

---

# GeoLLaVA-8K: Scaling Remote-Sensing Multimodal Large Language Models to 8K Resolution

---

Fengxiang Wang<sup>1</sup>, Mingshuo Chen<sup>2</sup>, Yueying Li<sup>1</sup>, Di Wang<sup>3,4</sup>, Haotian Wang<sup>1</sup>,  
Zonghao Guo<sup>5</sup>, Zefan Wang<sup>5</sup>, Boqi Shan<sup>6</sup>, Long Lan<sup>1</sup>, Yulin Wang<sup>5,\*</sup>,  
Hongzhen Wang<sup>5,\*</sup>, Wenjing Yang<sup>1,\*</sup>, Bo Du<sup>3,4</sup>, Jing Zhang<sup>3,\*</sup>

<sup>1</sup> College of Computer Science and Technology, National University of Defense Technology, China

<sup>2</sup> Beijing University of Posts and Telecommunications, China

<sup>3</sup> School of Computer Science, Wuhan University, China

<sup>4</sup> Zhongguancun Academy, China <sup>5</sup> Tsinghua University, China <sup>6</sup> Beihang University, China

## Abstract

Ultra-high-resolution (UHR) remote sensing (RS) imagery offers valuable data for Earth observation but pose challenges for existing multimodal foundation models due to two key bottlenecks: (1) limited availability of UHR training data, and (2) token explosion caused by the large image size. To address data scarcity, we introduce **SuperRS-VQA** (avg.  $8,376 \times 8,376$ ) and **HighRS-VQA** (avg.  $2,000 \times 1,912$ ), the highest-resolution vision-language datasets in RS to date, covering 22 real-world dialogue tasks. To mitigate token explosion, our pilot studies reveal significant redundancy in RS images: crucial information is concentrated in a small subset of object-centric tokens, while pruning background tokens (e.g., ocean or forest) can even improve performance. Motivated by these findings, we propose two strategies: *Background Token Pruning* and *Anchored Token Selection*, to reduce the memory footprint while preserving key semantics. Integrating these techniques, we introduce **GeoLLaVA-8K**, the first RS-focused multimodal large language model capable of handling inputs up to  $8K \times 8K$  resolution, built on the LLaVA framework. Trained on SuperRS-VQA and HighRS-VQA, GeoLLaVA-8K sets a new state-of-the-art on the XLRs-Bench. Datasets and code were released at [GeoLLaVA-8K](#).

## 1 Introduction

With the rapid development of Earth science, the collection, process, and representation of remote sensing (RS) data have become increasingly important. [1]. Among these data sources, satellite imagery is able to capturing extensive spatiotemporal information about Earth’s surface [2, 3], significantly enhancing our geographic understanding of this planet.

Recent advances in multimodal large language models (MLLMs) [4, 5, 6, 7, 8] have significantly improved visual understanding and reasoning, simultaneously facilitating remarkable scientific progress in geoscience for handling remote sensing data [9, 10, 11]. However, despite significant progress of MLLMs across both general and RS domains, current models still fall short in addressing real-world RS tasks, especially in the case of **ultra-high-resolution (UHR) scenarios**. For instance, even leading models like GPT-4o [12] and Qwen-VL series [13] are limited up to 4K resolution, resulting in limited performance (Accuracy < 0.45) on the XLRs-Bench [14], a recent RS evaluation benchmark featuring large image size (e.g., 8K-10K). This gap raises an urgent problem we aim to investigate in this paper:

---

\*Corresponding authors

Can we develop MLLMs tailored for RS data with resolutions far beyond current limits (e.g., up to  $8K \times 8K$ )?

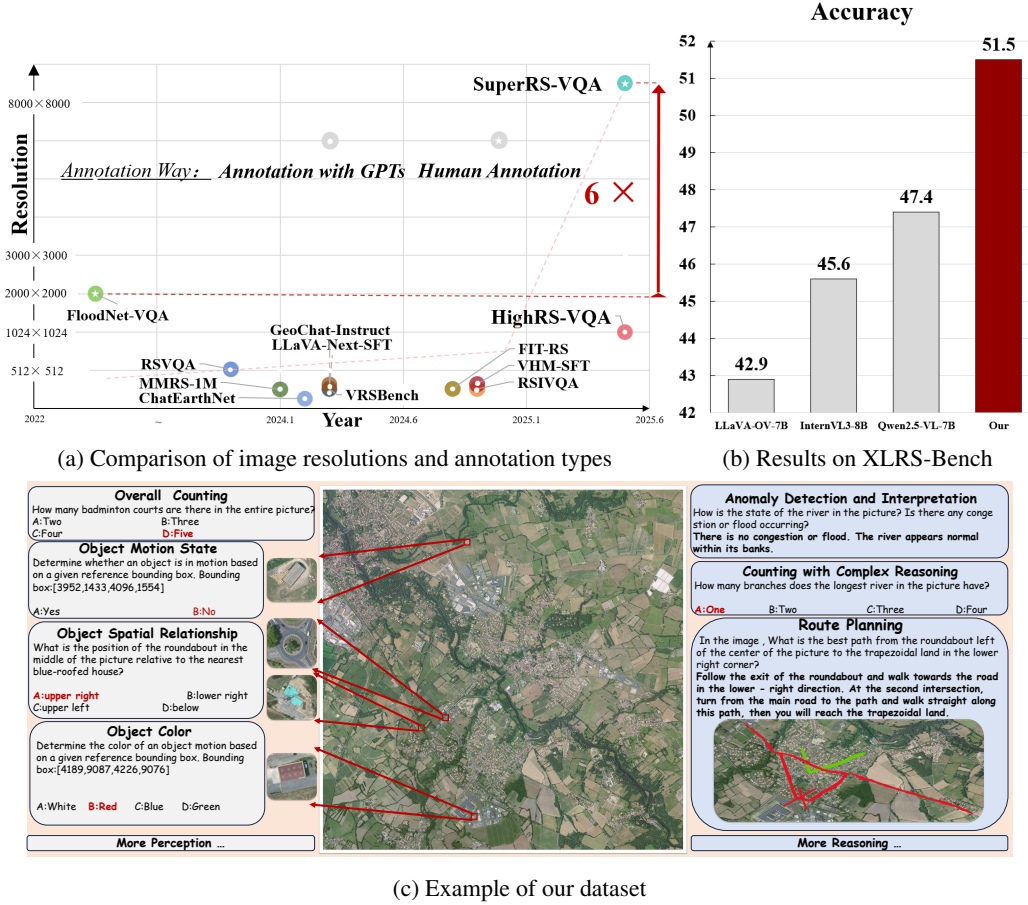


Figure 1: We introduce SuperRS-VQA and HighRS-VQA, the highest-resolution VQA datasets for MLLM training. GeoLLaVA-8K, trained on these datasets, significantly outperforms existing MLLMs on ultra-high-resolution remote sensing tasks.

In response to this question, we identify two critical challenges that remain underexplored: **(1) The lack of image-text training data for UHR RS images;** **(2) The massive number of visual tokens in UHR RS imagery increases training difficulty for MLLMs.** To address the first issue, we introduce two novel multimodal RS datasets featuring large image sizes: SuperRS-VQA (about  $8K \times 8K$ ) and HighRS-VQA (about  $2K \times 2K$ ). To our knowledge, they are the largest image-size RS vision-language datasets to date, covering 22 real-world subtasks and significantly surpassing previous RS image-text datasets in both scale and diversity.

For the second problem, from our intuition, the excessive sequence length of visual tokens may lead to two major issues in MLLM’s training: **(1) Expensive Computation Overhead:** Current models are not designed to operate at such scales. For example, directly adapting LLaVA to  $8K \times 8K$  inputs results in memory overflow. **(2) Low Semantic Density:** UHR RS imagery contains an abundance of homogeneous background tokens, which contribute little useful information. In contrast, semantically rich foreground tokens are sparse and risk being overlooked without effective token management.

To tackle these challenges, we propose two token compression strategies: Background Token Pruning and Anchored Token Selection, which focus on token aggregation and refinement to efficiently manage the large number of visual tokens introduced by UHR inputs. Building on these techniques, we develop GeoLLaVA-8K, a UHR-oriented, RS-specific MLLM capable of processing inputs up to 8K resolution. Experiments on large image-size benchmarks (e.g., XLRS-Bench) demonstrate that our model outperforms all existing open- and closed-source MLLMs, setting a new state-of-the-art.

In summary, our main contributions are as follows:

- (1) We curate SuperRS-VQA and HighRS-VQA, two RS image-text datasets covering 22 real-world subtasks for UHR scenes, featuring so far the largest image sizes in our knowledge.
- (2) We investigate the challenges in training MLLMs caused by the excessive visual tokens in UHR RS imagery. To address the problems of redundant backgrounds and the sparsity of semantically rich objects, we propose two strategies for token pruning and selection.
- (3) We develop GeoLLaVA-8K, the first RS MLLM tailored for UHR scenes, capable of processing inputs up to 8K. Experiments on representative UHR RS benchmarks demonstrate its superior performance compared to existing open- and closed-source MLLMs.

## 2 Related Work

**Remote Sensing Multimodal Datasets.** With the rapid progress of large multimodal models in general domains, the RS field has witnessed significant advancements in MLLM development [9, 10, 11, 15, 16], leading to the emergence of several RS-specific vision-language datasets. RSVQA [17] comprises image/question/answer triplets, with questions and answers derived from OpenStreetMap (OSM). RSIVQA [18] generates samples automatically using existing scene classification and object detection datasets. RSSA [19] targets hallucination, while FIT-RSRC [16] focuses on object relationship understanding. VRSBench [20] includes 29,614 images, 52,472 object references, and 123,221 QA pairs. Recent datasets for the SFT stage of MLLM training include GeoChat\_Instruct [9], ChatEarthNet [21], VHM\_sft [22], FIT-RS [16], and MMRS-1M [15]. However, most of these datasets are aggregated from existing sources and contain limited original annotations. Critically, their average image resolution remains below  $1K \times 1K$ , which is insufficient for real-world UHR RS tasks. Recently, XLRs-Bench [14] introduces the largest image-size RS benchmark to date (average  $8,500 \times 8,500$ ) for evaluation purposes only, underscoring the persistent lack of corresponding UHR training data in the RS MLLM field.

**Multimodal Large Language Model.** Leveraging advanced large language models (LLMs) like GPTs [5] and LLaMA [6], MLLMs have demonstrated strong visual understanding and reasoning capabilities [23, 24]. Proprietary models such as Gemini [4] and GPT-4o [5], along with open-source alternatives like Qwen-VL [8], InternLM-XComposer [25], MiniCPM [26], LLaVA [27], and MiniGPT-4 [7], show competitive performance but are typically limited to input resolutions of 2K-4K. To overcome this limitation, LLaVA-Next [28] processes images in patches and connects them via global tokens, Monkey [29] and LLaVA-UHD [30] compress patches to reduce redundancy, Cambrian [31] uses learnable queries for multi-scale interaction, and SLiME [32] applies dual compression to preserve global context. In the RS domain, several MLLMs have been developed. For example, GeoChat [9] enables multi-task RS dialogue based on the LLaVA-1.5 framework [33]. EarthGPT [15] unifies multisensor interpretation tasks by converting RS annotations into question-answer pairs. Meanwhile, LHRs-Bot [10] improves vision-language understanding through multi-level alignment and curriculum learning. Nevertheless, both general-domain and RS MLLMs struggle to effectively handle UHR RS imagery. Universal MLLMs lack domain adaptation and are unable to meet the higher resolution requirements of real-world RS applications. Although RS-specific MLLMs incorporate domain knowledge, they still operate at significantly lower resolutions, i.e., less than 1K, highlighting the urgent need for a UHR ( $8K \times 8K$ ) and domain-aligned MLLMs in the RS field.

## 3 Dataset

The RS domain currently lacks sufficient UHR image-text training data for MLLM development. To validate this judgment, we firstly unified existing RS image-text datasets with general-domain data in a 2:1 ratio for supervised fine-tuning (SFT). However, we observed a notable performance drop in LLaVA-Next-2K [34] on UHR benchmarks (XLRs-Bench [14]) in the Tab. 1, which we attribute to the limited resolution of current RS training datasets, highlighting the urgent need for UHR image-text pairs to support real-world RS applications.

To address this, following XLRs-Bench [14], we first manually annotated 12K UHR samples. Notably, existing MLLMs such as GPT-4o [12] either encounter memory overflow or generate low-quality outputs when processing UHR RS imagery. Therefore, we adopt a fully manual annotation

approach in this work. Nevertheless, due to the high time and labor costs, scaling manual annotation of UHR data is impractical. To address this, we develop a semi-automated annotation pipeline that leverages GPT-4o [12] alongside existing RS detection and segmentation datasets, generating 100K medium-to-high-resolution (MHR,  $2K \times 2K$ ) image-text pairs. To bridge the distribution gap between MHR and UHR data, we further apply an influence-based data selection method built on the LESS framework [35]. Next, we detailed the procedure of constructing datasets.

**Source Screening.** Our datasets have diverse and extensive image sources. We follow a key principle: to avoid using the same data sources as existing vision-language datasets as far as possible, ensuring diversity and effectiveness in training data. Specifically, we utilize datasets such as DeepGlobe [36], STAR [37], FAIR1M 2.0 [38], LoveDA [39], Inria [40], OpenSatMap [41], HRSCD [42], MiniFrance [3], and DOTA [2]. To minimize redundancy of image-text pairs, we deduplicate images within these datasets and remove overlaps with existing benchmark datasets like XLRB-Bench [14]. Nonetheless, it should be noted that, since generating MHR image-text pairs requires leveraging existing annotations with tools like GPT-4o, the overlap with prior VQA datasets is unavoidable.

**Annotation.** For UHR data, we formed a team of 5 MLLM and RS experts (Ph.D. holders or candidates) and 30 crowd-sourcing annotators (undergraduate and masters students). All samples were manually labeled and rigorously cross-validated over 40 days of annotation and 10 days of verification, resulting in a 12K UHR image-text dataset, named SuperRS-VQA. For MHR data, we developed a semi-automated annotation. Using task-specific prompts and existing annotations (e.g., bounding boxes in RS detection datasets), we generated text via GPT-4o [12]. Despite high token costs (>\$1,000), outputs still needed to be quality-checked by annotators.

**Data Selection Pipeline for MHR Data.** We further adopted an influence-based data selection pipeline (see Fig. 3) to improve the relevance of our dataset to UHR downstream tasks and ensure its cultivation of reasoning capabilities for models fine-tuned on it. This pipeline quantifies the exact contribution of each training data sample by measuring the change in validation performance when that sample is removed (i.e., a Leave-One-Out influence score). Concretely, we leverage gradient information from a warm-up fine-tuned model  $\theta$  trained on the crude dataset: for each candidate  $z \in Z$  among training examples and validation examples  $x \in X$ , we compute

$$\text{Inf}_{\text{SGD}}(z_i, X) = \max_{x \in X} \cos \langle \nabla \ell_\theta(z_i), \nabla \ell_\theta(x_i) \rangle \quad (1)$$

where  $\ell_\theta$  is the negative log-likelihood loss on a data sample. The above estimation is a simplified but empirically valid form of *Influence Function* [43], a widely used data valuation method which is a first-order approximation of performance change. Intuitively, a higher  $\text{Inf}_{\text{SGD}}$  indicates removing  $z_i$  from the training set would cause a larger, more directionally coherent shift in validation loss, and hence that  $z_i$  is more influential. Unlike heuristic selection methods based on the apparent semantics of VQA pairs, which often require laborious task-specific rule engineering, our approach directly measures data influence utilizing the preferences of a fine-tuned model.

We follow LESS [35] for a practical and efficient implementation of our method. The main challenge of using influence-based methods is gradient storage consumption, despite the Hessian calculation is already omitted in our formulation. Therefore, only gradients of LoRA adapters are calculated. Moreover, the gradients are projected to 8192-dimensional subspace via a fixed random matrix. We

**Table 1: Pilot Experiments of Dataset.** We keep the original pretraining data of LLaVA-Next-2K [34] (a variant of LLaVA-Next [28]) and we only modify the SFT (supervised fine-tuning) data in the second training phase. Accuracy is reported based on results from XLRS-Bench [14].

| Model      | SFT Data               |        |                    | Acc.  |
|------------|------------------------|--------|--------------------|-------|
|            | Data Source            | Volume | Average Resolution |       |
| LLaVA-Next | LLaVA-Next-2K_SFT [28] | 738K   | 512×512            | 44.6  |
| LLaVA-Next | LLaVA-Nex-2K_SFT [28]  | 738K   | 512×512            | 40.4↓ |
|            | GeoChat_Instruct [9]   | 283k   | 632×619            |       |
|            | ChatEarthNet [21]      | 6k     | 256×256            |       |
|            | VHM_sft [22]           | 150k   | 643×638            |       |
|            | FTT-RS [16]            | 1200k  | 512×512            |       |
|            | MMRS-1M [15]           | 308K   | 499×491            |       |
|            | VRSBench-train [20]    | 85k    | 512×512            |       |

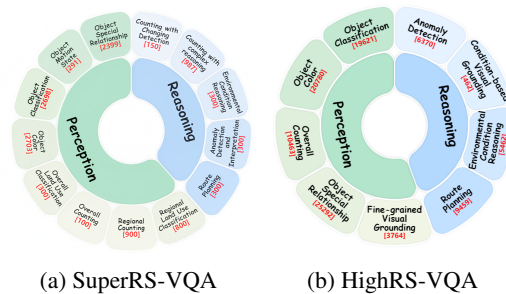


Figure 2: Our datasets have the perception and reasoning dimensions across 22 sub-tasks.



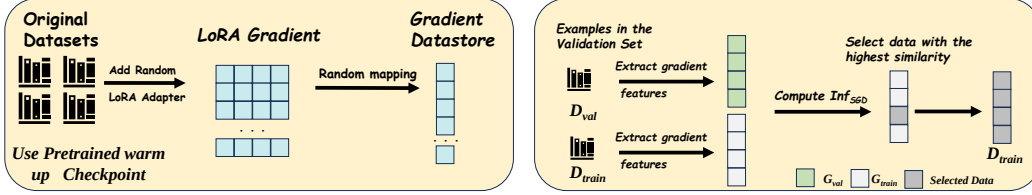


Figure 3: **Data Selection Pipeline for MHR Data.** In step 1, we train a warmup model on the crude dataset to acquire gradient features for both the training and validation sets. In step 2, we match training data candidates with the validation set and select the most influential samples.

note that this leverages the Johnson-Lindenstrauss Lemma [44] that random projection preserves the inner product of gradients, and the recent empirical validation [45] for data selection. The cosine similarity is used instead of the original inner product in common influence definitions because the gradient norm of a language model may have numerical biases [35]. After profiling gradients of the crude training set and the validation set, we rank and retain the top 70% most influential samples for downstream fine-tuning.

**Statistics.** Fig. 1 highlights the advantages and key details of our dataset, which offers the highest resolution among all RS image-text training datasets to date. Fig. 2 visualizes the capability dimensional coverage of both datasets. It can be seen that our datasets span a wider range of task dimensions, showing strong alignment with real-world scenarios. Finally, we summarize the key statistics of our datasets in Tab. 2, including average question/answer lengths, the diversity and quantity of annotated objects, average resolution, and other relevant attributes.

## 4 Method

Before training on our datasets, we investigate how the Low Semantic Density characteristic, i.e., excessive length visual tokens caused by UHR RS imagery, affects the performance of MLLMs. Specifically, Sections 4.1 and 4.2 separately analyze the low information density of RS imagery from two views: Section 4.1 reveals the redundancy of background tokens and their negative impact on MLLM modeling, while Section 4.2 explores the correlation between scarce objects visual tokens. In Section 4.3, we propose a targeted solution to overcome these challenges.

### 4.1 Overwhelming Background Tokens Hinder MLLM Fine-Tuning on RS Data

*Do background tokens dominate UHR RS imagery?*

In order to learn about a holistic recognition for the background ratio in UHR RS scenes, we conduct preliminary experiments, where the LLaVA-1.5 [33] is employed. Specifically, we randomly selected 100 images from the XLRs-Bench [14], covering natural backgrounds such as sea surfaces, forests, and fields, to provide a diverse set of test samples for the evaluation of background redundancy. Then, we extract  $64 \times 64$  small images through non-overlapped sliding windows, and sequentially input each into the

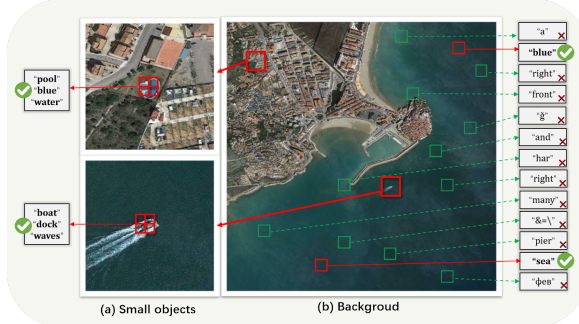


Figure 4: Examples of tokens and their positions that the *logit lens* yields in the late layers.

model. We evaluated the background rate of the UHR RS scene using two methods, following the previous work [46]: **Generative Description**: We prompt the model with “Describe the image” and analyze the generated description to extract key information. **Binary Polling**: We ask the model “Is this image mainly background?” where the background is defined as low-information natural areas (e.g., sea, deserts, vegetation), excluding artificial structures (e.g. buildings, roads, urban areas).

We developed a multi-level semantic parsing framework to precisely quantify MLLM’s responses. The framework separates the description from the evaluation content and processes it in five steps: Urban Semantic Recognition, Human-made Structure Analysis, Natural Element Classification, Semantic Categorization, and Probability Gradient Quantification. Details are shown in the appendix. Results show background coverage in RS images reaches up to **73.14%**.

To better understand the semantic information of visual tokens, we use the *logit lens* technique [47]. For each layer, we decode the activation at each token position via unembedding. Details are shown in the appendix. With the *logit lens* technique, we found that in large background areas, only a small portion of the information in the visual tokens is effectively mapped to specific semantic scenes, while most of the background tokens do not align with clear semantic representations (Fig. 4(b)).

*Do background tokens in RS imagery hinder MLLMs from effectively modeling UHR satellite images?*

Then, we feed the whole UHR RS imagery into MLLM, where the LLaVA-Next-2K [34], a variant of LLaVA-Next [28], is adopted to support higher resolution as far as possible. Initially, we directly perform the inference on XLRS-Bench [14], whose images have an average resolution of  $8K \times 8K$ . Naturally, the resulting long visual token sequences lead to out-of-memory (OOM) errors, aligning with the first challenge presented in Section 1.

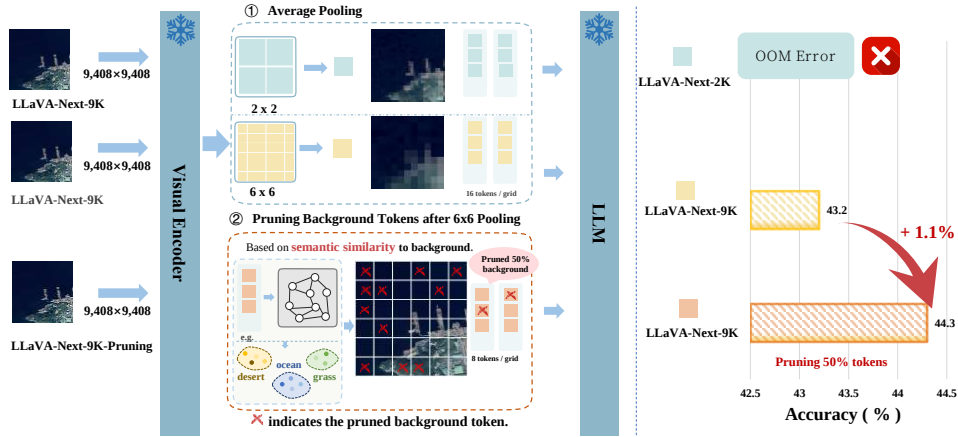


Figure 5: Halving the background tokens by half even resulted in improved performance.

To address the OOM issue, we increase the pooling layer size from the default  $2 \times 2$  to  $6 \times 6$ , allowing LLaVA-Next-2K to process  $9K \times 9K$  resolution, called LLaVA-Next-9K. To further reduce redundant background information, we use a background-aware token selection strategy. By calculating semantic similarity between visual tokens and typical background terms (e.g., ocean, desert), we score and rank tokens, then discard the top 50% most redundant ones. The process involves: (1) extracting embedding vectors for typical background terms; (2) calculating each token’s highest similarity score with these embeddings as its background score; and (3) ranking tokens by score and removing the top 50%. This approach reduce the background token count. The result is referred to as LLaVA-Next-9K-Pruning.

As Fig. 5 shows, LLaVA-Next-9K underperforms due to its lack of exposure to 9K-resolution training data, emphasizing the need for UHR RS datasets. Surprisingly, despite LLaVA-Next-9K-Pruning uses only half the tokens of LLaVA-Next-9K, it achieves better accuracy. This suggests that excessive background tokens not only introduce additional computational overhead but also impair performance. These findings underscore a critical insight: reducing background redundancy is critical to perform effective modeling under UHR RS images.

From the above two experiments, we confirm that at the token level, RS images exhibit significant background redundancy, *where key target information constitutes only a small fraction of the image, while most regions lack explicit semantic significance*. This presents a major challenge to the efficiency of multimodal models.

## 4.2 Scarce Object Tokens Drive MLLM Fine-Tuning on RS Data

After recognizing the background redundancy of UHR RS imagery in visual tokens, we turn to the second key aspect of understanding RS’s low semantic density: the localization of critical information.

*Whether essential information is concentrated in small targets and captured by corresponding visual tokens?*

We conduct ablation experiments to test whether the foreground information is concentrated in specific visual tokens. By removing selected tokens and observing the drop in recognition performance, we assess their importance. The images are sourced from XLRs-Bench, with two pre-processing steps applied to ensure reliability: High signal-to-noise sub-images and Hallucination control. Details are shown in the appendix. Following this procedure, the final dataset includes 1,189 VQA pairs. Fig. 6 provides an overview of our ablation. Details are shown in the appendix.

We define the token subset  $S$  for ablation using four settings: (1) **Object Tokens**: Tokens aligned with image patches that originally contain the target object; (2) **Object Tokens with Buffer**: Object tokens along with their surrounding tokens. (3) **Register Tokens**: Tokens whose norms deviate by more than two standard deviations from the mean, corresponding to the register tokens identified as encoding global image features [48]; (4) **Random Tokens**: A baseline in which  $n$  tokens are randomly ablated.

Tab. 3 shows that ablating object tokens significantly hinders the models ability to recognize targets. Generative decrease and VQA decrease experiments are following the previous work [46]. A larger percentage drop in performance indicates a greater impact of the ablation, meaning the model is more likely to answer incorrectly, thus suggesting that the ablated tokens contain more localized object information. Notably, with a comparable number of ablated tokens, removing object tokens consistently results in a larger performance degradation than random ablation, highlighting the precise localization of object-specific information. *logit lens* analysis [47] further reveals that visual tokens for scarce objects efficiently converge to accurate semantic representations (Fig. 4 (a)). Collectively, these findings underscore a key insight in RS: object tokens not only encode critical information but also align closely with actual visual features.

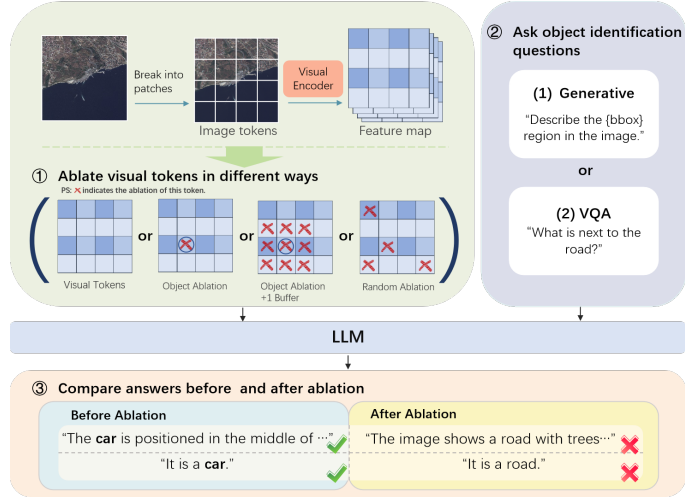


Figure 6: **Overview of object tokens ablation experiments.**

① We ablate some visual tokens that potentially contain object-specific information, ② prompt the model to describe the image, or answer objectspecific questions, then ③ measure the impact of token ablation by calculating the percentage of initially correct object identifications that become incorrect after ablation.

Table 3: **Performance degradation after token ablation.** “Avg. token counts” denotes the average number of tokens ablated per image for LLaVA-1.5.

| Ablation Type<br>(Avg. Token Count) | Generative<br>Decrease (%) | VQA<br>Decrease (%) |
|-------------------------------------|----------------------------|---------------------|
| Object (26.5)                       | 34.9                       | 24.8                |
| +1 Buffer (50.2)                    | 44.8                       | 32.6                |
| +2 Buffer (81.8)                    | 51.1                       | 40.6                |
| Register tokens (3.6)               | 18.1                       | 9.7                 |
| Random (5)                          | 6.7                        | 1.1                 |
| Random (30)                         | 14.5                       | 4.4                 |
| Random (50)                         | 17.5                       | 7.1                 |
| Random (80)                         | 25.6                       | 9.9                 |

### 4.3 Background Token Pruning and Anchored Token Selection

Through the above analysis, we have identified the Low Semantic Density challenge in UHR RS imagery, characterized by both background and foreground tokens, which poses a significant obstacle in training MLLMs. Based on these findings, we argue that an RS-specific MLLM capable of efficiently handling UHR imagery is both necessary and feasible. As shown in Fig. 7, we propose a two-step token selection strategy for background and object tokens in RS imagery. Specifically, building on our two high-resolution RS datasets: SuperRS-VQA and HighRS-VQA, we SFT existing MLLMs, where we select LLaVA-Next-2K [34], to create RS MLLMs. In practice, we initialize from LLaVA-Next-2K’s general-domain pretrained weights.

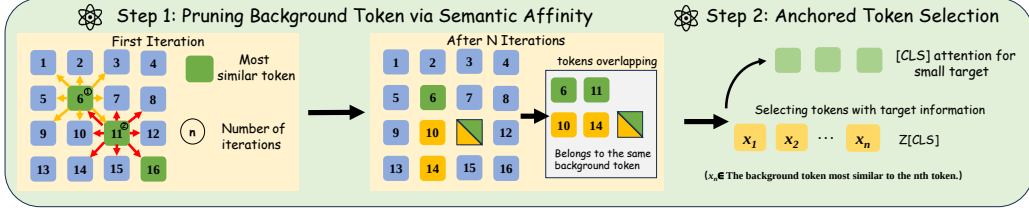


Figure 7: Our two-step method processes visual tokens during MLLMs SFT stage, between the visual encoder and the projection layer.

**Background Token Pruning via Semantic Affinity.** To address the redundancy of background tokens in visual token sequences, we propose an adaptive token clustering strategy for compressing background tokens. Specifically, we construct token-to-token associations by assigning each token  $\mathbf{p} = (u, v)$  to a neighboring token  $s$  with probability  $q_s(\mathbf{p})$ . Rather than applying this globally, we restrict associations to the local neighborhood  $\mathcal{N}_{\mathbf{p}}$ , ensuring:

$$\sum_{s \in \mathcal{N}_{\mathbf{p}}} q_s(\mathbf{p}) = 1 \quad (2)$$

After performing this computation once, we iteratively reapply the same process to the selected tokens for  $N$  steps, gradually forming an initial cluster. We assume that background tokens, such as those containing ocean features, exhibit high similarity and thus are suited for clustering. We perform this process for all tokens. Note that the clusters formed during later iterations may overlap with those from earlier steps. In such cases, we merge the overlapping clusters into a single one. This adaptive clustering strategy allows us to effectively group similar background tokens, e.g., ocean-related tokens, into one irregular cluster, and forest-related tokens into another, thus compressing redundant visual information more effectively.

**Anchored Token Selection for Scarce Object Retaining.** To prevent small-object or otherwise informative tokens from being lost in a background-oriented pruning stage, we introduce Anchored Token Selection (ATS). ATS leverages the attention map formed between the pretrained ViT’s [49] [CLS] token and the remaining image tokens after background pruning. Tokens receiving higher [CLS] to patch attention are deemed more semantically important and are kept, as they likely correspond to informative objects. Specially, the attention map  $\mathbf{a}_{[\text{CLS}]} \in \mathbb{R}^{1 \times n}$  from the [CLS] token  $\mathbf{z}_{[\text{CLS}]} \in \mathbb{R}^{1 \times d}$  to other patch tokens  $\mathbf{Z}_v \in \mathbb{R}^{n \times d}$  is computed by

$$\mathbf{a}_{[\text{CLS}]} = \text{Softmax} \left( \frac{\mathbf{z}_{[\text{CLS}]} \mathbf{W}_Q (\mathbf{Z}_v \mathbf{W}_V)^T}{\sqrt{d}} \right) = \text{Softmax} \left( \frac{\mathbf{q}_{[\text{CLS}]} \mathbf{K}_v^T}{\sqrt{d}} \right), \quad (3)$$

where  $n$  is the number of remaining image tokens,  $d$  is the dimension of hidden states, and  $\mathbf{W}_Q, \mathbf{W}_V$  are the query and key projection matrices of this encoder layer. Note that we utilize the attention map at the second-to-last layer of the visual encoder (*i.e.*, the output layer of CLIP-ViT in LLaVA-Next [28]). Given a compression ratio  $r$ , we calculate the final number of tokens to retain as  $R = n \times r$ . Ultimately, we select the top  $R$  tokens with the highest attention scores from the attention map  $\mathbf{a}_{[\text{CLS}]}$ , obtaining the final retained tokens  $\mathbf{Z}'_v \in \mathbb{R}^{R \times d}$ . After passing through a multi-modal projector  $g$ , the remaining image tokens are concatenated with language instructions  $\mathbf{H}_q$  and fed into the language model with trainable parameters  $f_\phi$  to generate the response  $\mathbf{X}_a$  with  $L$  text tokens in the auto-regression manner, where the probability of  $\mathbf{X}_a$  is computed by  $p(\mathbf{X}_a | g(\mathbf{Z}'_v), \mathbf{H}_q) = \prod_{i=1}^L f_\phi(\mathbf{x}_i | g(\mathbf{Z}'_v), \mathbf{H}_q, \mathbf{X}_{a < i})$ .



## 5 Experiments

We perform SFT training on the SuperRS-VQA and HighRS-VQA datasets, with a brief overview of the training details and results in this section. Exploratory and ablation studies are presented in Section 3 to clarify the research motivation. **Additional details, ablation studies of dataset and method, as well as case analyses, are provided in the appendix.**

### 5.1 Main Results

**Experimental Setup.** We use XLRs-Bench [14] for evaluation. The MLLMs evaluated on XLRs-Bench are grouped into three categories: (a) open-source MLLMs; (b) closed-source MLLMs and (c) the specialized RS model. For fair comparison, we used a zero-shot setting with uniform prompts for all MLLMs, including our work. The appendix details the architecture and parameter sizes of each open-source MLLMs, and includes additional results across various settings. Except for GeoChat which was evaluated using its native framework, all other models were evaluated using LMMS-Eval [50, 51]. Following XLRs-Bench [14], we evaluated the accuracy and reported of L-1 dimension for the VQA task, with L-3 and L-4 results available in the appendix.

Table 4: **Experimental results on the perception and reasoning dimensions on XLRs-Bench, with models ranked by average performance.** 'Avg.' represents the average accuracy across sub-tasks. We mark the highest score in **red**. The full name of sub-tasks can be found in the appendix.

| Method                      | Perception |      |      |      |      |      |      |      | Reasoning |      |      |      |      |      | Avg. |
|-----------------------------|------------|------|------|------|------|------|------|------|-----------|------|------|------|------|------|------|
| Sub-tasks (L-3 Capability)  | OC         | RC   | OLUC | RLUC | OCC  | OCL  | OMS  | OSR  | AD        | ECR  | RP   | RCCD | CCR  |      |      |
| Remote Sensing MLLMs        |            |      |      |      |      |      |      |      |           |      |      |      |      |      |      |
| GeoChat [9]                 | 16.7       | 29.0 | 2.0  | 23.0 | 21.1 | 16.8 | 35.0 | 24.2 | 33.0      | 43.0 | 10.0 | -    | 21.0 | 22.9 |      |
| Closed-source MLLMs         |            |      |      |      |      |      |      |      |           |      |      |      |      |      |      |
| GPT-4o [12]                 | 25.0       | 32.0 | 15.0 | 66.0 | 9.5  | 11.3 | 11.7 | 24.6 | 73.0      | 73.0 | 35.0 | 20.0 | 25.0 | 32.4 |      |
| GPT-4o-mini [52]            | 23.3       | 25.0 | 19.0 | 59.5 | 40.9 | 31.0 | 65.0 | 23.6 | 71.0      | 71.0 | 29.0 | 6.7  | 30.0 | 38.1 |      |
| Claude 3.7 Sonnet [53]      | 27.6       | 22.7 | 17.4 | 68.4 | 30.5 | 29.9 | 63.6 | 27.6 | 64.8      | 78.4 | 34.5 | 27.8 | 32.6 | 40.5 |      |
| Gemini 2.0 Flash [4]        | 41.7       | 45.0 | 38.0 | 73.5 | 34.6 | 27.6 | 61.7 | 32.0 | 73.0      | 82.0 | 43.0 | 30.0 | 51.0 | 48.7 |      |
| Open-source MLLMs           |            |      |      |      |      |      |      |      |           |      |      |      |      |      |      |
| InternLM-XComposer-2.5 [54] | 21.7       | 42.0 | 7.0  | 68.0 | 31.8 | 27.8 | 6.7  | 26.0 | 72.0      | 81.0 | 41.0 | 36.7 | 47.0 | 39.1 |      |
| LLaVA-Next [28]             | 26.7       | 40.0 | 5.0  | 67.0 | 28.8 | 32.8 | 66.7 | 30.0 | 69.0      | 78.0 | 27.0 | 35.0 | 36.0 | 41.7 |      |
| LLaVA-OneVision-7B [55]     | 25.0       | 38.0 | 2.0  | 69.5 | 35.9 | 35.3 | 65.0 | 25.2 | 76.0      | 83.0 | 24.0 | 43.3 | 36.0 | 42.9 |      |
| InternVL3-8B [56]           | 40.0       | 39.0 | 10.0 | 71.5 | 44.5 | 30.8 | 65.0 | 25.2 | 77.0      | 82.0 | 36.0 | 21.7 | 50.0 | 45.6 |      |
| Qwen2-VL-7B [13]            | 26.7       | 40.0 | 11.0 | 73.0 | 35.9 | 34.6 | 61.7 | 31.8 | 70.0      | 81.0 | 35.0 | 46.7 | 48.0 | 45.8 |      |
| LLaVA-OneVision-72B [55]    | 33.3       | 38.0 | 15.0 | 72.5 | 36.3 | 36.3 | 66.7 | 35.6 | 74.0      | 83.0 | 28.0 | 36.7 | 43.0 | 46.0 |      |
| InternVL2.5-8B [57]         | 38.3       | 37.0 | 10.0 | 77.0 | 33.4 | 35.5 | 65.0 | 21.6 | 73.0      | 83.0 | 34.0 | 50.0 | 43.0 | 46.2 |      |
| Qwen2.5-VL-7B [58]          | 33.3       | 40.0 | 31.0 | 77.0 | 40.6 | 40.5 | 66.7 | 36.2 | 68.0      | 72.0 | 27.0 | 38.3 | 45.0 | 47.4 |      |
| InternVL3-78B [56]          | 23.3       | 49.0 | 33.0 | 74.0 | 42.5 | 37.4 | 66.7 | 30.0 | 76.0      | 81.0 | 40.0 | 45.0 | 42.0 | 49.2 |      |
| Qwen2.5-VL-72B [58]         | 33.3       | 47.0 | 39.0 | 80.0 | 45.3 | 42.1 | 65.0 | 34.0 | 71.0      | 74.0 | 37.0 | 43.3 | 42.0 | 50.2 |      |
| GeoLLaVA-8K (Our)           | 26.7       | 38.0 | 49.0 | 69.0 | 41.6 | 31.6 | 65.0 | 35.0 | 67.0      | 78.0 | 66.0 | 50.0 | 52.0 | 51.5 |      |

**Main Results.** After fine-tuning on SuperRS-VQA and HighRS-VQA, our GeoLLaVA-8K delivers outstanding performance across various evaluation tasks. It not only outperforms domain-specific models but also surpasses all existing open- and closed-source models, including the latest Qwen2.5 and InternVL3. Remarkably, with just 7B parameters, GeoLLaVA-8K even outperforms Qwen2.5-VL-72B, the largest and best-performing open-source MLLMs. This impressive gain stems from the high-quality dataset and the targeted compression strategy designed for the low semantic density of RS imagery.

### 5.2 Further Analyses

**Effect of High-Resolution Data vs. Token Optimization Strategies** We conducted an ablation study to analyze the effects of high-resolution data and token optimization strategies. As shown in Table 6, using high-resolution datasets (SuperRS-VQA and HighRS-VQA) already improves performance over the baseline. When combined with Background Token Pruning (BTP) and Anchored Token Selection (ATS), the model achieves the best average accuracy of 51.5%. This shows that while high-resolution data enhances visual understanding, most performance gains come from token optimization, which focuses attention on semantically important regions.

Table 5: **FLOPs and latency of GeoLLaVA-8K under different compression ratios.**

| Compression | Tokens/Grid | Visual Tokens | TFLOPs | Latency (s/img) | Avg. |
|-------------|-------------|---------------|--------|-----------------|------|
| 16× (OOM)   | —           | —             | —      | —               | —    |
| 24×         | 24          | 14.0k         | 198.06 | 2.17            | 51.5 |
| 32×         | 18          | 10.5k         | 149.08 | 2.03            | 50.3 |
| 48×         | 12          | 7.1k          | 100.11 | 1.69            | 50.1 |
| 96×         | 6           | 3.6k          | 51.13  | 1.46            | 49.7 |

Table 6: Ablation study on the effect of high-resolution datasets and token optimization strategies.

| Subtasks                            | OC   | RC   | OLUC | RLUC | OCC  | OCL  | OMS  | OSR  | AD   | ECR  | RP   | RCCD | CCR  | Avg. |
|-------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| LLaVA-Next-2k (Baseline)            | 28.3 | 47.0 | 3.0  | 62.0 | 42.0 | 34.1 | 66.7 | 27.2 | 73.0 | 75.0 | 34.0 | 48.3 | 40.0 | 44.7 |
| + High-Resolution Dataset           | 25.9 | 41.0 | 50.0 | 68.0 | 40.4 | 28.5 | 64.9 | 33.4 | 58.0 | 77.0 | 57.0 | 45.0 | 50.0 | 49.1 |
| + Dataset + BTP(OOM)                | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | OOM  |
| + Dataset + BTP + ATS (GeoLLaVA-8K) | 26.7 | 38.0 | 49.0 | 69.0 | 41.6 | 31.6 | 65.0 | 35.0 | 67.0 | 78.0 | 66.0 | 50.0 | 52.0 | 51.5 |

**Computational Efficiency under Different Compression Ratios** We evaluated the computational efficiency of GeoLLaVA-8K by comparing FLOPs and inference latency across different token compression ratios. As shown in Table 5, higher compression ratios greatly reduce visual tokens and computational cost. At  $24\times$  compression, the model processes about 14K tokens with 198.06 TFLOPs and 2.17 s/image latency. Further compression to  $96\times$  lowers FLOPs to 51.13 and latency to 1.46 s/image, demonstrating the scalability and efficiency of our token optimization framework.

**Small Object Preservation and CLS Attention under Token Compression.** We analyzed the impact of token compression and CLS-based attention on small-object scenes using a subset of XLRS-Bench defined by COCOs area criteria. Small objects were defined as occupying less than 5% of the image area, and categorized into three levels: Extremely Small ( $<0.1\%$ ), Very Small ( $0.1\%-2.0\%$ ), and Normal Small ( $2.0\%-5.0\%$ ). A dedicated subset of 1,570 samples covering six tasks (classification, color, motion state, counting, spatial relations) was curated. As shown in Table 7, GeoLLaVA-8K outperformed LLaVA-Next across all tasks, with notable improvements in spatial reasoning, counting, and motion state. These results show that token optimization effectively preserves small-region semantics and reduces information loss. The CLS attention mechanism generally focuses on task-relevant small objects but occasionally favors a single dominant region, suggesting room for further improvement in multi-object focus and attention balance.

Table 7: Performance comparison on the small-object subset of XLRS-Bench.

| Dataset     | Object Classification | Object Color | Object Motion State | Regional Counting | Overall Counting | Object Spatial Relationship | Overall |
|-------------|-----------------------|--------------|---------------------|-------------------|------------------|-----------------------------|---------|
| LLaVA-Next  | 38.6                  | 29.4         | 62.0                | 35.0              | 22.5             | 20.0                        | 34.1    |
| GeoLLaVA-8K | 45.3                  | 34.4         | 70.0                | 40.0              | 32.5             | 35.0                        | 40.3    |

**Generalization on External Datasets** To evaluate the generalization of GeoLLaVA-8K, we tested it on the new ultra-high-resolution benchmark LRS-VQA [59]. Since LRS-VQA only provides open-ended QA annotations, we converted it to a multiple-choice format using an LLM [12] to generate three plausible distractors per question, followed by structural validation. GeoLLaVA-8K achieved 56.28% accuracy, outperforming LLaVA-Next [28] (55.07%) by 1.21%, confirming strong generalization to new datasets and formats.

Table 8: Generalization performance on LRS-VQA.

| Method             | Accuracy (%) | Evaluation Format |
|--------------------|--------------|-------------------|
| LLaVA-Next [28]    | 55.07        | MCQ               |
| GeoLLaVA-8K (ours) | <b>56.28</b> | MCQ               |

## 6 Conclusion

In this paper, we addressed two fundamental challenges in scaling vision-language models to UHR RS imagery: the lack of suitable training data and the computational burden of token explosion. To tackle these issues, we introduced two new UHR image-text datasets, SuperRS-VQA and HighRS-VQA, which greatly expand the data available for vision-language tasks on UHR RS images. We also proposed two novel token-efficient strategies: Background Token Pruning and Anchored Token Selection, which effectively reduce the number of visual tokens processed from 8K-resolution images while preserving essential information. Building on these contributions, we developed GeoLLaVA-8K, the first RS-specific MLLM that can directly handle inputs up to  $8K\times 8K$  resolution. GeoLLaVA-8K achieved state-of-the-art performance on the XLRS-Bench benchmark, outperforming both open- and closed-source MLLMs and demonstrating the effectiveness of our token-efficient approach. These results underscore the value of domain-adapted, token-efficient modeling and provide a new foundation for high-fidelity image understanding in RS.

**Limitation.** We have primarily focused on optical satellite imagery; evaluating GeoLLaVA-8K on other sensor modalities (e.g., synthetic aperture radar or multispectral images) will be important to ensure broader applicability.

**Acknowledgements:** This work was partially supported by the National Natural Science Foundation of China (No. 62372459, No.62376282, No.62225113 and No. 624B2109).

## References

- [1] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang. Crop classification based on feature band set construction and object-oriented approach using hyperspectral images. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 9(9):4117–4128, Sep. 2016.
- [2] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.
- [3] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, 111(9):3125–3160, 2022.
- [4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [5] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florenzia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.
- [6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [7] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.
- [8] 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.
- [9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.
- [10] 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.
- [11] 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.
- [12] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o>, 2024.
- [13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [14] Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *arXiv preprint arXiv:2503.23771*, 2025.

- [15] 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.
- [16] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.
- [17] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.
- [18] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.
- [19] Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaying Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, et al. H2rsvlm: Towards helpful and honest remote sensing large vision language model. *arXiv preprint arXiv:2403.20213*, 2024.
- [20] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024.
- [21] Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xiang Zhu. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. *Earth System Science Data Discussions*, 2024:1–24, 2024.
- [22] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaying 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.
- [23] Fengxiang Wang, Wanrong Huang, Shaowu Yang, Qi Fan, and Long Lan. Learning to learn better visual prompts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5354–5363, 2024.
- [24] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.
- [25] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.
- [26] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next1>, 2024.
- [29] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773, 2024.

- [30] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024.
- [31] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.
- [32] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [34] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.
- [35] Mengzhou Xia, Sathika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.
- [36] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018.
- [37] Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, et al. Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(5):6, 2024.
- [38] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao 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*, 184:116–130, 2022.
- [39] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.
- [40] Bohao Huang, Kangkang Lu, Nicolas Audeberr, Andrew Khalel, Yuliya Tarabalka, Jordan Malof, Alexandre Boulch, Bertr Le Saux, Leslie Collins, Kyle Bradbury, et al. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6947–6950. IEEE, 2018.
- [41] Hongbo Zhao, Lue Fan, Yuntao Chen, Haochen Wang, Xiaojuan Jin, Yixin Zhang, Gaofeng Meng, ZHAO-XIANG ZHANG, et al. Opensatmap: A fine-grained high-resolution satellite dataset for large-scale map construction. *Advances in Neural Information Processing Systems*, 37:59216–59235, 2024.
- [42] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019.
- [43] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
- [44] William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.



- [45] Sang Keun Choe, Hwijee Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, et al. What is your data worth to gpt? Ilm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*, 2024.
- [46] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.
- [47] Nostalgebraist. Interpreting gpt: The logit lens. <https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>, August 2020. Accessed: 2025-03-24.
- [48] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.
- [49] 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.
- [50] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkan Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.
- [51] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models. <https://github.com/EvolvingLMs-Lab/lmms-eval>, March 2024.
- [52] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence>, 2024.
- [53] Anthropic. Anthropic ai, 2023.
- [54] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
- [55] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [56] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, 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.
- [57] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [58] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [59] Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. *arXiv preprint arXiv:2503.07588*, 2025.

- [60] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [61] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our abstract and introduction accurately reflect our contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations and future directions at the end of the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: All experimental results are reproducible. Detailed information is provided in the appendix. We also plan to open-source our dataset and code to further contribute to the community.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide detailed descriptions of the dataset and experimental setup in the paper. Due to time constraints and submission size limits, the data and code cannot be included at this stage, but we commit to open-sourcing all datasets and code as soon as possible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All results are reproducible, with details in the appendix. We will also open-source our dataset and code to support future research.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We conducted extensive experiments in strict accordance with prior work in the field and commit to open-sourcing our code to ensure reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.



- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Similar to previous work, we used 1624 A100 GPUs as computing resources. Variations in computational resources do not affect our experimental results. Detailed information is provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We strictly adhered to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: **Positive Impacts.** The release of SuperRS-VQA, HighRS-VQA, and GeoLLaVA-8K empowers researchers and public sectors to monitor environmental

changes such as deforestation, coastal erosion, glacier retreat, and oil spills at an unprecedented 8K resolution. These tools also enable rapid identification of damaged infrastructure and flood zones after disasters, enhancing rescue efficiency and saving lives. Urban planners and agricultural professionals can utilize fine-grained object recognition and QA capabilities for energy savings, precision irrigation, and optimized land use. Background token pruning reduces memory and computational demands, allowing institutions in developing countries to train and fine-tune models on mid-range GPUs, fostering open science and education. The open, large-scale benchmark datasets also provide transparent, reproducible baselines for advancing multimodal research and system reliability. **Negative Impacts** Large-scale training and inference require significant energy, potentially raising carbon emissions if powered by non-renewable sources. Dataset imbalance in geography, season, or sensor type may degrade performance in underrepresented regions, reinforcing existing disparities. Fluent model-generated outputs may be misinterpreted as authoritative; if incorrect, they could mislead disaster response or environmental assessments.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We did not collect any potentially sensitive data from the internet. All provided data comply with established protocols in the remote sensing imagery community, ensuring there are no security concerns.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All data, models, and API utilized in this paper are publicly available and appropriately cited. The license and terms of use are properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide a new dataset, which will be released in the future for use by the research community.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

**15. Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**16. Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We only used LLMs to assist with writing, complying with the LLM policy. For benchmark evaluation, MLLMs were employed as assessment tools, following a widely accepted evaluation protocol in this field.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Appendix

### A.1 Overview of the Appendix

This appendix supplements the proposed **GeoLLaVA-8K** and our datasets (**SuperRS-VQA** and **HighRS-VQA**) with details excluded from the main paper due to space constraints.

The appendix is organized as follows:

- Sec. A.2: More implement details of GeoLLaVA-8K.
- Sec. A.3: The analysis on L-2 capability on GeoLLaVA-8K.
- Sec. A.4: Ablation studys of GeoLLaVA-8K.
- Sec. A.5: More implement details of Pilot Experiments.
- Sec. A.6: Visualizations of samples and challenging cases.
- Sec. A.7: Datasheets for the SuperRS-VQA and HighRS-VQA dataset.
- Sec. A.8: Discussion on limitations and societal impact.

### A.2 More Details of GeoLLaVA-8K

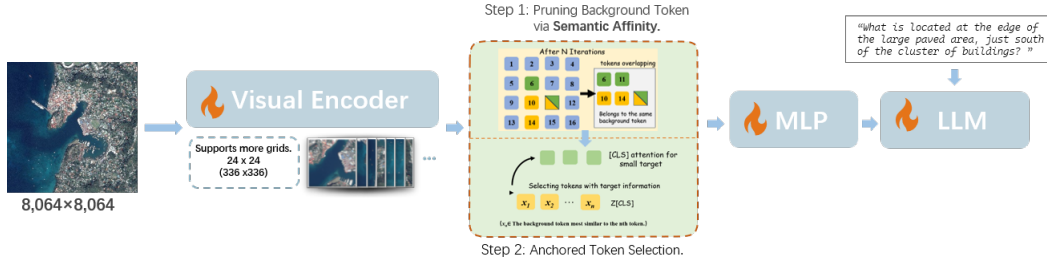


Figure 8: **Overall training framework.** We use SuperRS-VQA and HighRS-VQA datasets for SFT. Our two-step tokens compression method is performed between visual encoder and projection layer.

GeoLLaVA-8k is developed via full-parameter supervised fine-tuning (SFT) of LLaVA-Next-7B [28] on our SuperRS-VQA and HighRS-VQA datasets (see Fig. 8). We first expanded LLaVA-Next’s capacity from processing 7×7 visual grids (CLIP’s default 336×336 pixels per grid, ≈2K resolution) to 24×24 grids (≈8K resolution) by directly adjusting the relevant hyperparameters.

To address the explosion of visual tokens, we introduced a two-stage token-efficient mechanism. As described in Section 4.3, the original 2×2 pooling was replaced with a clustering-based strategy to merge and compress background tokens. We then applied Anchored Token Selection to further reduce tokens while preserving informative tokens. This approach achieves significantly higher compression: reducing 576 tokens per grid to 24 representative ones, achieving a 24× compression rate, which clearly outperforms the 4× reduction of original 2×2 pooling.

Tab. 9 presents the key training configuration of our GeoLLaVA-8K model. The model is trained on 81K UHR geospatial image-text pairs. We use different learning rates for the visual components (1e-6) and the projection layers interacting with the LLM (5e-6), and optimize the model using ZeRO-2 parallelism with a batch size of 32 for one training epoch.

Table 9: **Training Configuration of GeoLLaVA-8K.**

| Configuration | Parameter                       |
|---------------|---------------------------------|
| Resolution    | 8,064×8,064                     |
| Dataset       | 81,367 (SuperRS-VQA+HighRS-VQA) |
| Batch Size    | 16                              |
| LR: vision    | 1e-6                            |
| LR: proj. LLM | 5e-6                            |
| ZeRO stage    | ZeRO 2                          |
| Epoch         | 1                               |

Tab. 10 presents the detailed structure of XLRB-Bench, the benchmark dataset used to evaluate GeoLLaVA-8K. The dataset is hierarchically organized into two Level-1 categories: Perception and Reasoning, each comprising multiple Level-2 and Level-3 sub-tasks. All Level-3 sub-tasks are formatted as VQA with multiple-choice questions, totaling 3,040 samples. Perception tasks assess

fundamental visual understanding, while reasoning tasks evaluate higher-level cognitive abilities in geospatial contexts.

Table 10: **Characteristics of XLRS-Bench**, used as the benchmark for evaluating GeoLLaVA-8K. The full names of the Level-3 task abbreviations are also provided.

| L1-Task    | L2-Task                     | L3-Task                                 | Abbr. | Annotation Format | Number of Samples | Answer Type                     |
|------------|-----------------------------|-----------------------------------------|-------|-------------------|-------------------|---------------------------------|
| Perception | Counting                    | Overall Counting                        | OC    | VQA               | 60                | Multiple Choice(A/B/C/D)        |
|            |                             | Regional Counting                       | RC    | VQA               | 100               | Multiple Choice(A/B/C/D)        |
|            | Scene Classification        | Overall Land Use Classification         | OLUC  | VQA               | 100               | Multiple Choice(A/B/C/D)        |
|            |                             | Regional Land Use Classification        | RLUC  | VQA               | 200               | Multiple Choice(A/B/C/D)        |
|            | Object Spatial Relationship | Object Spatial Relationship             | OSR   | VQA               | 500               | Multiple Choice(A/B/C/D)        |
|            | Object Properties           | Object Classification                   | OCC   | VQA               | 800               | Multiple Choice(A/B/C/D)        |
|            |                             | Object Color                            | OCL   | VQA               | 800               | Multiple Choice(A/B/C/D)        |
|            |                             | Object Motion State                     | OMS   | VQA               | 60                | Multiple Choice(A/B for Yes/No) |
| Reasoning  | Route Planning              | Route Planning                          | RP    | VQA               | 100               | Multiple Choice(A/B/C/D)        |
|            | Anomaly Reasoning           | Anomaly Detection and Interpretation    | AD    | VQA               | 100               | Multiple Choice(A/B/C/D)        |
|            | Complex Reasoning           | Environmental Condition Reasoning       | ECR   | VQA               | 100               | Multiple Choice(A/B/C/D)        |
|            |                             | Counting with Complex Reasoning         | CCR   | VQA               | 100               | Multiple Choice(A/B/C/D)        |
|            | Spatiotemporal Reasoning    | Regional Counting with Change Detection | RCCD  | VQA               | 60                | Multiple Choice(A/B/C/D)        |

### A.3 Sub-tasks (L-2 capability) Results on GeoLLaVA-8K

Table 11: Experimental results on the perception and reasoning dimensions of VQA tasks, with models ranked by average performance. 'Avg' represents the average accuracy across sub-tasks.

| Method                      | Perception |                      |                             |                   | Reasoning |                   |                   |                          |
|-----------------------------|------------|----------------------|-----------------------------|-------------------|-----------|-------------------|-------------------|--------------------------|
| Sub-tasks (L-2 Capability)  | Counting   | Scene Classification | Object Spatial Relationship | Object Properties | Planning  | Anomaly Reasoning | Complex Reasoning | Spatiotemporal Reasoning |
| Remote Sensing MLLMs        |            |                      |                             |                   |           |                   |                   |                          |
| GeoChat [9]                 | 22.8       | 12.5                 | 24.2                        | 24.3              | 10.0      | 33.0              | 32.0              | -                        |
| Closed-source MLLMs         |            |                      |                             |                   |           |                   |                   |                          |
| GPT-4o [12]                 | 28.5       | 40.5                 | 24.6                        | 10.8              | 35.0      | 73.0              | 49.0              | 20.0                     |
| GPT-4o-mini [52]            | 24.2       | 39.3                 | 23.6                        | 45.6              | 29.0      | 71.0              | 50.5              | 6.7                      |
| Claude 3.7 Sonnet [53]      | 25.2       | 42.9                 | 27.6                        | 41.3              | 34.5      | 64.8              | 55.5              | 27.8                     |
| Gemini 2.0 Flash [4]        | 43.3       | 55.8                 | 32.0                        | 41.3              | 43.0      | 73.0              | 66.5              | 30.0                     |
| Open-source MLLMs           |            |                      |                             |                   |           |                   |                   |                          |
| InternLM-XComposer-2.5 [54] | 31.8       | 37.5                 | 26.0                        | 22.1              | 41.0      | 72.0              | 64.0              | 36.7                     |
| LLaVA-Next [28]             | 33.3       | 36.0                 | 30.0                        | 42.7              | 27.0      | 69.0              | 57.0              | 35.0                     |
| LLaVA-OneVision-7B[55]      | 31.5       | 35.8                 | 25.2                        | 45.4              | 24.0      | 76.0              | 59.5              | 43.3                     |
| LLaVA-OneVision-72B[55]     | 35.7       | 43.8                 | 35.6                        | 46.4              | 28.0      | 74.0              | 63.0              | 36.7                     |
| InternVL2.5-8B [57]         | 37.7       | 43.5                 | 21.6                        | 44.6              | 34.0      | 73.0              | 63.0              | 50.0                     |
| InternVL3-8B [56]           | 39.5       | 40.8                 | 25.2                        | 46.8              | 36.0      | 77.0              | 66.0              | 21.7                     |
| Qwen2.5-VL-7B [58]          | 33.3       | 42.0                 | 31.8                        | 44.0              | 35.0      | 70.0              | 64.5              | 46.7                     |
| Qwen2.5-VL-72B [58]         | 36.7       | 54.0                 | 36.2                        | 49.3              | 27.0      | 68.0              | 58.5              | 38.3                     |
| GeoLLaVA-8K (Our, 7B)       | 32.3       | 59.0                 | 35.0                        | 46.1              | 66.0      | 67.0              | 65.0              | 50.0                     |

This section highlights the performance of MLLMs across L-2 capabilities, while the related experimental results are shown in Tab. 11.

#### Superior Performance in Local Parsing

GeoLLaVA-8K’s superior performance in tasks such as Object Properties (46.1, surpassing the recent Qwen2.5-VL-7B) and Object Spatial Relationship (35.0, only lower than 72B models) demonstrates its ability to effectively parse local details in UHR imagery, laying a solid foundation for comprehensive and accurate ground object understanding. In UHR images, objects often exhibit substantial variations (e.g., buildings or vehicles in different colors), with diverse distribution patterns. GeoLLaVA-8K effectively preserves and leverages these subtle yet critical visual cues under challenging conditions, leading to superior analytical performance.

#### Robust Capability in Holistic Understanding.

GeoLLaVA-8K’s top performance in Scene Classification (59.0, 1st) and Complex Reasoning (65.0, tied for 2nd) highlight its ability to capture core semantics and underlying logic from UHR imagery at a macroscopic level. For example, the model can accurately determine a regions overall function, assess the stage of an engineering project, or infer plausible causes of anomalous phenomena by integrating rich visual cues, which is essential for unlocking the full potential of UHR imagery.

**Strong Potential in Dynamic Reasoning** The most distinct advantage of GeoLLaVA-8K lies in its ability to finely comprehend dynamic processes. Its top performance in Planning (66.0, 1st) and



Spatiotemporal Reasoning (50.0, 1st), highlights its effectiveness in modeling long-range dependencies for reasoning. This capability is essential for applications that require precise spatial and temporal understanding, such as urban expansion analysis, disaster response planning, and dynamic monitoring of natural resources, which are challenging for existing models.

#### A.4 Ablation studies of GeoLLaVA-8K

Table 12: **Results of ablation experiments on compression rates.** 'Avg.' represents the average accuracy across sub-tasks. We mark the highest score in **red**.

| Compression Ratio | OC   | RC   | OLUC | RLUC | OCC  | OCL  | OMS  | OSR  | AD   | ECR  | RP   | RCCD | CCR  | Avg.        |
|-------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------------|
| 16                |      |      |      |      |      |      |      |      |      |      |      |      |      | OOM         |
| 32                | 28.3 | 39.0 | 46.0 | 65.5 | 42.3 | 34.1 | 65.0 | 35.6 | 64.0 | 74.0 | 63.0 | 46.7 | 50.0 | 50.3        |
| <b>24 (Our)</b>   | 26.7 | 38.0 | 49.0 | 69.0 | 41.6 | 31.6 | 65.0 | 35.0 | 67.0 | 78.0 | 66.0 | 50.0 | 52.0 | <b>51.5</b> |

To investigate the key components of the methodology, we conducted comprehensive ablation studies.

**Visual Token Compression Strategy.** The initial ablation study, with detailed results in Tab. 12, focused on optimizing the visual token compression ratio. We evaluated three settings: 16, 24, and 32. Training with a ratio of 16 frequently led to out-of-memory (OOM) errors, even on multi-GPU setups (8 or 16 GPUs). Although the ratio of 32 used fewer tokens, it resulted in lower accuracy. In contrast, the 24-token setting achieved the best average accuracy (51.5) and enabled stable training on a single node with 8 GPUs striking an effective balance between performance and efficiency.

Table 13: **Results of ablation experiments on datasets.** 'Avg.' represents the average accuracy across sub-tasks. We mark the highest score in **red**.

| Datasets                        | OC   | RC   | OLUC | RLUC | OCC  | OCL  | OMS  | OSR  | AD   | ECR  | RP   | RCCD | CCR  | Avg.        |
|---------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------------|
| VRSBench-train [20]             | 13.3 | 28.0 | 11.0 | 39.5 | 30.1 | 24.6 | 36.7 | 23.6 | 61.0 | 70.0 | 34.0 | 45.0 | 35.0 | 34.8        |
| SuperRS-VQA                     | 20.0 | 37.0 | 34.0 | 63.0 | 41.4 | 30.3 | 65.0 | 33.4 | 73.0 | 79.0 | 63.0 | 46.7 | 56.0 | 49.4        |
| <b>SuperRS-VQA + HighRS-VQA</b> | 26.7 | 38.0 | 48.0 | 69.0 | 41.6 | 31.6 | 65.0 | 35.0 | 67.0 | 78.0 | 66.0 | 50.0 | 52.0 | <b>51.5</b> |

**Impact of Training Data Composition.** The second ablation study, summarized in Tab. 13, assessed the contributions of different data sources to the final model performance. We compared three primary scenarios:

- (i) training solely with the VRSBench-train dataset (which contains 123,221 synthetically generated samples);
- (ii) training solely with the SuperRS-VQA dataset (which contains 12,228 high-quality, human-annotated samples);
- (iii) training with a synergistic blend of the SuperRS-VQA dataset and the HighRS-VQA dataset (comprising 12,228 high-quality, human-annotated and 69,139 synthetically generated samples).

As shown in Tab. 13, although VRSBench contains approximately ten times more samples than SuperRS-VQA, training solely on SuperRS-VQA yields significantly better performance (49.4 vs. 34.8). This is likely attributed to the substantial resolution gap: VRSBench averages  $512 \times 512$  pixels, while SuperRS-VQA averages  $8,376 \times 8,376$ , and also reflects the higher annotation quality of SuperRS-VQA. A further accuracy boost to 51.4 was observed when incorporating the HighRS-VQA dataset, validating its complementary value and effectiveness.

#### A.5 More implement details of Pilot Experiments.

This section provides more implementation details of the experiments in Section 4.1-4.2.

##### Multi-level Semantic Parsing Framework in Sec.4.1.

The framework separates the description from the evaluation content and processes it in five steps:

- (1) **Urban Semantic Recognition**, which detects terms like “city” or “urban” and reduces the background probability if identified;
- (2) **Man-made Structure Analysis**, which evaluates semantic relationships between terms like “building,” “road,” and negation words;
- (3) **Natural Element Classification**, which calculates the weight of features like “water” and “forest”;
- (4) **Semantic Categorization**, which classifies image regions as “Dense,” “Sparse,” or “None”;
- (5) **Probability Gradient Quantification**, which assigns values based on semantic strength, such as 20% for urban semantics, 30% for composite man-made structures, and up to 95% for pure natural features.

**logit lens Technique in Sec.4.1.** For each hidden state  $h_l^i$  at position  $i$  in layer  $l$ , we project it into a probability space of tokenizer vocabulary and select the character with the highest logit. The original paper notes that visual-language models can optimize the alignment between visual and language modalities, enabling visual token representations in deeper layers to decode into language tokens that reflect object semantics. Fig. 9 shows additional examples using the *logit lens* technique to analyze the semantic content of visual tokens.

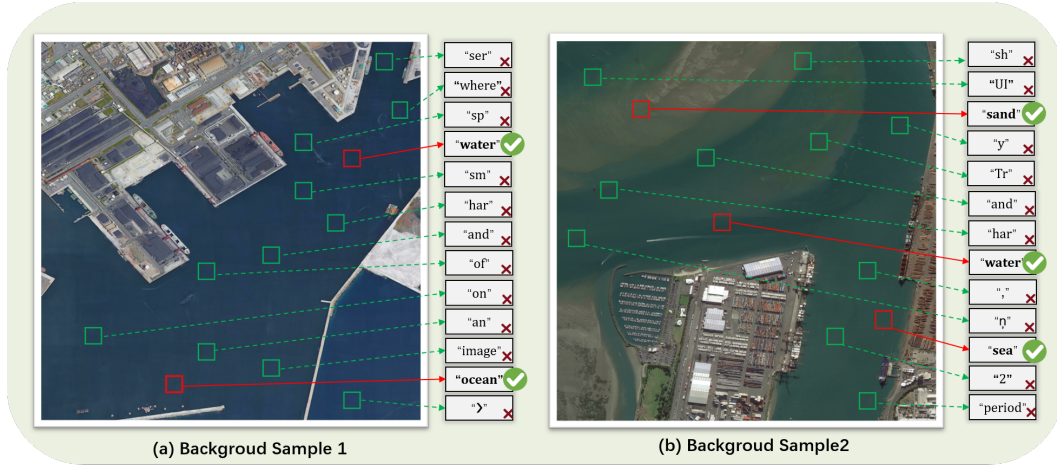


Figure 9: Visual tokens and their semantics detected by the *logit lens* technique.

**Reliability Image Prepare in Sec.4.2.** The images are sourced from XLRS-Bench, with two pre-processing steps applied to ensure reliability:

- (1) **High signal-to-noise sub-images:** Following [60], we apply target-centered cropping to increase the object coverage ratio relative to the cropped image size, thereby enhancing target visibility.
- (2) **Hallucination control:** Models may infer the presence of objects from context even when visual cues are removed. To counteract this, we generate two versions of each image: the original and a control version with the target masked by noise. We keep only those samples where the model correctly identifies the object in the original but fails in the occluded version, ensuring that recognition relies on visual evidence rather than context.

**Experimental details in Sec.4.2.** Token ablation is performed via substitution. Specifically, token positions corresponding to target objects are identified using XLRS-Bench grounding annotations. These tokens are then replaced with a fixed average embedding, computed as the mean of all visual tokens across the XLRS-Bench images used in this experiment.

We use two methods to assess the impact of token ablation:

1. **Generative Description:** The model is prompted with “Describe the image within the bounding box” using both the original embeddings ( $E_A$ ) and the ablated embeddings ( $E'_A$ ). We compare whether the object  $o$  appears in the generated descriptions. If it is mentioned with  $E_A$  but not with  $E'_A$ , the ablated tokens are deemed crucial for recognizing the object.

To ensure the response targets the annotated object, we include the bounding box in the prompt, given the frequent repetition of similar objects in remote sensing images.

2. **Visual Question Answering (VQA):** We used GPT-4o to generate detailed questions for 400 manually selected remote sensing images, drawn from the XLRS-Bench subset used in this experiment. For example, “What is located at the edge of the large paved area, just south of the cluster of buildings with flat roofs, and adjacent to the narrow green space between the two structures?” All questions were reviewed for clarity and to avoid explicitly naming the target (e.g., “boat”). The models answer is prefixed with “It is a ” and we compare its responses before and after ablation.

## A.6 Cases of GeoLLaVA-8K

In this section, we present representative examples from XLRS-Bench in Fig. 10. Each example compares the closed-source Gemini model with the latest open-source models: InterVL-3 and Qwen2.5-VL, across various VQA task formats. Extracting and interpreting objects from large-scale remote sensing imagery remains highly challenging; nonetheless, GeoLLaVA-8K exhibits strong performance in object recognition, localization, and counting under complex, real-world conditions.



How many **tennis courts** are there within the red circle?

(A) 1  
(B) 8  
(C) 2  
(D) 7

Gemini 2.0 Flash: (A) ✓  
InternVL3-8B: (A) ✓  
Qwen2.5-VL-72B: (C) ✗  
GeoLLaVA-8K(Our): (A) ✓

(a) Case of regional counting task.

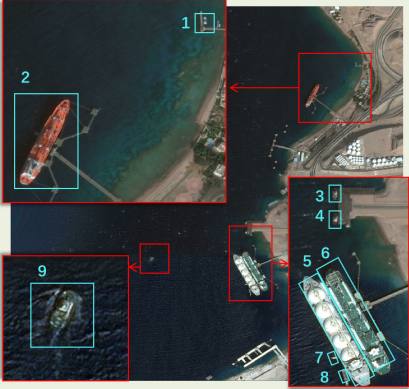


Identify the **object category** within the given bounding box in the satellite or aerial image. Image resolution: 10000 × 10000. Bounding box: [7472, 2102, 7494, 2128].

(A) The Gray Parking Garage  
(B) The Blue Swimming Pool  
(C) Blue Rooftops  
(D) The Green Swimming Pool

Gemini 2.0 Flash: (B) ✓  
InternVL3-8B: (C) ✗  
Qwen2.5-VL-72B: (C) ✗  
GeoLLaVA-8K(Our): (B) ✓

(b) Case of object classification task.



How many **boats** are there in the entire picture?

(A) 5  
(B) 9  
(C) 12  
(D) 15

Gemini 2.0 Flash: (A) ✗  
InternVL3-8B: (A) ✗  
Qwen2.5-VL-72B: (A) ✗  
GeoLLaVA-8K(Our): (B) ✓

(c) Case of overall counting task.

Figure 10: Cases of Different Task on Various MLLMs (Part 1)

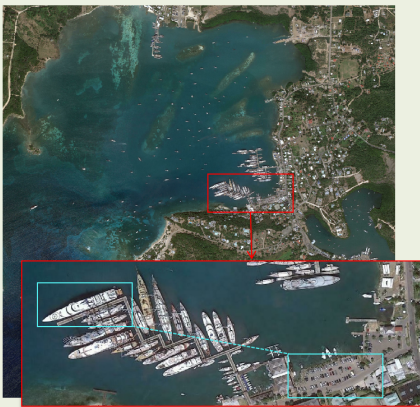


How many houses are there inside the red circle?

(A) 3  
(B) 4  
(C) 6  
(D) 7

Gemini 2.0 Flash: (A) ✗  
InternVL3-8B: (A) ✗  
Qwen2.5-VL-72B: (B) ✓  
GeoLLaVA-8K(Our): (B) ✓

(d) Case of regional counting task.



In the image, the **largest white vessel** docked is located in which direction relative to the **largest parking lot** in the harbor?

(A) To the left  
(B) Top left corner  
(C) Bottom right corner  
(D) Above

Gemini 2.0 Flash: (D) ✗  
InternVL3-8B: (D) ✗  
Qwen2.5-VL-72B: (D) ✗  
GeoLLaVA-8K(Our): (B) ✓

(e) Case of object spatial relationship task.

Figure 10: Cases of Different Task on Various MLLMs (Part 2)



## A.7 Datasheets

In this section, we document essential details about the proposed datasets and benchmarks following the NeurIPS Dataset and Benchmark guidelines and the template provided by Gebru *et al.* [61].

### A.7.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

1. “For what purpose was the dataset created?”

**A:** Ultra-high-resolution (UHR) remote sensing (RS) imagery offers valuable data for Earth observation but pose challenges for existing multimodal foundation models due to the key bottlenecks: limited availability of UHR training data. To address data scarcity, we introduce **SuperRS-VQA** (avg.  $8,376 \times 8,376$ ) and **HighRS-VQA** (avg.  $2,000 \times 1,912$ ), the highest-resolution vision-language datasets in RS to date, covering 22 real-world dialogue tasks.

2. “Who created the dataset (e.g., which team, research group) and on behalf of which entity?”

**A:** The dataset was created by the following authors:

- Anonymous authors

3. “Who funded the creation of the dataset?”

**A:** The dataset creation was funded by the affiliations of the authors involved in this work.

### A.7.2 Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU’s General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

1. “What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?”

**A:** The dataset primarily consists of ultra-high-resolution remote sensing images captured by satellites, along with their corresponding textual annotations. All datasets utilized in SuperRS-VQA and HighRS-VQA are publicly accessible and nonprofit.

2. “How many instances are there in total (of each type, if appropriate)?”

**A:** SuperRS-VQA and HighRS-VQA includes 81,367 VQA pairs. Details could be found in the Table 2 in main text.

3. “Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?”

**A:** The images in SuperRS-VQA and HighRS-VQA are sourced from existing detection and segmentation datasets, but all textual annotations were independently created by us.

4. “Is there a label or target associated with each instance?”

**A:** Yes, for these ultra-high-resolution images, we have provided VQA pairs instances.

5. “Is any information missing from individual instances?”

**A:** No, each individual instance is complete.

6. “Are relationships between individual instances made explicit (e.g., users movie ratings, social network links)?”

**A:** Yes, the relationship between individual instances is explicit.

7. “Are there recommended data splits (e.g., training, development/validation, testing)?”

**A:** The dataset is designed to train the UHR RS MLLMs.



8. *“Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?”*  
**A:** SuperRS-VQA and HighRS-VQA are self-contained and will be open-sourced on platforms like Hugging Face for easy use.
9. *“Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor/patient confidentiality, data that includes the content of individuals non-public communications)?”*  
**A:** No, all data are clearly licensed.
10. *“Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?”*  
**A:** No, SuperRS-VQA and HighRS-VQA do not contain any data with negative information.

### A.7.3 Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

1. *“How was the data associated with each instance acquired?”*  
**A:** The images in SuperRS-VQA and HighRS-VQA are sourced from existing detection and segmentation datasets. We enrich these ultra-high-resolution images with manual annotations, including 81,367 VQA pairs instances. Details are shown in the Section 3 in main text.
2. *“What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?”*  
**A:** We employed professional annotation and quality control teams to complete the annotations for VQA tasks in SuperRS-VQA. For MHR data, we developed a semi-automated annotation. Using task-specific prompts and existing annotations (e.g., bounding boxes in RS detection datasets), we generated text via GPT-4o. We further adopted an influence-based data selection pipeline to improve the relevance of our dataset to UHR downstream tasks and ensure its cultivation of reasoning capabilities for models fine-tuned on it.
3. *“If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?”*  
**A:** Please refer to the details listed in the main text Section 3.

### A.7.4 Preprocessing, Cleaning, and Labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the raw data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.

1. *“Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?”*  
**A:** Yes. To minimize the redundancy of image-text pairs, we deduplicate images within these datasets and remove overlaps with existing benchmark datasets like XLRS-Bench [14].
2. *“Was the ‘raw’ data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?”*  
**A:** Yes, raw data is accessible.
3. *“Is the software that was used to preprocess/clean/label the data available?”*  
**A:** Yes, the necessary software used to preprocess and clean the data is publicly available.

### A.7.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers make informed decisions, thereby avoiding potential risks or harms.

1. *“Has the dataset been used for any tasks already?”*

**A:** No.

2. *“Is there a repository that links to any or all papers or systems that use the dataset?”*

**A:** Yes, we will provide such links in the GitHub and the Huggingface repository.

3. *“What (other) tasks could the dataset be used for?”*

**A:** SuperRS-VQA and HighRS-VQA provide extensive annotations for VQA tasks. It could be used for training the MLLMs.

4. *“Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?”*

**A:** No.

5. *“Are there tasks for which the dataset should not be used?”*

**A:** N/A.

### A.7.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. *“Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?”*

**A:** The datasets will be made publicly accessible to the research community.

2. *“How will the dataset be distributed (e.g., tarball on website, API, GitHub)?”*

**A:** We will provide uperRS-VQA and HighRS-VQA in the GitHub and the Huggingface repository.

3. *“When will the dataset be distributed?”*

**A:** We will create a repository to release the data once the paper is officially published.

4. *“Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?”*

**A:** Yes, the dataset will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5. *“Have any third parties imposed IP-based or other restrictions on the data associated with the instances?”*

**A:** No.

6. *“Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?”*

**A:** No.

### A.7.7 Maintenance

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

1. *“Who will be supporting/hosting/maintaining the dataset?”*

**A:** The authors of this work serve to support, host, and maintain the datasets.

2. *“How can the owner/curator/manager of the dataset be contacted (e.g., email address)?”*

**A:** They can be contacted via the email addresses listed on the paper or webpage.

3. *“Is there an erratum?”*

**A:** There is no explicit erratum; updates and known errors will be specified in future versions.

4. *“Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?”*

**A:** Future updates (if any) will be posted on the dataset website.

5. *“Will older versions of the dataset continue to be supported/hosted/maintained?”*

**A:**

Yes. This initial release will be updated in the future, with older versions replaced as new updates are posted.

6. *“If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?”*

**A:** Yes, we will provide detailed instructions for future extensions.

## A.8 Limitation and Potential Societal Impact

In this section, we discuss the limitations and potential societal impact of this work.

### A.8.1 Potential Limitations (GeoLLaVA-8K)

- **Scope of Sensors:** GeoLLaVA-8K is tuned for satellite visiblelight imagery; its performance on SAR, multispectral or street-level data is unverified, limiting cross-domain generalisability.
- **Model Scale:** We used only a basic 7B model, while commercial models like Qwen and InternVL have reached 72B. We also aim to scale up model parameters for improved performance.

### A.8.2 Potential Positive Societal Impacts (GeoLLaVA-8K)

- **Fine-grained Environmental Intelligence:** Enables near-real-time QA over 8K imagery for detecting illegal logging, coastal erosion or oil spills, supporting evidence-based policy and SDG monitoring.
- **Disaster-Response Acceleration:** Rapid localisation of collapsed bridges, blocked roads or inundated zones can sharpen rescue logistics and reduce casualty rates.

### A.8.3 Potential Negative Societal Impacts (GeoLLaVA-8K)

- **Surveillance & Dual-Use Risk:** High-precision localisation of vehicles assets lowers the barrier for persistent monitoring and autonomous targeting.
- **Over-reliance on Automated QA:** Persuasive textual answers may be taken at face value; misclassification of disaster extent or land-use could misdirect resources.

### A.8.4 Potential Limitations (SuperRS-VQA & HighRS-VQA)

- **Sensor Homogeneity:** The datasets are dominated by RGB optical satellites; absence of SAR or hyperspectral samples constrains multimodal fusion research.
- **Annotation Consistency:** Crowdsourced VQA answers could contain subtle errors or regional terminology variance, introducing noise in supervision signals.

### A.8.5 Potential Positive Societal Impacts (SuperRS-VQA & HighRS-VQA)

- **Open Benchmark Catalyst:** By releasing the largest-image-size RS VQA corpora (up to 8,376E8,376), the datasets establish a transparent yard-stick for future UHR-aware algorithms.
- **Policy-Relevant Insights:** Rich dialogue tasks (22 subtasks) mirror real-world queries e.g., Count illegal fish-farms providing a sandbox for developing decision-support tools that aid sustainable development.

#### A.8.6 Potential Negative Societal Impacts (SuperRS-VQA & HighRS-VQA)

- **Privacy Concerns:** 153cm GSD tiles can reveal rooftop activity; malicious actors could deanonymise locations or monitor private property.
- **Environmental Footprint of Training:** Fine-tuning large models on tens of thousands of 8K images consumes significant energy; if replicated widely without checkpoint reuse, aggregate carbon emissions rise.