DescribeEarth: Describe Anything for Remote Sensing Images

Kaiyu Li*, Zixuan Jiang*, Xiangyong Cao†, Jiayu Wang, Yuchen Xiao, Deyu Meng, Zhi Wang

***Abstract*—Automated textual description of remote sensing images is crucial for unlocking their full potential in diverse applications, from environmental monitoring to urban planning and disaster management. However, existing studies in remote sensing image captioning primarily focus on the image level, lacking object-level fine-grained interpretation, which prevents the full utilization and transformation of the rich semantic and structural information contained in remote sensing images. To address this limitation, we propose Geo-DLC, a novel task of object-level fine-grained image captioning for remote sensing. To support this task, we construct DE-Dataset, a large-scale dataset contains 25 categories and 261,806 annotated instances with detailed descriptions of object attributes, relationships, and contexts. Furthermore, we introduce DE-Benchmark, a LLM-assisted question-answering based evaluation suite designed to systematically measure model capabilities on the Geo-DLC task. We also present DescribeEarth, a Multi-modal Large Language Model (MLLM) architecture explicitly designed for Geo-DLC, which integrates a scale-adaptive focal strategy and a domain-guided fusion module leveraging remote sensing vision-language model features to encode high-resolution details and remote sensing category priors while maintaining global context. Our DescribeEarth model consistently outperforms state-of-the-art general MLLMs on DE-Benchmark, demonstrating superior factual accuracy, descriptive richness, and grammatical soundness, particularly in capturing intrinsic object features and surrounding environmental attributes across simple, complex, and even out-of-distribution remote sensing scenarios. All data, code and weights are released at https://github.com/earth-insights/DescribeEarth.**

***Index Terms*—Image captioning, Remote sensing image, Multimodel model**

I. Introduction

REMOTE sensing images, continuously acquired from satellites, aerial platforms, and drones, provide an indispensable tool for monitoring and understanding the Earth. Its applications span environmental science [1], [2], urban planning [3], [4], disaster management [5], agriculture [6], and defense [7]. Automatically generating precise and detailed textual descriptions for specific features or phenomena within these images is crucial for unlocking their full potential, enabling more efficient analysis, rapid decision-making, and broader accessibility for both domain experts and the public [8]–[10].

Historically, the task of image description, often referred to as image captioning, primarily focused on generating a single,

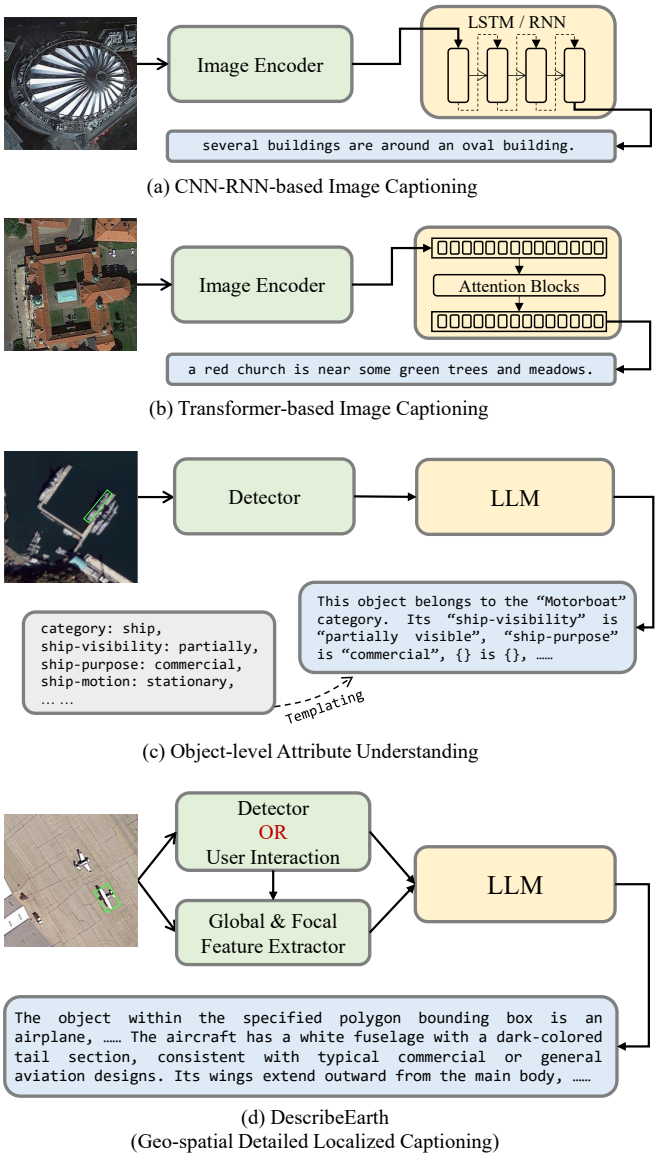


Fig. 1. Some methods proposed to resolve remote sensing image description related tasks. (a) and (b) are typical image captioning frameworks [8], [11]. (c) is a recent object description model [12], but it can only generate template-based descriptions across a limited number of categories and does not support interactive descriptions. (d) is our DescribeEarth model, which can generate detailed, open-ended and localized descriptions based on off-the-shelf detectors or user interactions.

This work is partially supported by the National Key R&D Program of China (2021ZD0112902), and China NSFC projects under contract 62272375, 12226004. *(Corresponding author: Xiangyong Cao)*

Kaiyu Li and Zhi Wang are with School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China (email: likyoo.ai@gmail.com, zhiwang@xjtu.edu.cn)

Zixuan Jiang is with College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an 710049, China (email: andrewjiang@stu.xjtu.edu.cn)

Xiangyong Cao, Jiayu Wang, and Yuchen Xiao are with School of Computer Science and Technology and Ministry of Education Key Lab For Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (email: caoxiangyong@xjtu.edu.cn, 2234112262@stu.xjtu.edu.cn, 3283879@qq.com)

Deyu Meng is with School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, Shaanxi, China, and Pazhou Laboratory (Huangpu), Guangzhou, Guangdong, China. (email: dymeng@mail.xjtu.edu.cn).

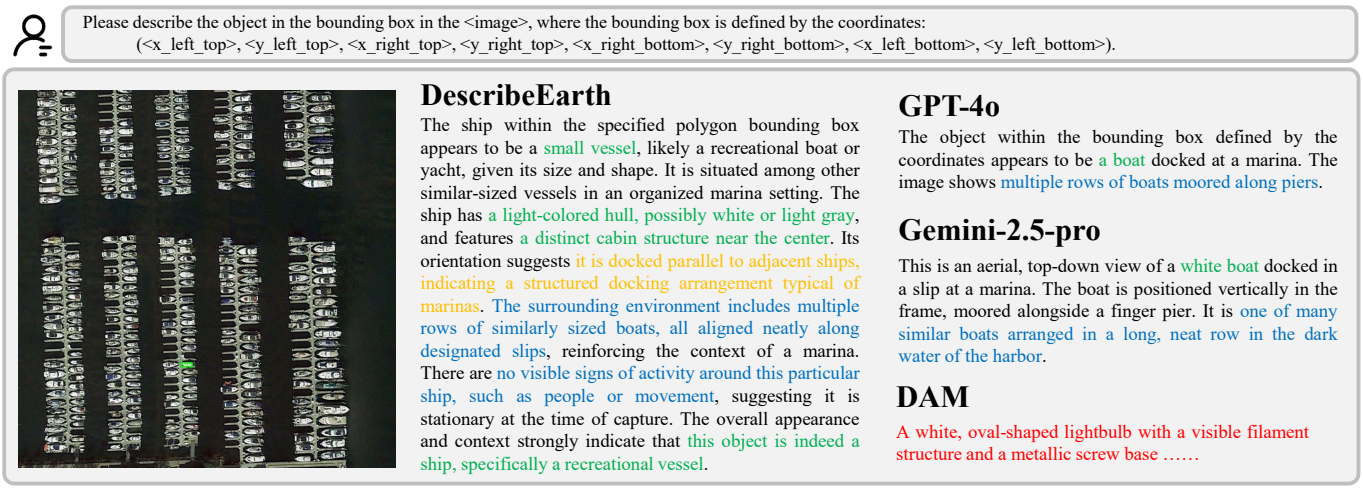


Fig. 2. Comparison of DescribeEarth, DAM [13], and other general MLLMs [14], [15] in describing remote sensing images. Green rendering represents intrinsic features of the target, yellow denotes spatial features, blue indicates contextual features, and red signifies incorrect descriptions. DescribeEarth delivers the most detailed and accurate object descriptions.

holistic sentence or paragraph for an entire image [16], [17]. Many methods have been proposed to solve this challenge, and some typical methods are presented in Fig. 1. Early approaches relied on architectures like CNN-RNN encoders [18]–[20] or attention-based mechanisms [21]–[23], often trained on small datasets and producing brief, coarse-grained sentences. With the advent of Multi-modal Large Language Models (MLLMs), these capabilities have significantly advanced, allowing for more coherent and contextually rich descriptions for general images [24]–[28]. In the remote sensing domain, early attempts at image captioning also focused on scene-level descriptions, often adapting general vision-language models (VLMs) to classify and provide coarse summaries of entire remote sensing images [29]–[32]. More recently, some efforts have moved towards generating region-specific descriptions. For instance, EagleVision [12] proposed an object-level attribute MLLM specifically for remote sensing, aiming to provide attribute-rich descriptions for detected objects. While EagleVision represents a step towards localized understanding in remote sensing, it often produces descriptions that are constrained by predefined attributes or templates, limiting the richness and open-ended detail necessary for comprehensive analysis. Moreover, the few training categories and non-interactive operations further limit its practicality.

The recent introduction of the Describe Anything Model (DAM) [33] marks a significant leap in Detailed Localized Captioning (DLC) for natural images and videos. DAM, through its focal prompt and localized vision backbone, adeptly balances local detail with global context, generating meticulously detailed descriptions for user-specified regions. Some recent work follows it and has also made significant contributions to this task [13], [34], [35]. While DAM and some MLLMs excel in the natural image domain, its direct application to remote sensing images reveals substantial performance gaps. As shown in Fig. 2, we present the description results of DAM and several general MLLMs on a remote sensing image. GPT-4o [14] and Gemini-2.5-pro [15] can only provide brief descriptions, lacking details of object attributes, spatial relationships, and contextual information. DAM has difficulty defining categories and providing precise descriptions for some remote sensing instances. This is primarily due to the differences between natural and remote sensing visual data, including unique perspectives (*e.g.*, nadir views), vast scale variations of objects, and the distinct semantic contexts relevant to geographical analysis. Consequently, general models like DAM, trained on natural image characteristics, struggle to accurately identify and provide detailed descriptions of objects and phenomena observed from remote sensing perspectives.

To bridge this critical gap and enable DLC for remote sensing images, a task we formally define as Geo-DLC, we need to address some challenges. Based on our analysis and building upon the inherent difficulties of localized captioning [33], we identify three obstacles for Geo-DLC:

- The development of Geo-DLC models is constrained by the lack of dedicated datasets. Existing remote sensing datasets typically offer labels for classification [36], bounding boxes for detection [37], masks for segmentation [38], or coarse descriptions for the entire image captioning [8], [9], [21], [39], but rarely provide the instance-level textual descriptions required for Geo-DLC. Manually creating such a dataset is impractical, requiring substantial resources and specialized geospatial expertise to accurately describe complex details and features.
- General VLMs, trained on natural images, possess limited recognition capabilities for remote sensing targets [40]–[43]. Objects in remote sensing images appear from unique perspectives, exhibit vast scale variations, and contain a significant number of small targets [44]. This makes them different from objects in natural images. Consequently, general models struggle to accurately identify and interpret even common objects when viewed from these specialized remote sensing perspectives, let alone capture their intricate details.
- Beyond training data, there is a lack of an evaluation

benchmark specifically designed for Geo-DLC. As stated in [33], traditional language metrics [45]–[47] and current LLM-based judgments [48] unfairly penalize models for generating correct details not present in an often incomplete reference caption.

To address these limitations and unlock the potential of detailed localized understanding in remote sensing, this paper introduces a comprehensive practice for Geo-DLC. Our contributions are summarized as follow:

- We introduce DE-Dataset, the first dataset for Geo-DLC. This dataset contains oriented bounding boxes (OBBs) for a diverse range of geographical objects, paired with instance-level detailed textual descriptions. DE-Dataset is constructed through a well designed data pipeline that leverages MLLMs and existing remote sensing object detection datasets, assisted by human verification. This enables efficient scaling of annotation to vast amounts of remote sensing images.
- We present DescribeEarth, a MLLM architecture explicitly designed for Geo-DLC task. To address the limited recognition capability of general models for remote sensing targets, DescribeEarth utilizes RemoteCLIP's features as a guiding prior [40] and integrates a novel visual feature fusion mechanism to effectively encode high-resolution details and remote sensing category prior of target regions while maintaining global context, leading to highly detailed and context-aware localized captions.
- We propose a high-quality benchmark, DE-Benchmark, specifically tailored for Geo-DLC. Following the rigorous attribute-based evaluation methodology of DAM, we meticulously design an evaluation dataset that moves beyond traditional reference-based metrics. This ensures models are appropriately rewarded for providing rich, accurate details pertinent to the remote sensing context and penalized for factual errors or irrelevant descriptions, thereby enabling comprehensive and fair assessment.

II. Related Work

A. Image Captioning

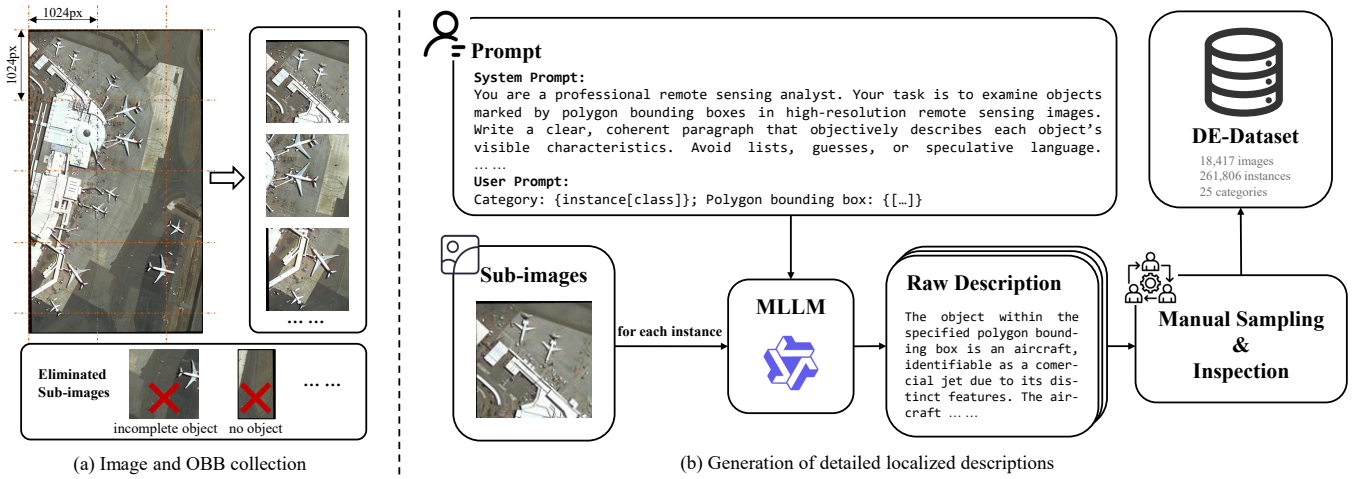
Image Captioning is a fundamental and critical research problem in the multi-modal field, which has attracted a lot of research due to its complexity. Early works were mainly inspired by natural language processing tasks such as machine translation [49], and adopted attention-based Encoder-Decoder architecture as the mainstream scheme [50]–[55]. These methods usually use region-based CNN (*e.g.*, ResNet [56]) or Transformer-based backbone [57] as the Encoder to extract local region features. Then RNN or LSTM is used as Decoder to generate natural language description. At this stage, a large number of studies focus on multi-scale attention and feature modeling in order to improve the understanding of images. However, limited by the expressive power of convolution operations, the model still has shortcomings in capturing the relationship between high-attention regions. Subsequently, several works noticed the advantages of graph structure in modeling element relationships and semantic dependencies [58]–[62] and began to try to construct scene graphs from images [63]–[69] or text [70], [71]. It is introduced into the Encoder-Decoder framework [72] to enhance the semantic alignment of image-text. These methods have made some progress in improving the accuracy of descriptions, but their generalization ability is still limited due to their reliance on small-scale fully supervised training.

With the rise of large-scale pre-trained models [73], researchers have begun to explore the introduction of pre-training paradigms into image captioning tasks [74]–[76]. Typical representatives such as ClipCap [77] significantly improved the generalization of the model by mapping the embeddings extracted by CLIP into the generative network. Since then, more works have focused on how to better align pre-trained visual features with text generation tasks to further improve the performance. At the same time, there are also studies that try to apply reinforcement learning and unsupervised learning methods to this task [78]–[80], focusing on image feature modeling and cross-modal alignment, but the completeness and length of the generated text are still limited by the Decoder architecture. In recent years, with the rapid development of MLLMs, the research paradigm of image captioning has further evolved. Some works directly leverage MLLMs to improve the richness and fluency of natural language output. For example, DLC task and DAM enable fine-grained natural language descriptions at the object level [33], but the performance of DAM is not ideal in remote sensing image scenes, and the detailed language representation at the object level still faces great challenges. In the field of remote sensing, EagleVision introduces object-level attribute-guided formatting descriptions. However, its templated output and limited categories constrain its ability to describe details and its generalizability [12].

B. Vision Language Foundation Models for Remote Sensing

With the rapid advancement of multimodal learning, VLMs have emerged as a critical paradigm for supporting diverse downstream tasks. Remote sensing, as a highly application-driven domain, has also witnessed a growing body of research leveraging VLMs. Within the contrastive learning framework represented by CLIP [73], numerous studies have sought to adapt the model to remote sensing scenarios. RemoteCLIP [40], GeoRSCLIP [81] and Git-RSCLIP [82] construct large-scale image–text datasets and fine-tune CLIP to enhance performance across multiple tasks. CRSR [83] and APPLeNet [84] introduce feature integration modules to strengthen cross-modal alignment, while ProGEO [85] and GRAFT [86] incorporate auxiliary information to improve both image encoding and modality interaction. S-CLIP [87] and RS-CLIP [88] further refine CLIP's domain adaptation through improved pseudo-labeling strategies and loss designs. Although these contrastive learning approaches achieve substantial performance gains, their scope remains confined to image–text matching, lacking the capability for natural language generation.

To address this limitation, recent studies have extended remote sensing applications to large-scale conversational VLM architectures. RS-LLaVA [89] builds upon CLIP as the vision



(a) Image and OBB collection

(b) Generation of detailed localized descriptions

Fig. 3. The production pipeline of our DE-Dataset. (a) is the tiling strategy used in dataset preprocessing. (b) is the generation process of localized description.

encoder and adapts the LLaVA [24] framework through instruction tuning, enabling dialog-based remote sensing VLMs. EarthGPT [90] and EarthMarker [91] optimize vision encoders to enhance multi-scale object representation, while VHM [32] and SkySenseGPT [92] mitigate CLIP's limited spatial awareness by constructing higher-quality image–text pairs. In parallel, several works such as RSGPT [93], LHRS-Bot [30], TEOChat [31], and UniRS [94] explore Vision–LLM connector architectures to improve multimodal perception. Beyond architectural advances, some methods have been proposed to address specific challenges of remote sensing VLMs. For instance, chain-of-thought reasoning and data augmentation strategies have been introduced to improve reasoning and generalization capabilities, as demonstrated in CPSeg [95], MGeo [96], and SpectralGPT [97].

C. Remote Sensing Datasets

Currently, most existing remote sensing VLMs are data-driven, and in general, different tasks rely on their specific datasets. For remote sensing visual question answering, EarthVQA [98] and CRSVQA [99] adopt manual annotation to ensure high-quality supervision. For object detection, datasets such as FAIR1M [100], DOTA [37], and DIOR [101] have been developed, while classification tasks are supported by AID [36], NWPU-RESISC45 [102], and UCM [103], and semantic segmentation relies on pixel-level datasets such as iSAID [104]. Building on these resources, several extended datasets have been proposed by combining or re-annotating existing ones to address new challenges, including RSVGD [105], RefSegRS [106], and SkyEye-968K [107]. Although these datasets cover fundamental tasks and some emerging applications, the demand for large-scale remote sensing vision-language datasets is growing rapidly with the development of MLLMs. To address this, researchers have explored automated annotation pipelines supported by general foundation models. GeoChat [29] designs system-level prompts to interact with Vicuna [108] for generating multi-turn question–answer pairs. HqDC-1.4M [32] employs Gemini-1.0-pro-vision to produce textual descriptions for large-scale remote sensing images, yielding extensive image–text pairs. FIT-RS [92] leverages TinyLLaVA-3.1B [109] to generate background descriptions of remote sensing images and integrates GPT-4 or GPT-3.5 for high-quality annotation. Despite these advances, high-quality fine-grained object-level datasets remain scarce, and current pipelines for generating object-level descriptions are still immature.

III. Task, Dataset and Benchmark

This section formally defines the Geo-DLC task, details the construction of DE-Dataset, and introduces the DE-Benchmark. These address the challenges in Section I.

A. Task Formulation

Following the framework of DLC as established in [33], the Geo-DLC task extends this concept to the remote sensing image. Unlike general image captioning which provides a holistic summary, Geo-DLC focuses on generating precise, comprehensive textual descriptions for specific geographical regions or objects within remote sensing images. Formally, given an input image I and a user-specified geographical Region of Interest (ROI) R within I (typically represented by various localization cues such as a bounding box, point, or binary mask), the objective is to generate a detailed textual description T centered on the instance. Such descriptions are required to cover both intrinsic object features and contextual environmental attributes, ensuring logical consistency and informativeness. This task can be formulated as:

$$T = Model(I, R) \tag{1}$$

The Geo-DLC task is challenging due to the inherent differences in object appearance from an aerial viewpoint, the vast scale variations, and the complex background contexts prevalent in remote sensing images. It demands a model to not only recognize objects under these unique conditions but also to elaborate on their fine-grained attributes and relationships, which existing general MLLMs often fail to achieve.

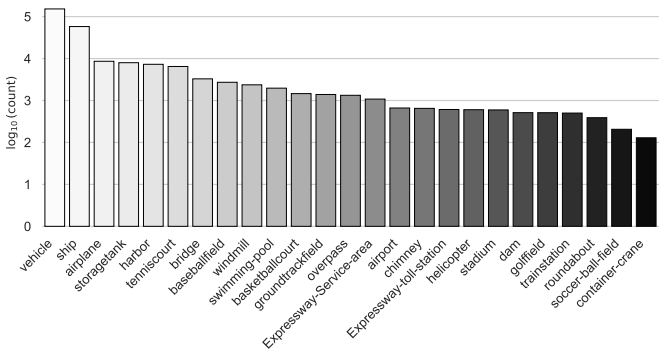


Fig. 4. The distribution of instance categories and their corresponding quantities (log_{10}) in DE-Dataset dataset.

(a) Simple sample (b) Complex sample (c) OOD sample

Fig. 5. Examples of different instance types in DE-Benchmark.

B. DE-Dataset

To support Geo-DLC task, we introduce DE-Dataset, the first Geo-DLC dataset constructed through a multi-stage process leveraging existing remote sensing datasets, MLLMs for automated annotation, and rigorous human quality control.

1) Data Construction Pipeline: We construct our DE-Dataset based on two oriented object detection datasets: DIOR [101] and DOTA [37]. To manage the large variation in image scales within DOTA and ensure consistent input for annotation, we perform a unified tiling strategy. Specifically, for large-scale images with spatial resolutions greater than 1024×1024, we partition them into non-overlapping 1024×1024 sub-images, as well as additional boundary tiles, to preserve the diversity of instance placements. All instance bounding boxes are then updated relative to the sub-image coordinates, and some low-quality sub-images are eliminated (*e.g.*, those containing only incomplete objects or no any objects, etc.). This preprocessing yields a refined collection of remote sensing sub-images and their corresponding OBBs, as shown in Fig. 3 (a).

To generate detailed localized captions at scale, we utilize a powerful MLLM, Qwen2.5-VL-32B [26], as an automated annotator. To alleviate cognitive load on the annotator and produce more accurate descriptions, we provide corresponding class name, sub-image, and its OBB for each object instance as input. We design prompts explicitly instructing the model to generate descriptions covering the following key aspects:

- Intrinsic features: appearance, structure, and other object-specific characteristics.
- Spatial features: orientation, position, and other localization attributes.
- Contextual features: surrounding environment, as well as signs of human or social activity.

In addition, to ensure dataset quality and reduce hallucinations, we impose strict linguistic constraints on the model output. Speculative or uncertain language is explicitly prohibited, ensuring that the generated descriptions remain logical, consistent, and factually grounded. Our prompt is shown in Fig. 3 (b). Furthermore, after the initial dataset is obtained, a random subset of the automatically generated captions, covering a diverse range of categories and complexities, is meticulously reviewed by human experts with remote sensing domain knowledge. These experts verify the factual correctness, descriptive richness, and adherence to the Geo-DLC task definition. Incorrect or low-quality captions are either revised, regenerated, or discarded. The overall construction pipeline is illustrated in Fig. 3.

2) Dataset Statistics: The DE-Dataset comprises 18,417 images and 261,806 instances, with an average of 14 instances per image. Each instance is accompanied by a detailed description averaging 119.83 words in length. The DE-Dataset contains 25 categories, with the category distribution shown in Fig. 4. Unlike traditional remote sensing image captioning datasets that offer only coarse-grained annotations or full-image descriptions [8], [9], [21], [39] and in contrast to object-level attribute datasets like EVAttrs-95K [12] which provide templated descriptions for a limited number of categories, DE-Dataset stands out. By adopting an extended paragraph form for descriptions, DE-Dataset offers a richer, more open-ended, and flexible representation of instances, encompassing both intrinsic attributes and contextual information without predefined templates. This design significantly enhances the practicality and versatility of the dataset for Geo-DLC tasks.

C. DE-Benchmark

To provide a robust and fair evaluation for Geo-DLC models, we construct DE-Benchmark, which adapts and extends the attribute-based evaluation protocol established in [13]. This protocol moves beyond traditional reference-based metrics by assessing factual accuracy and descriptive richness against a set of predefined attributes, a critical feature for Geo-DLC given the often incomplete nature of reference captions.

1) Preliminary Data Collection: Following the procedure used in DE-Dataset, we construct our benchmark based on the validation sets of DIOR and DOTA. Reusing the proposed data construction pipeline, we first generate detailed descriptions of the instances. For each category, several representative instances are manually selected to ensure coverage, resulting in preliminary evaluation samples. Considering the significant variations among remote sensing images across different regions, as well as issues such as dense distribution, large-scale changes, occlusion, and complex contextual backgrounds of some remote sensing objects, we further classify instances into simple and complex samples based on instance size, spatial location, occlusion level, contextual complexity, etc. Furthermore, to evaluate the model's generalization and open-set recognition ability, we additionally selected 10 out-of-

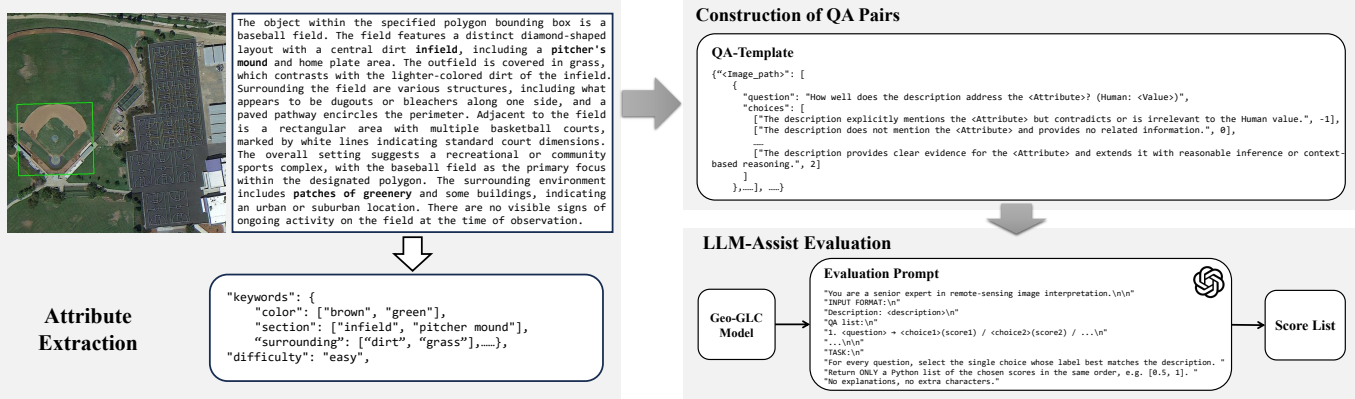


Fig. 6. The workflow of construction and evaluation for DE-Benchmark. We first manually extract key attributes and define difficulty for instances, which are then used to generate structured QA pairs with predefined scoring rubrics. Finally, an LLM evaluates the Geo-DLC model's generated descriptions against these QA pairs to produce a score.

distribution (OOD) categories from the xView [110] dataset which are not included in the DE-Dataset. Their descriptions are generated using the same pipeline, and the resulting samples are merged with the preliminary benchmark data. Some examples showing the different types of instances are shown in Fig. 5.

2) Attribute-Based QA Construction: For every instance in the DE-Benchmark, human experts carefully extract a set of attribute-value pairs. An attribute specifies a key feature that should be present in the description, *e.g.*, color, neighboring objects, spatial orientation. The corresponding value is the accurate, human-annotated information for that attribute. This manual annotation provides the ground truth for evaluating the factual accuracy and quality of model-generated captions, as shown on the left side of Fig. 6. Based on these attribute-value pairs, we create QA templates for evaluation. Each attribute leads to a question (*e.g.*, "How well does the description address the color?"). For each question, we design a set of detailed options. These options are scoring rubrics, not factual answers. They describe various levels of how well the model's output covers or presents the specific attribute. These rubrics include whether the description contradicts the human value, does not mention it, gives vague hints, supports it through implication, provides clear evidence, or extends it with reasonable inference. Each rubric option has a pre-assigned numerical score, allowing for a fine-grained assessment of the model's descriptive quality for each attribute.

3) LLM-Assist Evaluation: Evaluating detailed, open-ended descriptions in Geo-DLC is challenging. For this, we develop an LLM-Assist evaluation system that uses an LLM as a judge. Specifically, we adopt GPT-4.1, given its strong reasoning and language understanding. For each evaluation instance, the Geo-DLC model first generates its detailed localized caption. Then, this caption, along with its corresponding attribute-based QA pairs, is provided as input to the judgment LLM. The LLM is prompted to analyzes the model's description and, for each question, selects the rubric option that best matches how the attribute is addressed in the generated caption. Moreover, 5 general language quality questions further evaluate the description's grammatical correctness, logical flow, factual reliability, inferential capability, and conciseness. For structured analysis, all questions are categorized into four main types: "Appearance" (intrinsic visual characteristics), "Surrounding" (environmental and contextual information), "Language" (grammatical correctness, logical flow, and naturalness of descriptions), and "Usage" (functional properties and potential usage scenarios). The construction and evaluation process of DE-Benchmark is shown in Fig. 6.

IV. Methodology

To address the challenges of Geo-DLC mentioned above, *i.e.*, large-scale variations, limited recognition capabilities from the remote sensing perspective, and the lack of detailed and contextual descriptions, we propose DescribeEarth. Built upon the Qwen2.5-VL-3B-Instruct, DescribeEarth incorporates a scale-adaptive focal strategy to manage diverse object scales and a domain-guided fusion module (DFM) to integrate domain-specific semantic priors. These modules enhance the model's ability to generate highly detailed and accurate localized captions for remote sensing images. An overview of the DescribeEarth architecture is illustrated in Fig. 7.

A. Scale-adaptive Focal Strategy

Inspired by DAM's Focal Prompt strategy [33], which balances local detail with global context in natural images, we develop a scale-adaptive focal strategy for Geo-DLC. Remote sensing images, with their nadir views and extreme variations in object scales—from expansive aircraft runways to minute vehicles—present unique challenges for direct application of fixed-size focal crops. Our strategy improves DAM by dynamically generating a focal crop that, alongside the full original global image, ensures high-resolution encoding of the target region while maintaining crucial surrounding context, tailored for remote sensing characteristics. Here, the scale of each object is precisely determined by the longer side of its minimum enclosing axis-aligned bounding box. Based on this definition, we implement three distinct mechanisms for generating the focal crop:

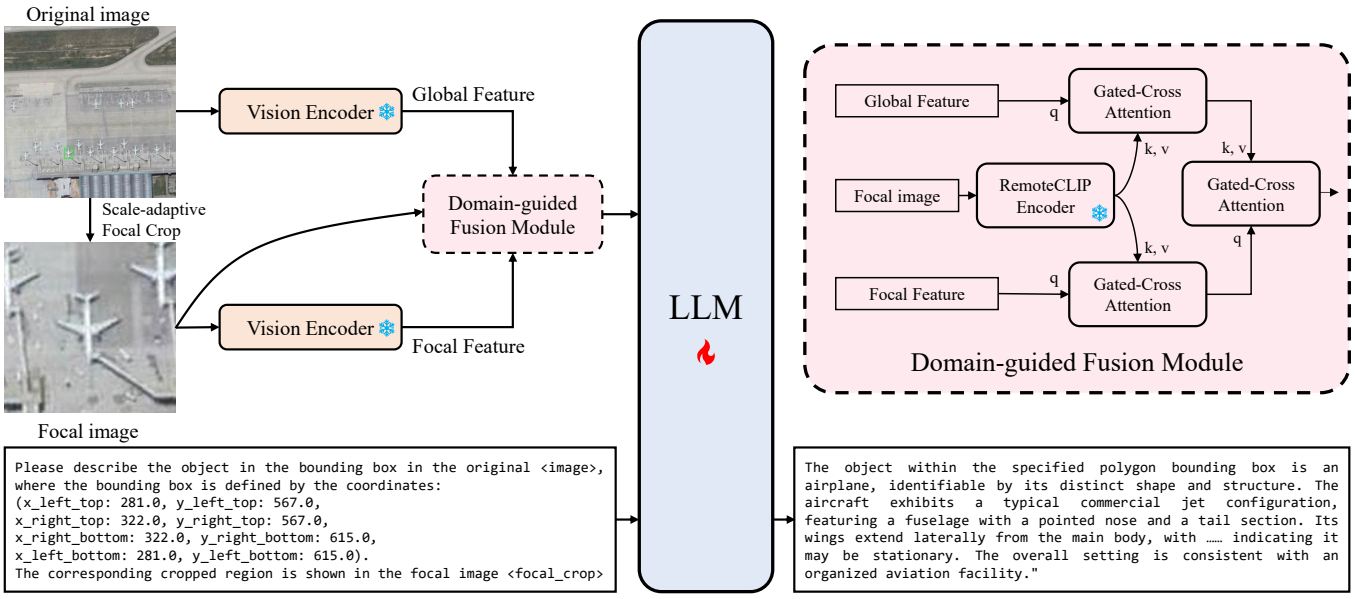


Fig. 7. The architecture of DescribeEarth. It uses a scale-adaptive focal strategy to extract both global and focal visual features from an original image. These features are then integrated with domain-specific priors from a RemoteCLIP encoder via DFM. The enriched visual representation, along with object bounding box coordinates embedded in the textual prompt, is then fed to the LLM for generating detailed localized captions.

- Large objects ($>$ 224 pixel): For objects whose original bounding box is already large enough to encompass sufficient fine-grained detail, the bounding box itself is directly cropped to form the focal region.
- Medium objects (112-224 pixel): A fixed 224×224 patch, centered at the bounding box centroid, is extracted as the focal crop. This patch size is designed to capture both fine object details and relevant surrounding context.
- Small objects ($<$ 112 pixel): A 112×112 square patch is extracted around the bounding box center. This tighter crop ensures that the limited pixels available for the target object receive maximum attention, emphasizing its details while retaining sufficient environmental cues to aid recognition and description.

All these generated focal crops, irrespective of their original scale, are then resized to a uniform resolution to match the input size of the vision encoder. This focal crop, alongside the original global image, then serves as input to our dual-path vision encoding pipeline, ensuring comprehensive feature extraction that leverages both fine-grained local details and broader contextual signals, as shown in Fig. 7.

B. Vision Encoding and Domain-guided Fusion

Both the global image and the focal cropped image are processed by a shared vision encoder, yielding global features and focal features. This encoder is initialized with the visual weights from the pre-trained Qwen2.5-VL-3B model and is kept frozen during training to leverage its general visual understanding. Weight sharing enforces representational consistency and improves parameter efficiency, enabling the model to align cross-scale representations in a unified feature space.

To further enhance semantic richness, we introduce DFM. This module integrates domain-specific priors from RemoteCLIP [40], which is adept at understanding remote sensing visuals. Specifically, focal crops are also encoded by a frozen RemoteCLIP encoder and then projected via a linear layer to obtain the domain-specific feature. The fusion of these diverse visual cues (global, focal, and domain-specific) is managed through a hierarchical gated cross-attention mechanism [111], as shown in Fig. 7. First, global and focal features interact with the domain-specific features in parallel gated cross-attention modules. This fusion aims to semantically enrich both the global and focal representations with remote sensing-specific knowledge. Following this, the enhanced global features are used to refine the focal features in a subsequent cross-attention operation. This operation allows the fine-grained, localized information from the focal features to be further enriched and contextualized by the broader global representation. This ensures that the final visual representation, which is passed to the LLM, is both highly detailed for the localized region and deeply grounded in its wider geographical context.

C. Integration with the LLM

For integrating localization with the LLM in Geo-DLC, DescribeEarth adopts a strategy distinct from DAM, which directly encodes mask information within its vision backbone via dedicated mask embedding layers. Recognizing that direct manipulation of visual input with masks can alter the vision encoder's input shape or distribution, potentially requiring extensive fine-tuning or impacting its general understanding, we instead convey the ROI by embedding OBB coordinates directly into the LLM's textual prompt. This design preserves the visual input stream to our shared vision encoder in its pristine, standard image patch format, allowing the powerful, pre-trained vision encoder (from Qwen2.5-VL-3B) to remain frozen and untouched. This avoids new visual biases and ensures robust leveraging of its established general representations. Furthermore, embedding OBB coordinates as text

TABLE I
PERFORMANCE COMPARISON ACROSS CATEGORIES, DIFFICULTY LEVELS, AND QUESTION TYPES (IN %). THE NUMBERS IN PARENTHESES INDICATE THE GAP BETWEEN OUR DESCRIBEEARTH MODEL AND OTHER MODELS. POSITIVE VALUES IN GREEN, NEGATIVE IN RED. "FT" INDICATES FINE-TURNING ON OUR DE-DATASET.

	Key	DescribeEarth	Qwen2.5-VL-3B (ft)	Qwen2.5-VL-3B	GPT-4o	Gemini-2.5-Pro	DAM
Category	Baseball field	**75.75**	62.00 (+13.75)	66.00 (+9.75)	56.00 (+19.75)	71.25 (+4.50)	72.00 (+3.75)
	Tennis court	**78.25**	77.50 (+0.75)	61.00 (+17.25)	63.00 (+15.25)	60.00 (+18.25)	56.50 (+21.75)
	Stadium	75.50	**76.50** (-1.00)	76.00 (-0.50)	69.00 (+6.50)	51.25 (+24.25)	59.00 (+16.50)
	Harbor	**86.34**	84.79 (+1.55)	72.16 (+14.18)	73.45 (+12.89)	60.57 (+25.77)	60.82 (+25.52)
	Golf field	**68.85**	65.98 (+2.87)	59.22 (+9.63)	67.42 (+1.43)	46.11 (+22.74)	60.86 (+7.99)
	Ground track field	**83.43**	77.53 (+5.90)	74.44 (+8.99)	74.44 (+8.99)	70.79 (+12.64)	64.89 (+18.54)
	Airport	**74.31**	71.30 (+3.01)	61.81 (+12.50)	71.30 (+3.01)	42.59 (+31.72)	61.34 (+12.97)
	Expressway service area	56.40	58.53 (-2.13)	55.62 (+0.78)	**61.63** (-5.23)	37.40 (+19.00)	51.74 (+4.66)
	Overpass	68.81	**72.14** (-3.33)	68.10 (+0.71)	65.48 (+3.33)	60.24 (+8.57)	56.43 (+12.38)
	Expressway toll station	65.74	65.97 (-0.23)	53.47 (+12.27)	**67.59** (-1.85)	41.90 (+23.84)	53.01 (+12.73)
	Airplane	78.80	**80.00** (-1.20)	64.67 (+14.13)	75.00 (+3.80)	48.37 (+30.43)	70.92 (+7.88)
	Storage tank	80.00	**81.50** (-1.50)	75.25 (+4.75)	73.00 (+7.00)	65.00 (+15.00)	63.75 (+16.25)
	Train station	**83.78**	79.52 (+4.26)	68.88 (+14.90)	73.94 (+9.84)	57.18 (+26.60)	62.23 (+21.55)
	Ship	**83.33**	83.33 (0.00)	65.44 (+17.89)	73.04 (+10.29)	56.37 (+26.96)	65.69 (+17.64)
	Windmill	71.24	**72.04** (-0.80)	69.62 (+1.62)	67.20 (+4.04)	61.02 (+10.22)	60.48 (+10.76)
	Basketball court	53.64	56.82 (-3.18)	50.00 (+3.64)	**65.45** (-11.81)	39.09 (+14.55)	50.68 (+2.96)
	Chimney	63.29	63.06 (+0.23)	60.59 (+2.70)	**66.44** (-3.15)	38.29 (+25.00)	59.01 (+4.28)
	Dam	**73.41**	72.73 (+0.68)	69.77 (+3.64)	72.50 (+0.91)	38.18 (+35.23)	64.77 (+8.64)
	Bridge	68.87	68.63 (+0.24)	68.14 (+0.73)	**71.57** (-2.70)	42.65 (+26.22)	63.73 (+5.14)
	Building	67.71	65.10 (+2.61)	**70.31** (-2.60)	61.46 (+6.25)	41.67 (+26.04)	16.67 (+51.04)
	Construction site	65.82	54.59 (+11.23)	65.31 (+0.51)	**79.08** (-13.26)	43.88 (+21.94)	55.10 (+10.72)
	Damaged building	56.63	55.10 (+1.53)	53.57 (+3.06)	**62.24** (-5.61)	40.82 (+15.81)	21.43 (+35.20)
	Facility	**88.37**	68.02 (+20.35)	83.14 (+5.23)	72.67 (+15.70)	46.51 (+41.86)	41.86 (+46.51)
	Locomotive	**72.22**	61.11 (+11.11)	56.67 (+15.55)	63.33 (+8.89)	45.56 (+26.66)	21.67 (+50.55)
	Roundabout	**69.89**	69.35 (+0.54)	64.52 (+5.37)	64.78 (+5.11)	50.81 (+19.08)	54.57 (+15.32)
	Helicopter	67.38	**70.24** (-2.86)	62.38 (+5.00)	60.71 (+6.67)	39.52 (+27.86)	60.71 (+6.67)
	Large vehicle	**73.16**	70.79 (+2.37)	59.74 (+13.42)	63.68 (+9.48)	34.74 (+38.42)	59.74 (+13.42)
	Pylon	67.07	57.32 (+9.75)	60.37 (+6.70)	**70.12** (-3.05)	48.78 (+18.29)	21.34 (+45.73)
	Shed	62.24	62.76 (-0.52)	62.24 (0.00)	**67.86** (-5.62)	40.82 (+21.42)	44.90 (+17.34)
	Tower	66.28	64.53 (+1.75)	66.28 (0.00)	**71.51** (-5.23)	46.51 (+19.77)	20.93 (+45.35)
	Tower crane	64.58	63.54 (+1.04)	54.17 (+10.41)	**68.75** (-4.17)	33.33 (+31.25)	0.00 (+64.58)
	Vehicle lot	80.23	**87.79** (-7.56)	65.70 (+14.53)	71.51 (+8.72)	44.19 (+36.04)	37.79 (+42.44)
Difficulty Level	Simple	**74.19**	68.07 (+6.12)	66.81 (+7.38)	70.24 (+3.95)	53.30 (+20.89)	62.20 (+11.99)
	Complex	**70.24**	63.76 (+6.48)	62.19 (+8.05)	65.51 (+4.73)	46.77 (+23.47)	58.54 (+11.70)
	OOD	**68.78**	63.75 (+5.03)	63.59 (+5.19)	68.78 (0.00)	43.01 (+25.77)	28.22 (+40.56)
Question Type	Appearance	42.26	**53.77** (-11.51)	31.43 (+10.83)	35.49 (+6.77)	13.94 (+28.32)	31.94 (+10.32)
	Surround	**54.41**	37.86 (+16.55)	35.68 (+18.73)	43.99 (+10.42)	11.32 (+43.09)	6.46 (+47.95)
	Language	**98.77**	98.07 (+0.70)	96.88 (+1.89)	98.14 (+0.63)	85.43 (+13.34)	86.88 (+11.89)
	Usage	35.49	32.37 (+3.12)	31.03 (+4.46)	**44.64** (-9.15)	13.62 (+21.87)	20.98 (+14.51)

provides the LLM with explicit, numerical spatial information, enabling it to directly reason about the precise location, extent, and crucial orientation of remote sensing objects, which is often paramount for geometrically accurate descriptions. OBBs are inherently superior to horizontal bounding boxes for delineating arbitrarily oriented structures prevalent in nadir-view image, and their textual representation allows the LLM to comprehend these complex geometries without visual perturbation. This text-based localization input is also highly flexible, naturally compatible with various text-conditioned tasks and user interaction paradigms, paving the way for easier integration with future multi-modal applications without altering the visual processing pipeline.

During inference, user interaction for specifying the ROI is highly flexible. Common inputs such as clicks or interactive bounding boxes are first pre-processed into a unified OBB representation. This conversion leverages segmentation models like SAM [112], ensuring that even imprecise user inputs can be transformed into a standardized, precise OBB format. This processed OBB, along with the visual tokens, forms the textual prompt which is then fed to the LLM. The LLM then performs autoregressive decoding, leveraging both these rich visual features and the explicit OBB textual information, to generate the fine-grained, detailed Geo-DLC caption.

V. EXPERIMENTS

A. Experimental Setup

The DescribeEarth model was trained using two A800 GPUs. The basic architecture is built on Qwen2.5-VL-3B. The vision encoder component of DescribeEarth is initialized from the pre-trained Qwen2.5-VL-3B model and is kept frozen throughout the training process to preserve its established general visual understanding capabilities. To incorporate domain-specific knowledge, RemoteCLIP, a pre-trained ViT-B architecture, is utilized; its weights are also kept frozen, and intermediate outputs are employed rather than only final encoder features to enrich the visual representation. The LLM and the vision-language connector (MLP) within DescribeEarth are initialized with pre-trained Qwen2.5-VL-3B weights. Training was conducted with a batch size of 4, employing gradient accumulation every 4 steps, and a single epoch was run under a fine-finetuning strategy. The global image resolution is set to 448×448, while the focal image resolution is set to 224×224.

B. Main Results

Table I presents a detailed performance comparison of DescribeEarth against several state-of-the-art closed-source multimodal large language models (GPT-4o, Gemini-2.5-pro), a DLC model (DAM), and the original Qwen2.5-VL-3B-Instruct on the DE-Benchmark. Overall, DescribeEarth consistently achieves strong performance across various categories and difficulty levels. It outperforms the best closed-source model by a notable margin of 3.95% on simple instances and 4.73% on complex instances. While GPT-4o shows competitive performance on some categories (*e.g.*, it is slightly better on "Basketball court" and "Construction site" categories, and achieves 68.78%, matching DescribeEarth's score on OOD samples, indicating its broad generalization ability), DescribeEarth significantly surpasses all other compared models, demonstrating its specialized effectiveness for remote sensing images. Specifically, DescribeEarth shows substantial gains in categories like "Harbor", "Ground track field", "Facility", and "Locomotive", which are highly relevant and frequent objects in remote sensing. DAM, designed for natural images, struggles significantly on remote sensing data, with gaps as large as +64.58% relative to DescribeEarth for "Tower crane", highlighting the domain adaptation challenge.

When analyzing performance across different question types, DescribeEarth shows particular strengths in "Surrounding" descriptions, with a remarkable gain of 10.42% over GPT-4o and 43.09% over Gemini-2.5-pro. This indicates the effectiveness of our design in integrating contextual environmental attributes. Furthermore, it exhibits strong performance in "Appearance" and "Language" type questions, with respective gains of +6.77% and +0.63% over GPT-4o. These results underscore DescribeEarth's ability to generate rich, factually accurate, and grammatically sound localized descriptions, which are critical for geospatial analysis. Although GPT-4o outperforms DescribeEarth in "Usage" related questions, DescribeEarth's overall robustness and domain-specific excellence make it a superior choice for Geo-DLC.

C. Ablation Studies

1) Effect of Domain Guidance: To understand the influence of domain-specific guidance on model performance, we evaluated three distinct settings, as presented in Table II. The feature-level domain guidance setting, which is the default configuration for DescribeEarth, leverages RemoteCLIP features directly integrated into the fusion module, effectively providing a domain-specific prior. The "No Guidance" setting omits this prior, relying solely on direct fusion of global and focal features via gated cross-attention. The text-prompt domain guidance setting, in contrast, appends RemoteCLIP predictions as a simple textual hint ("RemoteCLIP suggests it's <category>") to the prompt during both training and inference. The results clearly demonstrate that feature guidance yields the best overall performance, indicating the crucial role of domain-specific categorical priors in enhancing localized captioning. Text guidance performs better than "No Guidance" on simple instances, suggesting that even a simple textual hint can aid in straightforward recognition tasks. However, in more complex and OOD scenarios, text guidance performs worse than No Domain Guidance. This indicates that directly relying on RemoteCLIP's raw category predictions can introduce misleading information when these predictions are inaccurate for challenging or unfamiliar objects. Our feature guidance strategy, by contrast, integrates RemoteCLIP's visual features more subtly within the fusion module, avoiding direct reliance on potentially erroneous explicit category predictions and thus maintaining robustness across diverse scenarios.

TABLE II
COMPARISON OF DIFFERENT DOMAIN GUIDANCE STRATEGIES ON DE-BENCHMARK (IN %). NUMBERS IN PARENTHESES DENOTE THE PERFORMANCE GAP. POSITIVE VALUES IN GREEN, NEGATIVE IN RED.

	Feature Guidance	No Guidance	Text Guidance
Simple	**74.41**	72.03 (+2.38)	73.50 (+0.91)
Complex	**70.04**	68.72 (+1.32)	68.43 (+1.61)
OOD	**66.48**	65.67 (+0.81)	64.90 (+1.58)

TABLE III
COMPARISON OF DIFFERENT FEATURE FUSION STRATEGIES. NUMBERS IN PARENTHESES DENOTE THE PERFORMANCE GAP. POSITIVE VALUES IN GREEN, NEGATIVE IN RED.

	DFM	GCA	Concat	Q-Former
Simple	74.41	72.03 (+2.38)	**75.20** (-0.79)	65.97 (+8.44)
Complex	**70.04**	68.72 (+1.32)	69.78 (+0.26)	64.82 (+5.22)
OOD	**66.48**	65.67 (+0.81)	65.07 (+1.41)	64.85 (+1.63)

2) Effect of Feature Fusion Strategy: Table III explores the impact of different feature fusion strategies, with our DFM achieving the highest scores at both complex and OOD levels. In comparison, the pure gated cross attention strategy (denoted as "GCA" in Table III), which directly fuses global and focal features, performs worse than DFM across all levels (*e.g.*, 72.03% for simple instances, 2.38% lower than DFM). The "Concat" strategy, which simply concatenates global and focal features, yields 70.31% for simple instances, appearing competitive with GCA. However, its performance significantly degrades for complex and especially OOD instances, showing a greater decline compared to DFM. This indicates that simple feature concatenation is insufficient to capture the complex relationships and domain-specific insights required for Geo-DLC, particularly when dealing with novel or challenging instances, where deep semantic interaction during fusion is critical. Similarly, employing a Q-Former [113] as the fusion module resulted in significantly lower performance than DFM, further emphasizing the efficacy of our DFM design in integrating domain-specific priors to enrich visual representations.

3) Effect of Vision Module Training: We further investigate whether training the vision encoder affects model performance, which is a subtle difference in training between DescribeEarth and DAM. DAM integrate localization information by altering the visual input stream itself, typically by adding a mask as a fourth input channel. This input format necessitates fine-tuning the vision encoder to adapt to the new data distribution. Conversely, DescribeEarth is designed to decouple localization from the visual encoder. We provide location information (*i.e.*, OBB coordinates) textually,

TABLE IV
COMPARISON OF WHETHER THE VISUAL ENCODER IS FROZEN (IN %). NUMBERS IN PARENTHESES DENOTE THE PERFORMANCE GAP. POSITIVE VALUES IN GREEN, NEGATIVE IN RED.

	Frozen	Trainable
Simple	**74.41**	73.17 (+1.24)
Complex	70.04	**70.27** (-0.23)
OOD	**66.48**	64.96 (+1.52)

TABLE V
ABLATION OF THE FOCAL STRATEGY AND GLOBAL IMAGE'S INPUT RESOLUTION (IN %). NUMBERS IN PARENTHESES DENOTE THE PERFORMANCE GAP. POSITIVE VALUES IN GREEN, NEGATIVE IN RED.

Resolution	**448×448**	448×448	224×224
Focal strategy	**Scale-adaptive**	Fixed-size	Scale-adaptive
Simple	74.19	73.59 (+0.60)	**74.41** (-0.22)
Complex	**70.24**	69.58 (+0.66)	70.04 (+0.20)
OOD	**68.78**	66.54 (+2.24)	66.48 (+2.30)

directly to the LLM. This approach leaves the visual input pristine and in-distribution for the pre-trained vision encoder. Consequently, we can freeze the vision encoder, preserving its powerful, pre-trained representations without the risk of degradation or over-fitting. To validate this design choice, we compare two settings: the fully frozen vision encoder and the trainable vision encoder. The results in Table IV affirm our strategy. Freezing the encoder not only enhances generalization to simple (+1.24%) and OOD (+1.52%) instances but also maintains the integrity of the pre-trained visual backbone. While a trainable encoder shows a marginal improvement in complex scenarios (0.23%), the overall benefits of our freezing strategy make it a more parameter-efficient and robust approach for the Geo-DLC task.

4) Effect of Scale-Adaptive Focal Strategy: Our scale-adaptive focal strategy is meticulously designed to balance the encoding of fine-grained target details with essential surrounding context. In this strategy, we employ a hierarchical cropping approach for objects of different sizes to ensure the model perceives both the object itself and its environment. As listed in Table V, scale-adaptive focal strategy proves crucial for Geo-DLC, significantly outperforming the fixed-size strategy (directly cropping the object's bounding box). This confirms that dynamically adjusting the focal crop size is critical for accommodating the vast scale variations in remote sensing image. To further enhance this capability, we investigate the effect of the global image resolution. The results in Table V confirm that utilizing a higher-resolution global image (448×448) consistently improves performance, showing gains on complex ($+0.20\%$) and OOD instances ($+2.30\%$). This demonstrates that a clearer global context is instrumental for improving the effectiveness of the focal strategy, allowing for more accurate and context-aware descriptions.

D. Integration with Detection Models

Beyond user-interactive queries, a key application of DescribeEarth is its integration into fully automated pipelines for large-scale geospatial analysis. This capability allows for the automatic enrichment of remote sensing image with dense, semantic annotations. As illustrated in Fig. 8, the workflow operates in two main stages. First, an off-the-shelf object detection model or region proposal network is applied to a remote sensing image to identify and localize objects of interest, outputting their respective OBBs. Second, each detected OBB, along with the original image, is sequentially fed into our DescribeEarth model, which then generates a detailed, context-aware textual description for the specific object within that box. This integration creates a powerful and scalable solution for automatically annotating vast amounts of remote sensing data. It moves beyond simple class labels, enriching geospatial datasets with semantic, human-readable information that can be indexed and searched, thereby significantly enhancing the capabilities of geographic information systems and enabling advanced analytical queries.

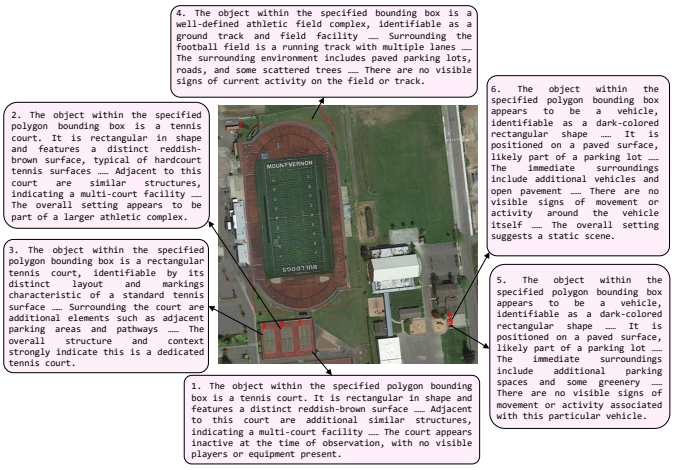


Fig. 8. Pipeline for automated image captioning

VI. CONCLUSION

In this paper, we present a comprehensive framework to advance Geo-DLC, a critical task for the semantic interpretation of remote sensing imagery. To enable meaningful progress, we introduce DE-Dataset, the first large-scale dataset specifically designed for this task, featuring instance-level OBBs paired with rich, fine-grained descriptions. Recognizing the limitations of traditional evaluation methods, we also develop DE-Benchmark, an attribute-based and LLM-assisted protocol that ensures a fair and nuanced assessment of a model's ability to generate factually accurate and contextually relevant captions. These resources are instrumental in the development and validation of our DescribeEarth model, which demonstrates a marked improvement in interpreting complex geospatial objects over state-of-the-art closed-source alternatives. Ultimately, this work establishes a new standard for Geo-DLC and provides a robust foundation to spur the development of more sophisticated vision-language models capable of unlocking deeper insights from Earth observation data for automated analysis, monitoring, and intelligence applications.

REFERENCES

[1] J. Chen, S. Chen, R. Fu, D. Li, H. Jiang, C. Wang, Y. Peng, K. Jia, and B. J. Hicks, "Remote sensing big data for water environment monitoring: Current status, challenges, and future prospects," *Earth's Future*, vol. 10, no. 2, p. e2021EF002289, 2022.

[2] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sensing of Environment*, vol. 301, p. 113924, 2024.

[3] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[4] D. Yu and C. Fang, "Urban remote sensing with spatial big data: A review and renewed perspective of urban studies in recent decades," *Remote Sensing*, vol. 15, no. 5, p. 1307, 2023.

[5] N. Casagli, E. Intrieri, V. Tofani, G. Gigli, and F. Raspini, "Landslide detection, monitoring and prediction with remote-sensing techniques," *Nature Reviews Earth & Environment*, vol. 4, no. 1, pp. 51–64, 2023.

[6] E. Benami, Z. Jin, M. R. Carter, A. Ghosh, R. J. Hijmans, A. Hobbs, B. Kenduiywo, and D. B. Lobell, "Uniting remote sensing, crop modelling and economics for agricultural risk management," *Nature Reviews Earth & Environment*, vol. 2, no. 2, pp. 140–159, 2021.

[7] T. Harrison and M. Strohmeyer, *Commercial space remote sensing and its role in national security*. Center for Strategic & International Studies, 2022.

[8] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[9] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "Nwpu-captions dataset and mlca-net for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[10] X. Huang, K. Lu, S. Wang, J. Lu, X. Li, and R. Zhang, "Understanding remote sensing imagery like reading a text document: What can remote sensing image captioning offer?" *International Journal of Applied Earth Observation and Geoinformation*, vol. 131, p. 103939, 2024.

[11] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8458–8472, 2021.

[12] H. Jiang, J. Yin, Q. Wang, J. Feng, and G. Chen, "Eaglevision: Object-level attribute multimodal llm for remote sensing," *arXiv preprint arXiv:2503.23330*, 2025.

[13] X. Xiao, Y. Zhang, T.-H. Nguyen, B.-T. Lam, J. Wang, L. Zhao, J. Hamm, T. Wang, X. Li, X. Wang *et al.*, "Describe anything in medical images," *arXiv preprint arXiv:2505.05804*, 2025.

[14] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[15] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[16] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023.

[17] I. A. Albadarneh, B. H. Hammo, and O. S. Al-Kadi, "Attention-based transformer models for image captioning across languages: An in-depth survey and evaluation," *Computer Science Review*, vol. 58, p. 100766, 2025.

[18] L. Zhou, Y. Zhang, Y.-G. Jiang, T. Zhang, and W. Fan, "Re-caption: Saliency-enhanced image captioning through two-phase learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 694–709, 2019.

[19] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image captioning with end-to-end attribute detection and subsequent attributes prediction," *IEEE Transactions on Image processing*, vol. 29, pp. 4013–4026, 2020.

[20] X. Ye, S. Wang, Y. Gu, J. Wang, R. Wang, B. Hou, F. Giunchiglia, and L. Jiao, "A joint-training two-stage method for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[21] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[22] S. Chang and P. Ghamisi, "Changes to captions: An attentive network for remote sensing change captioning," *IEEE Transactions on Image Processing*, vol. 32, pp. 6047–6060, 2023.

[23] X. Zhang, A. Jia, J. Ji, L. Qu, and Q. Ye, "Intra-and inter-head orthogonal attention for image captioning," *IEEE Transactions on Image Processing*, 2025.

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[25] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 296–26 306.

[26] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[27] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann *et al.*, "Phi-4 technical report," *arXiv preprint arXiv:2412.08905*, 2024.

[28] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[29] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840.

[30] D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, "Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model," in *European Conference on Computer Vision*. Springer, 2024, pp. 440–457.

[31] J. A. Irvin, E. R. Liu, J. C. Chen, I. Dormoy, J. Kim, S. Khanna, Z. Zheng, and S. Ermon, "Teochat: A large vision-language assistant for temporal earth observation data," *arXiv preprint arXiv:2410.06234*, 2024.

[32] C. Pang, X. Weng, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, S. Wang, L. Feng, G.-S. Xia *et al.*, "Vhm: Versatile and honest vision language model for remote sensing image analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6381–6388.

[33] L. Lian, Y. Ding, Y. Ge, S. Liu, H. Mao, B. Li, M. Pavone, M.-Y. Liu, T. Darrell, A. Yala *et al.*, "Describe anything: Detailed localized image and video captioning," *arXiv preprint arXiv:2504.16072*, 2025.

[34] Y.-L. Vu, D.-T. Duong, T.-B. Duong, A.-K. Nguyen, T.-H. Nguyen, L. T. P. Nguyen, J. Xing, X. Li, T. Wang, U. Bagci *et al.*, "Describe anything model for visual question answering on text-rich images," *arXiv preprint arXiv:2507.12441*, 2025.

[35] W. Lin, X. Wei, R. An, T. Ren, T. Chen, R. Zhang, Z. Guo, W. Zhang, L. Zhang, and H. Li, "Perceive anything: Recognize, explain, caption, and segment anything in images and videos," *arXiv preprint arXiv:2506.05302*, 2025.

[36] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[37] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[38] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[39] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International conference on computer, information and telecommunication systems (Cits)*. IEEE, 2016, pp. 1–5.

[40] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[41] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, "Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[42] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, "Skyscript: A large and semantically diverse vision-language dataset for remote sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5805–5813.

[43] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "Satclip: Global, general-purpose location embeddings with satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 4347–4355.

[44] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, "Position: Mission critical–satellite data is a distinct modality in machine learning," in *Forty-first International Conference on Machine Learning*, 2024.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[46] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[47] S. Banerjee and A. 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*, 2005, pp. 65–72.

[48] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on llm-as-a-judge," *arXiv preprint arXiv:2411.15594*, 2024.

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015.

[50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2015, pp. 3156–3164. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298935

[51] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2018, pp. 6077–6086. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00636

[52] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Manhattan, New York, USA: Springer International Publishing, 2018, pp. 711–727.

[53] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*. Palo Alto, California, USA: AAAI Press, 2018, pp. 6837–6845.

[54] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Manhattan, New York, USA: IEEE, 2019, pp. 4634–4643. [Online]. Available: http://doi.org/10.1109/ICCV.2019.00473

[55] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2020, pp. 10 971–10 980. [Online]. Available: http://doi.org/10.1109/CVPR42600.2020.01098

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2016, pp. 770–778. [Online]. Available: http://doi.org/10.1109/CVPR.2016.90

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[58] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2016, pp. 4651–4659. [Online]. Available: http://doi.org/10.1109/CVPR.2016.503

[59] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2016, pp. 203–212. [Online]. Available: http://doi.org/10.1109/CVPR.2016.29

[60] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2017, pp. 1141–1150. [Online]. Available: http://doi.org/10.1109/CVPR.2017.127

[61] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Manhattan, New York, USA: IEEE, 2017, pp. 4904–4912. [Online]. Available: http://doi.org/10.1109/ICCV.2017.524

[62] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, ser. Thematic Workshops '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 305–313. [Online]. Available: https://doi.org/10.1145/3126686.3126717

[63] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2017, pp. 3097–3106. [Online]. Available: http://doi.org/10.1109/CVPR.2017.330

[64] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Computer Vision – ECCV 2018*. Manhattan, New York, USA: Springer International Publishing, 2018, pp. 690–706. [Online]. Available: http://doi.org/10.1007/978-3-030-01246-5_41

[65] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2017, pp. 3107–3115. [Online]. Available: http://doi.org/10.1109/CVPR.2017.331

[66] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Manhattan, New York, USA: IEEE, 2017, pp. 1270–1279. [Online]. Available: http://doi.org/10.1109/ICCV.2017.142

[67] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2019, pp. 6612–6621. [Online]. Available: http://doi.org/10.1109/CVPR.2019.00678

[68] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2019, pp. 1969–1978. [Online]. Available: http://doi.org/10.1109/CVPR.2019.00207

[69] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2017, pp. 3298–3308. [Online]. Available: http://doi.org/10.1109/CVPR.2017.352

[70] Y.-S. Wang, C. Liu, X. Zeng, and A. Yuille, "Scene graph parsing as dependency parsing," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 397–407. [Online]. Available: http://doi.org/10.18653/v1/N18-1037

[71] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision – ECCV 2016*. Manhattan, New York, USA: Springer International Publishing, 2016, pp. 382–398. [Online]. Available: https://doi.org/10.1007/978-3-319-46454-1_24

[72] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, 2019, pp. 10 685–10 694. [Online]. Available: http://doi.org/10.1109/CVPR.2019.01094

[73] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[74] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[75] M. Barraco, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "The unreasonable effectiveness of clip features for image captioning: An experimental analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 4662–4670. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00512

[76] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou, "Xgpt: Cross-modal generative pre-training for image captioning," in *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Proceedings, Part I*. Berlin, Heidelberg: Springer, 2021, pp. 786–797. [Online]. Available: https://doi.org/10.1007/978-3-030-88480-2_63

[77] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021. [Online]. Available: https://arxiv.org/abs/2111.09734

[78] J. Gu, S. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Manhattan, New York, USA: Springer International Publishing, 2018, pp. 519–535. [Online]. Available: http://doi.org/10.1007/978-3-030-01246-5_31

[79] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Manhattan, New York, USA: IEEE, Jun. 2019, pp. 4125–4134. [Online]. Available: http://doi.org/10.1109/CVPR.2019.00425

[80] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, "Improving image captioning with conditional generative adversarial nets," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8142–8150, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4823

[81] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, "Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

[82] C. Liu, K. Chen, R. Zhao, Z. Zou, and Z. Shi, "Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model," *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[83] Z. Li, W. Zhao, X. Du, G. Zhou, and S. Zhang, "Cross-modal retrieval and semantic refinement for remote sensing image captioning," *Remote Sensing*, vol. 16, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/1/196

[84] M. Singha, A. Jha, B. Solanki, S. Bose, and B. Banerjee, "Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[85] C. Mao and J. Hu, "ProGEO: Generating Prompts through Image-Text Contrastive Learning for Visual Geo-localization," *arXiv e-prints*, p. arXiv:2406.01906, Jun. 2024.

[86] U. Mall, C. P. Phoo, M. K. Liu, C. Vondrick, B. Hariharan, and K. Bala, "Remote sensing vision-language foundation models without annotations via ground remote alignment," 2023.

[87] S. Mo, M. Kim, K. Lee, and J. Shin, "S-CLIP: Semi-supervised vision-language learning using few specialist captions," in *Advances in Neural Information Processing Systems*, vol. 36, 2024, pp. 61 187–61 212.

[88] X. Li, C. Wen, Y. Hu, and N. Zhou, "Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103497, 2023.

[89] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[90] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.

[91] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, J. Li, and X. Mao, "Earthmarker: A visual prompting multi-modal large language model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[92] J. Luo, Z. Pang, Y. Zhang, T. Wang, L. Wang, B. Dang, J. Lao, J. Wang, J. Chen, Y. Tan *et al.*, "Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding," *arXiv preprint arXiv:2406.10100*, 2024.

[93] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025.

[94] Y. Li, W. Xu, G. Li, Z. Yu, Z. Wei, J. Wang, and M. Peng, "Unirs: Unifying multi-temporal remote sensing tasks through vision language models," *arXiv preprint arXiv:2412.20742*, 2024.

[95] L. Li, "Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 513–522.

[96] R. Ding, B. Chen, P. Xie, F. Huang, X. Li, Q. Zhang, and Y. Xu, "Mgeo: Multi-modal geographic language model pre-training," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 185–194.

[97] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *arXiv preprint arXiv:2311.07113*, 2023.

[98] J. Wang, Z. Zheng, Z. Chen, A. Ma, and Y. Zhong, "Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 5481–5489.

[99] M. Zhang, F. Chen, and B. Li, "Multistep question-driven visual question answering for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[100] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022.

[101] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

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

[103] Y. Yang and S. 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*, 2010, pp. 270–279.

[104] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 28–37.

[105] Y. Zhan, Z. Xiong, and Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[106] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "Rrsis: Referring remote sensing image segmentation," *arXiv preprint arXiv:2306.08625*, 2023.

[107] Y. Zhan, Z. Xiong, and Y. Yuan, "Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 221, pp. 64–77, 2025.

[108] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[109] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, and L. Huang, "Tinyllava: A framework of small-scale large multimodal models," *arXiv preprint arXiv:2402.14289*, 2024.

[110] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018.

[111] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[112] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[113] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.