii) Differences***: As shown in Fig. 8, under the point prompt, EarthGPT-X generates a caption emphasizing the spatial relation, which reflects its local context-awareness. Under the box prompt, the model highlights the absolute spatial location, showing a region-level grounding behavior. In contrast, the free-form prompt yields a more generic description, focusing on object-level attributes (*Note that these examples are qualitative illustrations, not unique or deterministic results, since different prompt placements may lead to varied outputs*).

E. Zero-shot Pixel-level Grounding and Reasoning

All pixel grounding and reasoning experiments are conducted in a zero-shot setting, and the capability is obtained without any task-specific training.

1) Different Scenes: To evaluate the robustness under complex scenes, we select RS images with multiple diverse targets. As shown in Fig. 9, in the port scene, we mark a ship and a tiny vehicle to examine the model's ability to localize small objects within a cluttered background. EarthGPT-X not only successfully detects these targets but also separates them

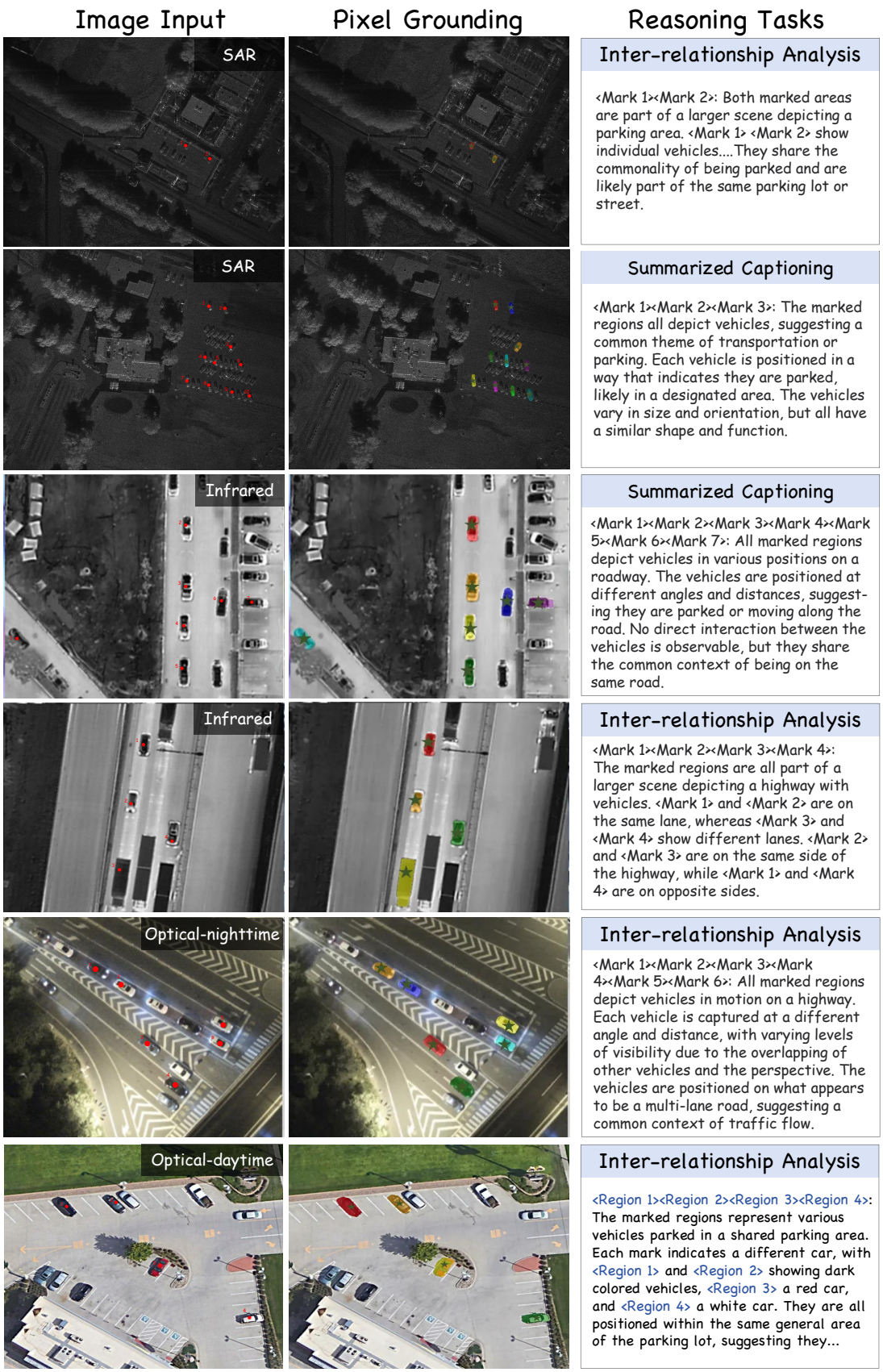


Fig. 10. Examples of EarthGPT-X performing small vehicle recognition and reasoning with pixel-level grounding on multi-source imagery (i.e., SAR, infrared, optical–daytime, and optical–nighttime). ***Better to zoom in to view clearly.***

from the surroundings. In another urban sports center scene, five spatially separated objects (e.g., swimming pool, tennis court, and vehicle) are interpreted. EarthGPT-X demonstrates exceptional performance in precise object recognition and pixel-grounding.

2) Different Image Modalities: As shown in Fig. 10, we further evaluate on small vehicle targets across heterogeneous image modalities, including optical, infrared, and SAR. The test data is from VisDrone-DroneVehicle [106] and SAR_Vehicle_Detection_Dataset [107]. Despite the significant domain gaps among these modalities, EarthGPT-X consistently identifies and segments small and tiny vehicles, showing its strong generalization and adaptability to heterogeneous RS image types.

VII. Conclusion

In this paper, we propose EarthGPT-X, a flexible and versatile MLLM specifically designed for understanding RS imagery from multiple sources. Technically, the hybrid signal mutual understanding mechanism is developed to enhance the interplay between dense visual features, sparse free-form visual prompts, and text instructions, facilitating a multi-grained understanding of complex RS imagery. Furthermore, a novel cross-domain one-stage fusion training strategy is proposed to generalize general-domain knowledge into RS multi-source scenarios. Through the designed visual prompt learning framework and training strategy, EarthGPT-X achieves multi-level visual understanding at the image, region, and pixel levels, etc., allowing for precise and intelligent analyses in real-world applications. In the future, we plan to enhance the image captioning capability of EarthGPT-X through task-aware fine-tuning and adaptive prompt optimization. Furthermore, we aim to expand EarthGPT-X towards a wider range of tasks and more advanced visual reasoning capabilities across diverse modalities, ultimately strengthening its versatility and effectiveness in comprehensive RS imagery analysis.

References

[1] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[2] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[3] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023.

[4] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

[5] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*, 2023.

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[7] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[8] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *CVPR*, pages 27831–27840, 2024.

[9] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.

[10] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[12] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024.

[13] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[14] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, Jun Li, and Xuerui Mao. Earthmarker: A visual prompting multi-modal large language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[15] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[16] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European conference on computer vision*, pages 52–70. Springer, 2024.

[17] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *CVPR*, pages 12914–12923, 2024.

[18] Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.

[19] Wei Zhang, Miaoxin Cai, Tong Zhang, Guoqiang Lei, Yin Zhuang, and Xuerui Mao. Popeye: A unified visual-language model for multi-source ship detection from remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. early access.

[20] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024.

[21] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6381–6388, 2025.

[22] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14303–14313, 2025.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[24] OpenAI. Chatgpt. https://chat.openai.com, 2023a.

[25] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[26] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *CVPR*, pages 18030–18040, 2022.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 36:34892–34916, 2023.

[29] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[30] Dongyang Liu, Renrui Zhang, Longtian Qiu, et al. SPHINX-x: Scaling data and parameters for a family of multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32400–32420. PMLR, 21–27 Jul 2024.

[31] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[32] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[33] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[34] Qwen Team. Qwen2.5-vl, January 2025.

[35] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[36] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[37] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model. In *Findings of ACL*, 2025.

[38] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

[39] Jie Zhou, Yongxiang Liu, Bowen Peng, Li Liu, and Xiang Li. Madinet: Mamba diffusion network for sar target detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[40] Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024.

[41] Xinyu Wang, Liming Yuan, Haixia Xu, and Xianbin Wen. Csds: End-to-end aerial scenes classification with depthwise separable convolution and an attention mechanism. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10484–10499, 2021.

[42] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019.

[43] Mingjin Zhang, Yuchun Wang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. Irsam: Advancing segment anything model for infrared small target detection. In *European Conference on Computer Vision*, pages 233–249. Springer, 2024.

[44] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023.

[45] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[46] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[47] Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. Geopix: A multimodal large language model for pixel-level image understanding in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[48] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

[49] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[51] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models, 2022.

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[55] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[57] Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. In *CVPR*, pages 13980–13990, 2024.

[58] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[59] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *2016 International conference on computer, information and telecommunication systems (Cits)*, pages 1–5. IEEE, 2016.

[60] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11):2321–2325, 2015.

[61] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[62] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36:8815–8827, 2023.

[63] SUN Xian, WANG Zhirui, SUN Yuanrui, DIAO Wenhui, ZHANG Yue, and FU Kun. Air-sarship-1.0: High-resolution sar ship detection dataset. *Journal of Radars*, 8(6):852–863, 2019.

[64] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021.

[65] Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *arXiv preprint arXiv:2403.06534*, 2024.

[66] Wang Zhirui, Kang Yuzhuo, Zeng Xuan, WANG Yuelei, ZHANG Ting, and SUN Xian. Sar-aircraft-1.0: High-resolution sar aircraft detection and recognition dataset. *Journal of Radars*, 12(4):906–922, 2023.

[67] Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. A sar dataset of ship detection for deep learning under complex backgrounds. *Remote Sensing*, 11(7), 2019.

[68] J Chen, Z Huang, R Xia, B Wu, L Sheng, L Sun, and B Yao. Large-scale multi-class sar image target detection dataset-1.0. j. radars 2022.

[69] InfiRay. Sea-shipping. http://openai.iraytek.com/apply/Sea_shipping.html/, 2021.

[70] InfiRay. Infrared-security. http://openai.iraytek.com/apply/Infrared_security.html/, 2021.

[71] InfiRay. Aerial-mancar. http://openai.raytrontek.com/apply/Aerial_mancar.html/, 2021.

[72] InfiRay. Double-light-vehicle. http://openai.raytrontek.com/apply/Double_light_vehicle.html/, 2021.

[73] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[74] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[75] OpenAI. Gpt-4 technical report, 2023.

[76] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[77] Jianwei Yang et al. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[78] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

[79] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[80] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[81] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[82] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[83] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[84] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[85] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019.

[86] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

[87] Yi Yang and Shawn Newsam. 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*, pages 270–279, 2010.

[88] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[89] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[90] Ziyi Lin, Dongyang Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Wenqi Shao, Keqin Chen, Jiaming Han, et al. Sphinx: A mixer of weights, visual embeddings and image scales for multi-modal large language models. In *European Conference on Computer Vision*, pages 36–55. Springer, 2024.

[91] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, July 2024. [Online]. Available:https://internvl.github.io/blog/2024-07-02-InternVL-2.0.

[92] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[93] Binqiang Wang, Xiaoqiang Lu, Xiangtao Zheng, and Xuelong Li. Semantic descriptions of high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1274–1278, 2019.

[94] Xiangrong Zhang, Xin Wang, Xu Tang, Huiyu Zhou, and Chen Li. Description generation for remote sensing images using attribute attention mechanism. *Remote Sensing*, 11(6):612, 2019.

[95] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

[96] Junjue Wang, Zihang Chen, Ailong Ma, and Yanfei Zhong. Capformer: Pure transformer for remote sensing image caption. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 7996–7999. IEEE, 2022.

[97] Yukun Cao, Jialuo Yan, Yijia Tang, Zhenyi He, Kangle Xu, and Yu Cheng. Aware-transformer: A novel pure transformer-based model for remote sensing image captioning. In *Computer Graphics International Conference*, pages 105–117. Springer, 2023.

[98] João Daniel Silva, João Magalhães, Devis Tuia, and Bruno Martins. Large language models for captioning and retrieving remote sensing images, 2024.

[99] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[100] OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/, August 2024. Accessed: 2025-09-17.

[101] Ke Li, Di Wang, Haojie Xu, Haodi Zhong, and Cong Wang. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[102] Jiashun Suo, Tianyi Wang, Xingzhou Zhang, Haiyang Chen, Wei Zhou, and Weisong Shi. Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data*, 10(1):227, 2023.

[103] Tong Zhang, Yin Zhuang, He Chen, Guanqun Wang, Lihui Ge, Liang Chen, Hao Dong, and Lianlin Li. Posterior instance injection detector for arbitrary-oriented object detection from optical remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[104] Yin Zhuang, Yuqing Liu, Tong Zhang, Liang Chen, He Chen, and Lianlin Li. Heterogeneous prototype distillation with support-query correlative guidance for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[105] Minglang Qiao, Mai Xu, Lai Jiang, Peng Lei, Shijie Wen, Yunjin Chen, and Leonid Sigal. Hypersor: Context-aware graph hypernetwork for salient object ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):5873–5889, 2024.

[106] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2022.

[107] whu csl. SAR Vehicle Detection Dataset. https://github.com/whu-csl/SAR_vehicle_detection_dataset. Accessed: Sep. 17, 2025.