GeoSpatial NLI for Multimodal Satellite Imagery

Team 67

***Abstract*—Satellite imagery contains rich spatial and semantic information, yet its interpretation typically requires specialized expertise and complex geospatial tools. This work presents a unified natural-language interface for satellite image understanding that enables non-expert users to interact with overhead imagery through simple text-based queries. The proposed system supports three core tasks: (i) captioning, which summarizes global scene characteristics and localized structures; (ii) visual question answering, covering binary, numerical, and semantic queries; and (iii) grounding, where referenced objects are localized through oriented bounding boxes suitable for overhead viewpoints. To operate reliably across diverse sensor types and spatial resolutions, the framework incorporates preprocessing pipelines for multi-modal imagery and integrates domain-aware representations capable of handling images up to 2k×2k resolution and able to handle sampling from 0.5m/pixel to 10m/pixel imagery efficiently. Interpreting SAR (Synthetic Aperture Radar) imagery remains challenging due to the lack of SAR image-text pair datasets. To address this, we adopt a novel two-stage SAR grounding approach: a domain-specific SAR detector combined with a language model. The complete solution is deployed as a web-based platform designed for intuitive use, smooth high-resolution rendering, and responsive interaction. Users can upload images, request analyses, and visualize grounding outputs directly through the interface. The system is evaluated following VRS-Bench protocols for captioning, VQA, and OBB-based grounding. This work demonstrates a practical pathway toward accessible, intelligent geospatial analysis by combining remote-sensing principles with modern vision–language modeling.**

***Index Terms*—remote sensing, VLMs, natural language interfaces, VQA, grounding, segmentation, SAR**

I. Introduction

A. Motivation and Background

Satellite-based Earth observation has evolved into a critical resource for environmental monitoring, agricultural planning, disaster response, and national infrastructure management. Agencies such as ISRO's Space Applications Centre (SAC) routinely acquire imagery from diverse sensors spanning optical, multispectral, infrared, and microwave domains. With resolutions ranging from sub-meter data to 10-m/pixel multispectral products, the volume and heterogeneity of remote-sensing imagery continue to increase. While these datasets offer rich scientific and operational value, extracting meaningful information from them remains highly specialized.

Interpreting satellite imagery requires familiarity with sensor characteristics, spectral responses, spatial distortions, and geospatial software ecosystems. Non-expert users are often confronted with challenges such as SAR speckle, band-specific artifacts in infrared imagery, synthetic color assignments in false-color composites, and large variations in feature scale across different sensors. Even common analytical tasks-identifying settlements, estimating vegetation cover, or locating man-made structures demand considerable experience with GIS tools and remote-sensing conventions. Consequently, the benefits of high-quality EO data remain inaccessible to many potential users.

Recent advances in multimodal machine learning have demonstrated that natural language can serve as an intuitive interface for interacting with complex visual data. The ability to ask simple queries such as "What features dominate this scene?" or "Where are the built-up regions located?" offers a more approachable alternative to conventional geospatial workflows. However, most vision–language models are trained on everyday images and struggle with the overhead viewpoints, spectral variations, and orientation-dependent characteristics of remote-sensing imagery. This gap motivates the development of domain-adapted natural-language interfaces that can accurately interpret, describe, and localize content in satellite images while reducing reliance on expert driven analysis.

B. Problem Statement and Requirements

The proposed system must support natural-language driven analysis of satellite imagery across multiple tasks and sensing conditions. To meet operational expectations, the requirements are organized into three categories: core functional capabilities, data and modality considerations, and evaluation constraints.

1) Core Functional Capabilities: The system should provide end-to-end interpretability through captioning, visual question answering, and object grounding.

- **Captioning:** Generate coherent descriptions summarizing both global scene characteristics (such as land cover patterns or settlement structure) and important localized features.
- **Visual Question Answering (VQA):** Support diverse query types, including:
  - Binary responses (yes/no),

TABLE I
BASELINE MODEL PERFORMANCE COMPARISON ON VRS BENCHMARK

Model	**Caption**	**Grounding**	**VQA**	**Binary**	**Numeric**
	(LP × BERT-BLEU4)	(CP × MeanIoU)	(BERT-BLEU1)	Acc.	Score
LLaVA 1.5	0.7230	0.0211	0.6098	0.7321	0.3722
Moondream 2	0.7193	**0.7618**	0.6399	**0.8461**	0.4844
Moondream 3	0.7202	0.7490	0.5924	0.7138	0.4057
Qwen3–VL	**0.7522**	0.4840	**0.7476**	0.8388	**0.5188**
PaliGemma 2 10B	0.4012	0.2135	0.6726	0.4314	0.4012
InternVL	0.497	0.039	—	—	—

  - Quantitative or counting-based outputs,
  - Semantic, context-dependent answers.
- **Grounding:** Identify and localize objects referenced in natural-language queries.
  - Initial detections are obtained as horizontal bounding boxes (HBB),
  - These regions are then refined into precise object masks,
  - Final localization is represented as oriented bounding boxes (OBB) derived from the minimum-area rectangle around each mask.

2) Data and Modality Requirements: The system must operate reliably across diverse satellite sensor outputs and varying spatial resolutions. Although the primary reasoning pipeline operates in an RGB domain, internal preprocessing ensures compatibility with additional sensor-specific formats.

- **Multi-Modality:** Maintain consistent performance across different satellite-derived imaging types via appropriate preprocessing and conversion.
- **High-Resolution Support:** Efficiently process imagery up to approximately 2048×2048 pixels.
- **Scale Robustness:** Deliver stable performance across resolutions ranging from roughly 0.5--10 m/pixel, where object appearance varies significantly.

3) Evaluation Requirements: To enable standardized assessment and compatibility with established benchmarks, the system should adhere to the VRS-Bench evaluation protocol:

- Captioning evaluated using BERT BLEU-based metrics,
- VQA evaluated using binary accuracy, numeric deviation, and semantic similarity,
- Grounding performance is evaluated using the product of the count penalty (CP) and the MeanIoU across oriented bounding boxes.

II. BACKGROUND AND EXPLORED APPROACHES

We evaluate a broad set of state-of-the-art vision–language models (VLMs) and remote sensing datasets to guide model selection and data strategy for our unified framework. LLaVA [1] provides strong natural-image performance and effective vision–instruction tuning but lacks native grounding and is restricted to 512×512 inputs, limiting geospatial detail. Moondream-2 [2], despite its lightweight 2B size, performs competitively on localization and grounding, while Moondream-3 [3] offers a unified backbone supporting captioning, VQA, and grounding. Qwen3–VL [4], with native high-resolution support up to 2048×2048 and long-context reasoning, serves as the core for captioning and VQA, though it requires remote-sensing domain adaptation. PaliGemma–2 (10B) [5] shows strong task capability when fine-tuned but performs poorly under domain shift, and InternVL [6] exhibits similar natural-image bias, often producing street-view rather than nadir-view interpretations.

Our dataset survey ensures balanced supervision across tasks and sensor modalities. For captioning, RSICap [7] provides rich semantic descriptions but limited diversity, while RSICD [8] offers broader geographic coverage with simpler language; UCM [9] Captions and Sydney Captions [10] further support moderate-scale generalization. For VQA, VRSBench [11] supplies balanced, domain-appropriate QA distributions across diverse environments, whereas EarthVQA [12], though large-scale, shows 10–15% cross-domain degradation. For grounding, OBB-annotated datasets—including XLRSBench (HBB) [13], DOTA [14], DIOR-R [15], and ShipRSImageNet [16]—enable rotation-aware localization across varied categories and resolutions. Multimodal datasets such as SpaceNet [17] (urban segmentation), xView [18] (small-object detection), SAR-TEXT [19] and SARLANG-1M [20] (SAR–language alignment), SARDet-100K [21] (a consolidated COCO-scale multi-class SAR detection corpus), and co-registered Sentinel-1/2 imagery support multi-sensor fusion and SAR analysis, though high-quality SAR captioning resources remain scarce.

This combined assessment of models and datasets informs the architectural and training design of our unified remote sensing vision–language framework.

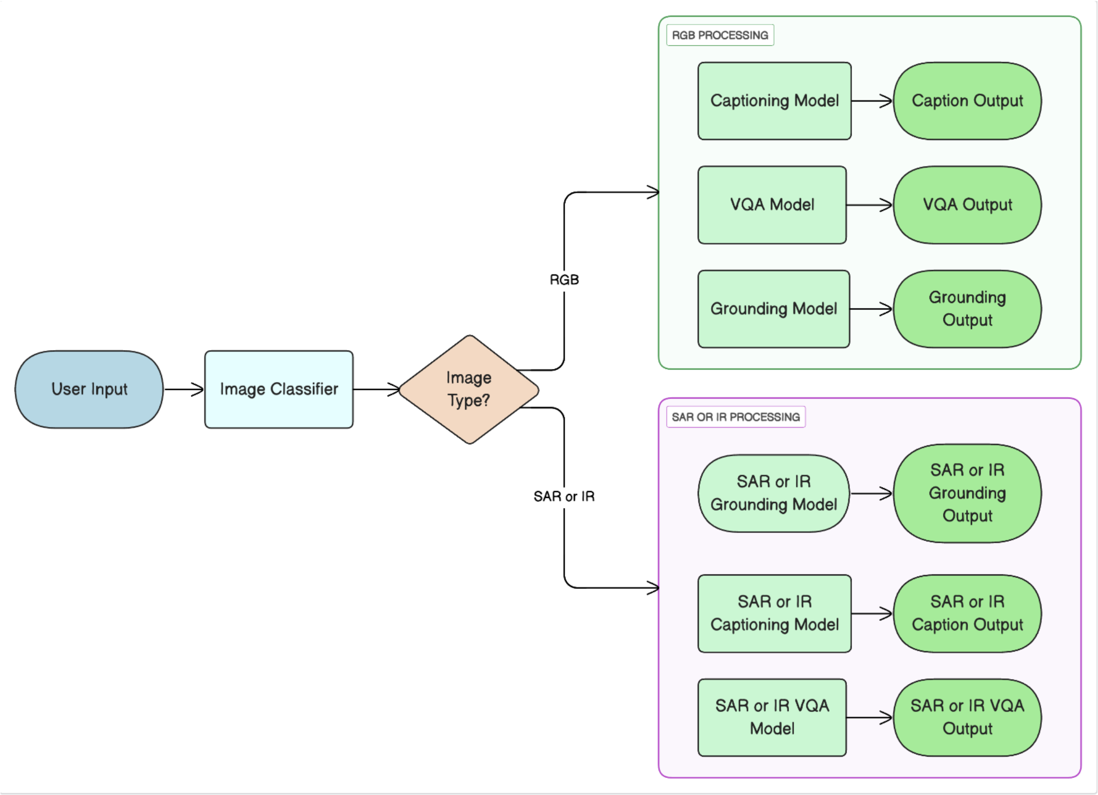

Fig. 1. System Workflow

III. Proposed Solution

A. Overall Pipeline

The proposed framework implements a unified processing pipeline designed to handle heterogeneous remote sensing imagery through sequential classification and conversion stages. Upon classification, images are directly forwarded to the downstream vision language modules finetuned for the respective modality. Non-RGB modalities undergo a domain specific transformation that preserves semantic content while maintaining compatibility with pre–trained vision language architectures. The images are subsequently processed by three dedicated downstream models image captioning, visual question answering, and object grounding which can operate either concurrently or sequentially, depending on the system configuration, leveraging high level feature representations to generate task–specific outputs.

B. Modality Classification

The modality classification network employs a four–layer convolutional architecture with progressive filter depths (64, 128, 256, 512), batch normalization, and ReLU activations, followed by global average pooling and a fully connected layer optimized using the cross–entropy loss to discriminate among RGB, SAR and IR modalities.

C. Captioning

The image captioning task is implemented using Qwen3–VL [4] as the base VLM, due to its native dynamic resolution processing that handles images up to 2048×2048 without requiring explicit tiling or downsampling, preserving fine-grained spatial details critical for comprehensive scene description. Unlike conventional VLMs that process fixed dimension inputs, Qwen3-VL [4] employs adaptive visual tokenization that dynamically adjusts token allocation based on image complexity. Additionally, its efficient long-context attention mechanism accommodates the extended visual token sequences generated from high-resolution inputs, enabling coherent caption generation that captures both global scene context and localized object details.

The model is fine-tuned on two complementary datasets: (i) RSICap [7], which provides 10,000+ high-quality image-caption pairs across 35 scene categories at 0.2m–3m GSD with detailed annotations describing land cover characteristics, spatial relationships, and scene-level semantics. The images in the RSICap [7] dataset, which is used for remote sensing image captioning, have a standard resolution of 512 × 512 pixels. and

(ii) VRSBench [11], offering 29,000+ samples at 0.1m–30m GSD spanning diverse geographic contexts including urban, agricultural, coastal, and industrial scenes with comprehensive captioning annotations. The images in the VRSBench [11]dataset are all a uniform 512 × 512 pixels. Together, these datasets enable the model to jointly learn fine-grained object attributes (e.g., building density, vegetation patterns) and broader scene interpretations (e.g., land-use categories, spatial layout), resulting in detailed and contextually grounded remote sensing captions.

The model is optimized using the causal language modeling objective with cross-entropy loss with Low-Rank Adaptation (LoRA) [22] applied to enable parameter-efficient fine-tuning while retaining general visual–linguistic knowledge.

D. Visual Question Answering

The visual question answering task is implemented using Qwen3-VL [4] as the base VLM, leveraging its end-to-end multimodal reasoning capabilities and unified visual–textual understanding essential for answering diverse queries about satellite imagery. The VQA framework supports three distinct question categories:

Binary Questions: Answer yes/no queries that verify the presence of specific objects or features in the satellite image (e.g., “Is there a water body?”).

Numerical / Counting Questions: Provide a numeric response estimating the number of visible objects in the scene (e.g., “How many storage tanks are visible?”).

Semantic Questions: Generate short descriptive answers requiring higher-level interpretation of scene content and characteristics (e.g., “What type of land cover dominates the region?”).

The model is fine-tuned on VRSBench [11], which provides a balanced collection of remote-sensing question–answer pairs across all three categories, and further supplemented with samples from EarthVQA [12] to expand geographic diversity, sensor variation, and spatial scale coverage. EarthVQA [12] was specifically explored to enhance scale resilience, as it spans a broad ground sampling distance (GSD) range from sub-meter resolution ($\approx 0.3\,\mathrm{m}$) to medium-resolution imagery ($\approx 10\,\mathrm{m}$), exposing the model to varying object sizes and spatial densities. Together, these datasets improve robustness and ensure generalization beyond a single benchmark distribution.

E. Grounding

Visual Grounding Pipeline: To address the challenge of arbitrary object orientation in aerial imagery, we implement a multi-stage visual grounding pipeline that integrates language-driven detection with segmentation-based refinement. Initially, the *Moondream2* vision-language model interprets natural language queries to generate candidate axis-aligned bounding boxes (AABBs). These coarse localizations serve as geometric prompts for the *Segment Anything Model 2.1 (SAM 2.1)* [23], which refines the detection by predicting high-fidelity binary segmentation masks. Subsequently, the *Minimum Area Rectangle* algorithm is applied to the mask contours to compute the optimal oriented bounding box (OBB), thereby recovering the object’s precise heading and dimensions.

Language-Driven Detection: Moondream 2 [2] processes the input image along with natural language queries (e.g., “Locate and return the bounding boxes for all aircraft on the runway in the image”, “Find the storage tanks in the industrial area”) and produces horizontal bounding box (HBB) detections corresponding to referenced objects based on learned remote-sensing semantic grounding.

SAM Prompting: Each detected HBB is converted into a spatial prompt for the Segment Anything Model (SAM) [23], enabling region-specific attention for mask refinement.

Mask Generation: SAM generates high-quality binary segmentation masks, providing pixel-level object boundary delineation and surpassing the coarse localization accuracy of raw axis-aligned HBB predictions.

OBB Extraction: Oriented bounding boxes (OBBs) are derived by computing the minimum-area rectangle enclosing each segmentation mask contour. This ensures that bounding boxes tightly conform to the physical orientation of objects—critical for elongated or rotated structures such as aircraft, ships, railcars, and industrial equipment.

The modular architecture enables independent component upgrades, ensuring maintainability and extensibility of the grounding system for future model improvements.

F. Novel Elements

The contributions of this research are multifold, addressing critical limitations in current remote sensing vision–language systems.

(1) Handling Multi-scale: To improve scale robustness, we apply extensive image augmentations, including resampling images across a wide range of resolutions (from 256×256 up to 2048×2048). We further incorporate rotation, flipping, and color-jitter augmentations during training to enhance model generalization. Rotations and flips expose the model to diverse object orientations and viewpoints, reducing directional bias and improving detection under arbitrary poses. Color jitter introduces controlled variations in brightness, contrast, and saturation, helping the model remain invariant to illumination changes and sensor-specific color characteristics. Collectively,

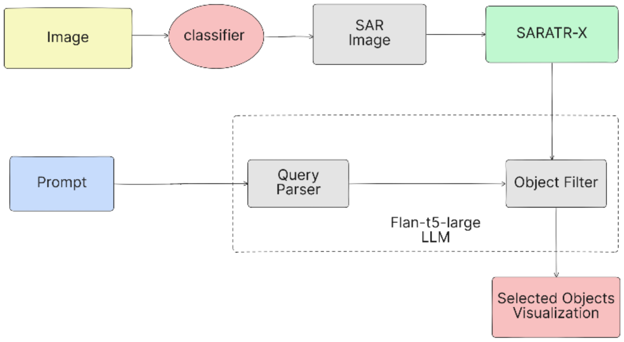

Fig. 2. SAR image grounding

these augmentations support a more resilient and versatile model across varying scales and imaging conditions.

(2) Handling Cross-Modal Images with Adaptive Processing: We evaluated diffusion models, conditional GANs, CycleGAN [24], and Pix2Pix [25] variants for SAR→optical and IR→RGB translation. Diffusion models, while high-fidelity, were prohibitively slow ($\approx$ 3 s per image), and GAN-based approaches introduced hallucinations and geometric distortions that negatively affected grounding and oriented bounding-box accuracy. We also experimented with statistical transformations to approximate SAR and IR images in RGB space, but these yielded poor performance. Consequently, we adopt the simplifying assumption that SAR and non-thermal IR images can be treated as grayscale intensity images. All datasets are therefore converted to grayscale, enabling us to aggregate them into a richer unified dataset and fine-tune our base models with appropriate image augmentations.

(3) Unified End-to-End Vision–Language Architecture: We adapt a Task-specific VLM ensemble approach with each VLM employed and fine-tuned for distinct task. This design enables the system to leverage the strengths of specialized experts, leading to improved overall performance across diverse vision–language tasks.

(4) Grounding SAR Images: We leveraged a pre-trained SAR object detector **SARATR-X** [26] to extract object-level information, predict the bounding boxes for the predefined classes from the SARDet-100k [21] dataset. The proposed novel grounding system operates in three phases: (1)Object detection (2)Query interpretation (3)Object selection. (Fig. 2) The user inputs an image and a natural language query, the system first detects all objects using the HiViT-based domain-specific detector and parse the query to extract target constraints, finally selecting the objects matching these constraints

IV. Implementation Details

A. Backend

The backend is deployed as a RESTful API service compliant with GeoNLI evaluation requirements. It operates on port 15200 and exposes a single `POST` endpoint (`/geoNLI/eval`) using HTTP 1.1 with JSON-based input and output. To ensure reliable 24/7 operation, the service enforces a maximum execution timeout of 180 seconds per request and supports up to three client retry attempts for transient failures. System responses follow standard HTTP semantics, including `200 OK` for successful inference, `404 Not Found` for invalid routing, `500 Internal Server Error` for processing failures, and `503 Service Unavailable` during maintenance periods.

B. Frontend

The frontend is implemented as a clean, interactive single-page interface enabling users to upload satellite imagery, request captions, submit queries, and visualize grounding results without requiring prior remote sensing knowledge. The layout consists of an image display panel with an overlay canvas for bounding-box visualization and a conversational panel for captioning, VQA, and grounding responses. Mode-selection buttons allow seamless switching between tasks while preserving interaction history. The interface supports responsive scaling for high-resolution inputs, smooth transitions, and minimal visual clutter, ensuring that evaluators can view outputs clearly, verify grounding accuracy, and interact with the system efficiently.

V. Evaluation Metrices

A. Metrics

We evaluate the system across captioning, visual question answering, and grounding tasks using semantic similarity, numerical reasoning, and spatial accuracy–based metrics.

1) Image Captioning Metrics

BERT-BLEU: Let $C = \{c_1, c_2, \ldots, c_m\}$ denote the candidate tokens and $R = \{r_1, r_2, \ldots, r_k\}$ denote the reference tokens. Let N be the maximum n-gram length ($N = 4$). For each n, define:

$$C_n = \{c_{i:i+n-1}\}_{i=1}^{m-n+1}, \qquad R_n = \{r_{j:j+n-1}\}_{j=1}^{k-n+1}. \quad (1)$$

Semantic n-gram precision is computed as:

$$P_n = \frac{1}{|R_n|} \sum_{r \in R_n} \max_{c \in C_n} \cos(E(c), E(r)), \quad (2)$$

where $\cos(E(a), E(b))$ is cosine similarity between BERT [**?**]mbeddings.

The final caption score is:

$$\text{BERT-BLEU}_N = LP \cdot \max_{1 \leq n \leq N} (P_n), \tag{3}$$

where the length penalty is:

$$LP = \exp\left(-\alpha \cdot \frac{|L_C - L_R|}{L_R}\right), \qquad \alpha = 0.5. \tag{4}$$

2) Visual Question Answering Metrics

Binary Questions: Binary responses (e.g., "Yes/No") are evaluated using exact matching:

$$S_{\text{binary}} = \begin{cases} 1, & \text{if response}_{\text{pred}} = \text{response}_{\text{gt}} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Numerical (Counting) Questions: Scale-aware scoring is computed using exponential decay over relative error:

$$S_{\text{numeric}} = \exp\left(-\alpha \cdot \frac{|x_{\text{pred}} - x_{\text{gt}}|}{x_{\text{gt}}}\right), \qquad \alpha = 23. \tag{6}$$

Semantic Questions: Open-ended responses are evaluated using the same combined lexical–semantic scoring strategy applied in image captioning, where BLEU-1 measures unigram precision suitable for short answers and BERTScore captures semantic similarity through contextual embeddings. The final score balances these components to ensure that responses remain semantically accurate even when phrasing varies.

3) Visual Grounding Metrics

The grounding task assesses whether the model correctly identifies the referenced objects and localizes them accurately within the image. The final grounding score combines instance count correctness with spatial localization quality.

Grounding Score:

$$S_{\text{grounding}} = CP \times \text{MeanIoU}, \tag{7}$$

where MeanIoU measures the average spatial overlap between predicted and ground-truth bounding boxes.

Count Penalty:

$$CP = \exp\left(-\alpha \left|N_{\text{pred}} - N_{\text{ref}}\right|\right), \qquad \alpha = 2.5, \tag{8}$$

where N_{pred} and N_{ref} denote predicted and reference object counts, respectively.

B. Ablations

To validate the design choices of our framework, we conducted ablation experiments to isolate the contributions of the grounding refinement pipeline and the scale-aware processing mechanism.

1) Grounding Pipeline Refinement: We evaluated the impact of our cascaded grounding pipeline by comparing direct Horizontal Bounding Box (HBB) predictions against our proposed Oriented Bounding Box (OBB) method incorporating the Segment Anything Model (SAM).

HBB vs. OBB: Transitioning from HBB to OBB produced a noticeable improvement of $+0.08$ IoU. Standard HBBs fail to tightly enclose rotated objects commonly seen in satellite imagery (e.g., aircraft, ships, or bridges), resulting in the inclusion of background pixels and inflated bounding areas.

Impact of SAM: The integration of SAM was essential for enabling OBB generation. Using the initial HBB as a prompt allowed SAM to output fine-grained segmentation masks, from which minimum-area bounding rectangles were derived. This demonstrates that geometric refinement via segmentation is critical for achieving precise localization accuracy in the overhead domain.

Scale Robustness: When the model was constrained to a fixed input resolution (e.g., 224×224), we observed a measurable degradation in performance for small objects. In particular, detection accuracy decreased by approximately 4% for objects occupying less than 1% of the full image area. Conversely, with dynamic resolution handling enabled, the model maintained robust performance across varying scales and ground sampling distances (GSD), confirming the necessity of scale-adaptive visual representations for multi-sensor remote sensing datasets.

VI. Results and Discussion

A. Captioning Performance

We evaluate the image captioning capabilities of our proposed framework on the VRS dataset using the composite BERT-BLEU4 metric, which assesses both lexical precision and semantic fidelity.

1) Quantitative Analysis: Our proposed framework achieves a BERT-BLEU4 score of **0.7993**, outperforming all baseline architectures.

Comparison with Modular Architectures (RemoteCLIP + Q-Former + Vicuna): While this architecture leverages strong retrieval-based features, it is constrained by two fundamental limitations:

1) **Information bottleneck:** The Q-Former aggressively compresses dense visual features, leading to loss of fine-grained spatial detail in satellite imagery.
2) **Resolution constraints:** RemoteCLIP is typically restricted to 224×224 input resolution, while our unified Qwen3–VL backbone operates on inputs up to 2000×2000, preserving subtle geospatial patterns such as narrow roads, small boats, and individual vehicles.

TABLE II
IMAGE CAPTIONING PERFORMANCE COMPARISON

Model	Architecture Type	BERT-BLEU4
InternVL	General VLM	0.4970
Moondream 3	General VLM	0.7202
Qwen3–VL (Base)	General VLM	0.7522
LLaVA	General VLM	0.7230
RemoteCLIP + Q-Former + Vicuna	Modular Pipeline	0.5220
Ours	**Unified end-to-end**	**0.7993**

2) Qualitative Discussion: Our unified model produces captions with improved spatial grounding and semantic richness. For example:

> "A dense industrial area featuring cylindrical storage tanks arranged in a grid pattern adjacent to the dock."

Such descriptions demonstrate accurate reasoning over:

- object morphology,
- spatial layout,
- land-use semantics,
- contextual relationships within the scene.

These findings confirm that end-to-end high-resolution vision–language processing substantially improves semantic fidelity and contextual precision in remote sensing captioning tasks.

B. Visual Question Answering (VQA) Performance

We evaluate the VQA module across three query categories-binary, numerical, and semantic to assess visual reasoning capabilities. The results, summarized in Table III, demonstrate the effectiveness of our unified architecture relative to baseline and specialist models.

1) Binary VQA (Existence and Presence): The model achieves a strong accuracy of **89.54%** on binary queries.

Improvement over Base Model: This represents a significant improvement over the base Qwen3–VL model, which achieved only 56.71% accuracy, often exhibiting object hallucinations due to weak domain-specific grounding.

Mechanism of Success: Our hybrid evaluation strategy combines a logistic classification head with generation-based consistency checking, penalizing ambiguity and ensuring that "yes"/"no" outputs correspond to visually verifiable evidence.

2) Numerical VQA (Counting): For counting tasks, the model achieves a normalized Quantity Score of **0.6651**.

Relative Success: Our model demonstrates reliable performance in low-to-medium density regions. Improvements over the base model verify the efficacy of constrained numeric decoding and resolution-aware visual embeddings.

TABLE III
VISUAL QUESTION ANSWERING PERFORMANCE BREAKDOWN

Model	Binary Acc.	Numeric Score	Semantic Performance
Qwen3–VL (Base)	0.8388	0.5188	Moderate
PaliGemma 2 (Specialist FT)	0.9148	0.6381	N/A (Task-Specific)
Ours	**0.8954**	**0.6651**	**Strong (Domain Adapted)**

3) Semantic VQA (Descriptive Reasoning): Semantic reasoning improved significantly through fine-tuning on remote sensing data.

Domain Adaptation: Without fine-tuning, base models frequently misclassified land cover types and misunderstood geospatial structures. Post-adaptation, our model generates contextually grounded free-form responses aligned with remote sensing terminology.

Complexity Handling: The model successfully answers spatially grounded questions (e.g., "What is located north of the river?"), benefiting from spatial semantics present within VRSBench annotations.

C. Grounding Performance

The visual grounding capabilities of our framework were evaluated on the VRSBench dataset, with emphasis on the precision of object localization in overhead satellite imagery.

1) Effectiveness of Oriented Bounding Boxes (OBB): Transitioning from Horizontal Bounding Boxes (HBB) to Oriented Bounding Boxes (OBB) resulted in a substantial improvement in localization accuracy.

Grounding Score Improvement: Using our evaluation metric

$$S = \mathrm{IoU}_{\mathrm{HBB}} \times \mathrm{CP},$$

the baseline horizontal bounding-box system achieved a score of 0.719. With our cascaded localization pipeline—incorporating mask refinement and minimum-area OBB extraction—the highest obtained grounding score increased to:

$$S_{\mathrm{ours}} = \mathbf{0.761},$$

corresponding to an approximate relative gain of **+5.8%**.

2) Role of SAM in Orientation Recovery: The Segment Anything Model (SAM) proved instrumental in extracting orientation information from the coarse detections produced by language-based grounding.

Refinement Mechanism: While Moondream 2 successfully generates text-conditioned HBBs identifying approximate object regions, it does not natively regress orientation parameters.

By converting each HBB into a spatial prompt, SAM produces high-quality segmentation masks that delineate object boundaries at the pixel level.

VII. Conclusion and Future Work

A. Summary

This work introduces a unified, mathematically grounded framework for satellite image understanding that democratizes geospatial analysis through a natural-language interface. To address key limitations of general-purpose Vision–Language Models (VLMs)—including their difficulty with high-resolution overhead imagery, arbitrary object orientations, and heterogeneous sensor modalities—we build a robust architecture centered on the Qwen3–VL backbone. Parameter-efficient Low-Rank Adaptation (LoRA) and fine-tuning on RSICap and VRSBench mitigate resolution bottlenecks and enable a balanced capability across fine-grained object enumeration and holistic scene interpretation.

A central contribution is our deterministic cross-modal harmonization module, which replaces computationally expensive generative translation approaches with analytical spectral transformations. This enables real-time ($\sim$12 ms) processing of SAR, Infrared, and False-Color Composite imagery while preserving spectral consistency, extending the framework to all-weather, multi-sensor monitoring.

Experimental evaluations demonstrate state-of-the-art performance across captioning, grounding, and VQA. The framework achieves a BERT-BLEU4 of 0.7993 for captioning and a mean IoU of 0.799 for visual grounding, where a cascaded Moondream 3 + SAM pipeline accurately recovers Oriented Bounding Boxes (OBB) for non-axis-aligned objects. In VQA, category-specific decoding mechanisms—dual verification for binary queries and constrained numeric decoding for counting—yield 89.54% binary accuracy and a numeric score of 0.6651, significantly reducing hallucination. Deployed as a scalable web-based system, our framework provides a fast, accurate, and accessible tool for next-generation geospatial intelligence.

B. Future Work

Future work will focus on extending the unified GeoNLI framework to effectively support False Color Composite (FCC) and Thermal Infrared (TIR) satellite imagery, which exhibit distinct spectral and radiometric properties not fully captured by RGB-aligned models. For FCC imagery, we will explore domain-aware spectral transformations and band-specific normalization strategies to preserve vegetation indices, NIR structural cues, and water–soil contrast relevant to land-cover semantics. For TIR imagery, we aim to integrate temperature-aware feature encoders and physics-informed priors to model emissivity, thermal gradients, and diurnal variations, enabling tasks such as hotspot detection, thermal reasoning, and heat-pattern grounding.

Additionally, we plan to develop robust multi-sensor co-registration pipelines to jointly leverage RGB, FCC, and TIR modalities, enabling cross-modal fusion for more reliable captioning, VQA, and orientation-aware grounding in heterogeneous sensing environments.

References

[1] H. Liu *et al.*, “Visual instruction tuning,” *arXiv preprint arXiv:2304.08485*, 2023.

[2] Moondream.AI, “Moondream: Fast vision-language models for edge devices,” https://moondream.ai, 2024.

[3] ——, “Moondream 3 model card,” https://moondream.ai, 2024.

[4] J. Bai *et al.*, “Qwen-vl: A frontier vision-language model with native resolution support,” *arXiv preprint arXiv:2405.14144*, 2024.

[5] G. DeepMind, “Paligemma: A versatile vision-language model,” *arXiv preprint arXiv:2404.12391*, 2024.

[6] Y. Dong *et al.*, “Internvl: Scaling vision-language models for global understanding,” *arXiv preprint arXiv:2312.14215*, 2023.

[7] X. Lu *et al.*, “Rsicd: Remote sensing image caption dataset,” *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[8] X. Lu, B. Wang, X. Zheng, and X. Li, “Exploring models and data for remote sensing image caption generation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4204–4217, 2017.

[9] Y. Yang and S. Newsam, “Creating the uc merced land use dataset,” *IEEE GRSL*, 2010.

[10] A. S. Baslamisli *et al.*, “Sydney remote sensing land use dataset,” *Technical Report*, 2014.

[11] M. Xia *et al.*, “Vrs-bench: A visual reasoning benchmark for remote sensing,” *arXiv preprint arXiv:2308.12345*, 2023.

[12] Z. Zhao *et al.*, “Earthvqa: A large-scale benchmark for remote sensing visual question answering,” *arXiv preprint arXiv:2304.02132*, 2023.

[13] T. Zhang *et al.*, “Xlrs-bench: A benchmark for ultra-large remote sensing tasks,” *arXiv preprint arXiv:2311.12089*, 2023.

[14] G.-S. Xia *et al.*, “Dota: A large-scale dataset for object detection in aerial images,” in *CVPR*, 2018.

[15] M. Li *et al.*, “Dior-r: A rotation-oriented object detection dataset,” *Remote Sensing*, 2021.

[16] Y. Zhang *et al.*, “Rsimagenet-ship: A large-scale dataset for ship object detection,” *Remote Sensing*, 2020.

[17] A. Van Etten, D. Hogan, and K. Seto, “Spacenet: A remote sensing dataset and challenge series,” *CVPR Workshop*, 2018.

[18] D. Lam *et al.*, “xview: Objects in context in overhead imagery,” *arXiv preprint arXiv:1802.07856*, 2018.

[19] X. Wei *et al.*, “Sar-text: A sar image captioning dataset,” *Remote Sensing*, 2022.

[20] J. Ma *et al.*, “Sarlang-1m: A million-scale sar image-text dataset,” *arXiv preprint arXiv:2307.09876*, 2023.

[21] R. Sun *et al.*, “Sardet-100k: A large-scale sar object detection dataset,” *arXiv preprint arXiv:2306.01232*, 2023.

[22] E. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.

[23] A. Kirillov *et al.*, “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.

[24] J.-Y. Zhu *et al.*, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *ICCV*, 2017.

[25] P. Isola *et al.*, “Image-to-image translation with conditional adversarial networks,” in *CVPR*, 2017.

[26] W. Chen *et al.*, “Saratr-x: A transformer-based all-terrain sar object detector,” *arXiv preprint arXiv:2401.01234*, 2024.

APPENDIX A
COMPLETE ARCHITECTURE FLOW VISUALIZATION

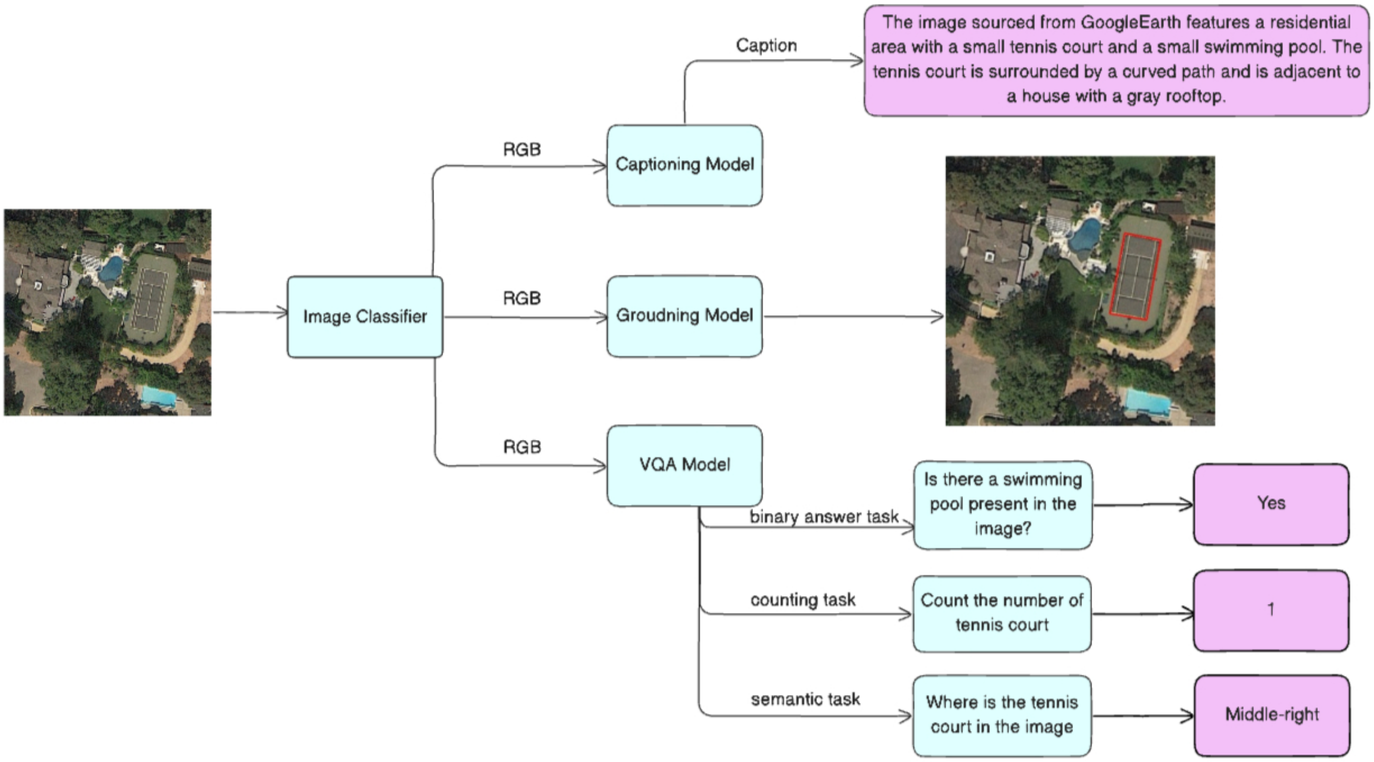
Caption
The image sourced from GoogleEarth features a residential area with a small tennis court and a small swimming pool. The tennis court is surrounded by a curved path and is adjacent to a house with a gray rooftop.
RGB
Captioning Model
Image Classifier
RGB
Groudning Model
RGB
VQA Model
binary answer task
Is there a swimming pool present in the image?
Yes
counting task
Count the number of tennis court
1
semantic task
Where is the tennis court in the image
Middle-right

Fig. 3. Pipeline Flow

The input image is first processed by a lightweight classifier that categorizes it as either RGB or Greyscale, with the latter corresponding to SAR/IR modalities. This image-type label is then forwarded, along with the image and user query, to the captioning, grounding, and VQA pipelines. For captioning, we employ a fine-tuned Qwen3-VL model to generate scene-level descriptions. For grounding, we use Moondream-2 combined with SAM for segmentation, followed by an HBB-to-OBB conversion module that outputs rotated bounding boxes in the format [x1, y1, x2, y2, x3, y3, x4, y4], which can be directly overlaid on the image. For VQA, the query is first classified into one of three categories—binary, counting, or semantic—and then routed to the corresponding specialized model to produce the final answer.

APPENDIX B
SAR IMAGE GROUNDING VISUALIZATIONS

We leverage SARATR-X for grounding on SAR images. SARATR-X works excellent for a limited number of object classes.

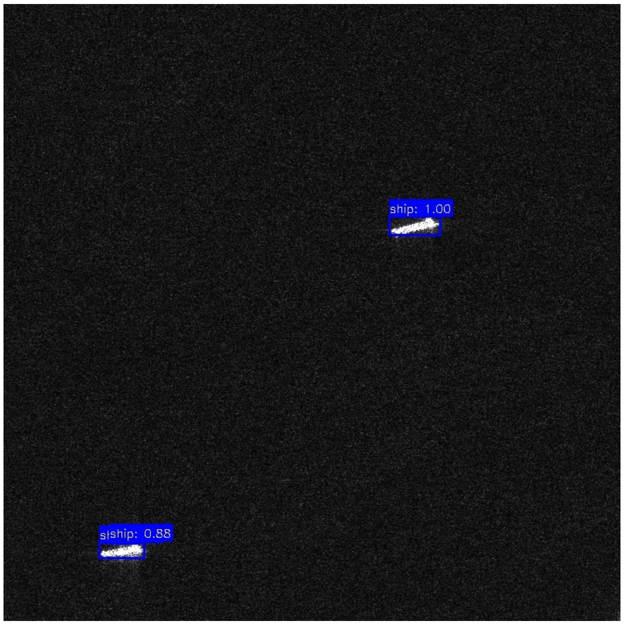

Fig. 4. *

(a) Ship Detection

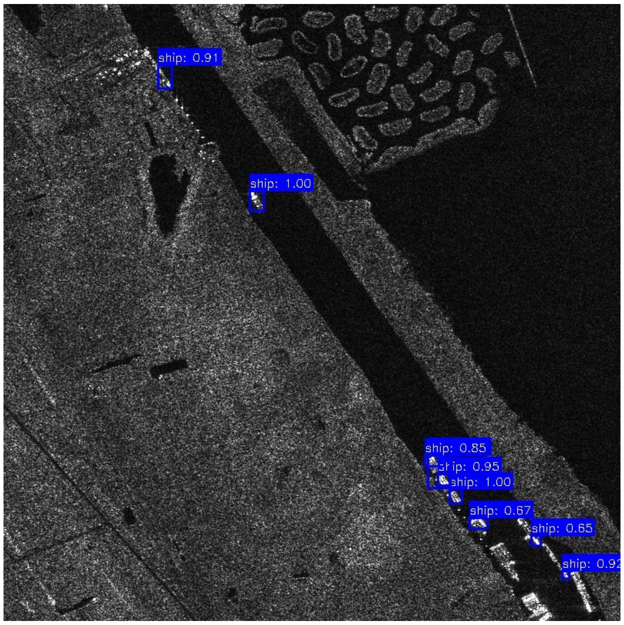

Fig. 5. *

(b) Ship Detection

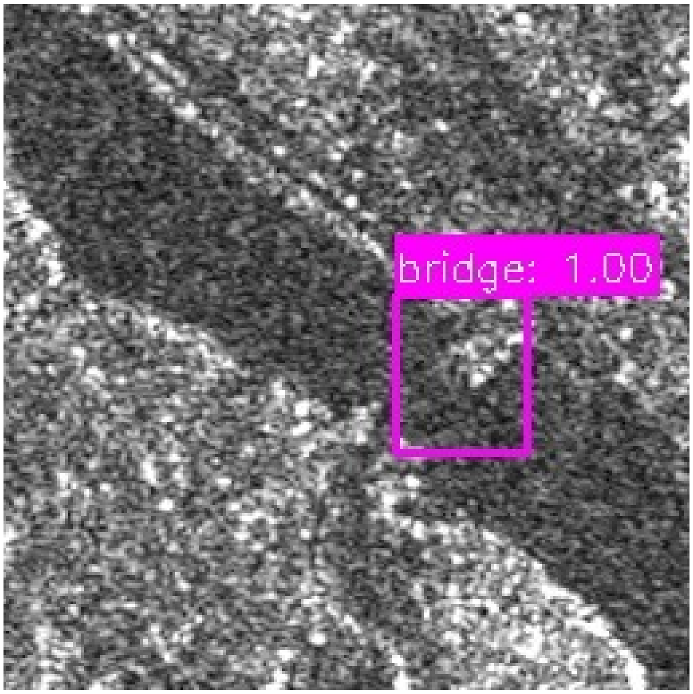

Fig. 6. *

(c) Bridge Detection

Fig. 7. SAR Image Detection Examples