DRISHTI : Deep Remote-Sensing Intelligence for Semantic Hybrid Text-Image Understanding

***Abstract*—The proliferation of satellite constellations and high-resolution aerial platforms has generated unprecedented volumes of Remote Sensing (RS) imagery, yet effective natural language interaction with such data remains a significant challenge. We introduce DRISHTI (Deep Remote-sensing Intelligence for Semantic Hybrid Text-Image understanding), a unified Vision-Language Model (VLM) framework enabling intuitive natural language interaction with RS imagery across varied resolutions, sensor modalities, and downstream tasks. DRISHTI addresses three critical gaps in existing RS-VLM research. First, we contribute DRISHTI-GCV, a large-scale, difficulty-aware dataset suite comprising approximately 180,000 curated samples spanning Grounding, Captioning, and Visual Question Answering (VQA) tasks, aggregated from multiple RS benchmarks with explicit annotations for image resolution, query complexity, and object scale. Second, we propose a two-stage Curriculum Learning paradigm: the first stage adapts a Low-Rank Adaptation (LoRA)-tuned VLM backbone to general RS semantics; the second stage specializes the model for fine-grained tasks using Direct Preference Optimization (DPO) alignment. Third, we introduce a dedicated numeric question handling pipeline wherein a lightweight question-type router dispatches counting and area estimation queries to a Segment Anything Model 3 (SAM3)-based module with pyramidal tiling. Comprehensive experiments on VRSBench, RSVQA, FIT-RSFG, and real-world satellite/aerial RGB, SAR, IR, and False Color Composite imagery demonstrate that DRISHTI consistently matches or surpasses state-of-the-art RS-VLMs, achieving substantial gains in caption quality (+40% BERT-BLEU), VQA accuracy (+21% over GeoChat), and counting performance while maintaining interactive latency.**

I. Introduction

The advent of large-scale Earth Observation (EO) programs, including the European Space Agency's Copernicus Sentinel constellation, NASA's Landsat Continuity Mission, and commercial very-high-resolution (VHR) platforms such as Maxar WorldView and Planet's SkySat fleet, has resulted in the continuous generation of petabytes of remote sensing (RS) imagery every year. This deluge of geospatial data underpins critical applications ranging from urban planning and infrastructure monitoring to disaster response, agricultural assessment, and climate change analysis. However, the traditional paradigm of manual image interpretation by domain experts cannot scale to meet increasing operational demands, motivating the development of automated and semi-automated analysis pipelines.

While sensing capabilities continue to advance rapidly, accessibility to information extracted from RS data remains limited by the need for specialized tools and technical expertise. As a result, non-expert users, including policymakers, planners, and first responders, often face delays in deriving actionable insights from satellite imagery. This growing gap between data collection and knowledge extraction highlights the need for more intuitive and scalable mechanisms for interacting with complex geospatial data.

Recent progress in Vision-Language Models (VLMs), liu2024visual has demonstrated the feasibility of querying visual content using natural language, enabling tasks such as image captioning, visual grounding, and Visual Question Answering (VQA). Large-scale architectures such as CLIP, BLIP-2, and LLaVA (Large Language and Vision Assistant) have achieved strong performance on natural-image benchmarks by aligning textual and visual representations during pre-training. However, applying these models directly to RS imagery introduces significant challenges. Satellite images differ fundamentally from everyday photographs in scale, sensor modality, scene geometry, and object density, leading to frequent hallucination, imprecise localization, and weak numerical reasoning when existing VLMs are deployed in this domain.

Remote sensing scenes introduce several unique challenges. Data spans multiple spatial resolutions, from sub-meter optical imagery to coarse multispectral products, requiring explicit multi-scale reasoning. Scenes often contain thousands of arbitrarily oriented, densely packed objects that strain object-centric representations. Furthermore, RS imagery extends beyond RGB data and includes modalities such as Synthetic Aperture Radar (SAR), thermal infrared, and multispectral or hyperspectral sensors, each with distinct physical characteristics. Critically, many RS applications require quantitative spatial reasoning, including object counting, size estimation, and spatial relationship analysis, all of which remain underdeveloped in most off-the-shelf VLMs.

To address these limitations, we introduce **DRISHTI** (Deep Remote-sensing Intelligence for Semantic Hybrid Text-Image Understanding), a unified framework designed to bring natural language reasoning into high-resolution, multimodal satellite imagery. DRISHTI integrates language grounding, dense object understanding, and quantitative analysis within a single architecture. Rather than treating remote sensing as a simple domain adaptation problem, our approach explicitly models scale variation, object density, and geometric structure. As a result, DRISHTI enables accurate and interpretable interaction with geospatial data through natural language.

Our key contributions are summarized as follows:

- **DRISHTI-GCV: A Difficulty-Aware RS Dataset.** We introduce **DRISHTI-GCV**, a unified dataset of $\sim$170K samples for grounding, captioning, and VQA, integrating multiple RS benchmarks with annotations for resolution

(224^2-2048^2), query difficulty, object size, and spatial complexity. This enables curriculum learning and fine-grained evaluation across diverse RS conditions.

- **Curriculum Learning with LoRA and DPO Alignment** We propose a two-stage training strategy where LoRA-based fine-tuning adapts the VLM to RS semantics, followed by DPO-based specialization on difficult samples using a BERT-BLEU reward to reduce hallucinations. This allows an 8B model to outperform larger baselines.
- **Grounding via the Adaptive Hierarchical Grounding Net- work (AHG-Net)** We introduce a query-routing mechanism that directs numeric questions to a SAM3-based segmentation pipeline with multi-scale tiling and geometric post-processing, enabling accurate counting and area estimation beyond LLM-only reasoning.

II. Literature Review

A. Vision-Language Models for Remote Sensing

Adapting Vision-Language Models (VLMs) to remote sensing (RS) is challenging due to ultra-high resolution imagery, dense object layouts, oriented targets, and complex spatial relationships. Early RS-specific models extend general-purpose VLMs with spatial awareness and high-resolution encoding. GeoChat [1] builds upon LLaVA with a high-resolution CLIP encoder and region-based conversational grounding. LHRS-Bot [2] improves spatial reasoning using curriculum-based alignment over the LHRS-Align corpus, while SkyEyeGPT [3] unifies captioning, VQA, and grounding by injecting topic tokens into an EVA-G + LLaMA-2 backbone.

These models rely on RS-targeted datasets such as VRS-Bench [4], which provides image patches annotated with captions, oriented bounding boxes, and over 120k VQA pairs, complemented by RSICD [5] and RSVQA [6] for multi-resolution supervision. To address large image sizes, pyramidal tiling strategies preserve both fine detail and global context more effectively than uniform splitting, which tends to break long-range spatial dependencies.

B. High-Resolution Multimodal Representation Learning

The Qwen family provides a strong foundation for high-resolution multimodal reasoning. Qwen3-VL [7] introduces dynamic resolution handling and Multimodal Rotary Position Embeddings (M-RoPE), enabling fine-grained spatial encoding across image and video inputs. This allows the model to preserve geometric structure and object locality, which is essential in RS imagery.

Dynamic resolution enables computation to focus on relevant image regions without fully upsampling large images, improving efficiency while maintaining spatial fidelity. However, most existing VLMs still focus on RGB imagery and lack unified support for SAR, multispectral, and infrared modalities, limiting real-world RS applicability.

C. Captioning and Visual Question Answering in Remote Sensing

Remote sensing captioning and VQA demand structured semantic understanding, quantitative reasoning, and precise spatial description. Unlike natural images, RS data involves complex terrain interactions and object distributions. Benchmarks such as RSICD [5] and VRSBench [4] facilitate training, but hallucination and spatial inconsistency remain common failure modes.

Curriculum-based instruction tuning in LHRS-Bot [2] improves robustness, yet linguistic alignment remains underexplored in RS. Techniques such as Direct Preference Optimization (DPO) and BERT-BLEU scoring [8] have shown promise in improving semantic faithfulness and fluency but are rarely applied in RS-specific pipelines.

D. Visual Grounding and Open-Vocabulary Detection

Open-vocabulary grounding enables language-conditioned localization in large-scale imagery. Rex-Omni [9] introduces universal proposal networks for dense detection, while VLM-FO1 [10] connects LVLMs to detection heads. LiSAT [11] enables language-driven pixel-level reasoning for counting and area estimation, addressing key quantitative tasks in RS.

Grounding DINO [12] advances open-set detection via a cross-modal decoder and language-guided query selection, improving generalization to unseen classes. However, most grounding models struggle with cluttered scenes and rotated targets that dominate aerial imagery, motivating domain-specialized approaches.

III. Dataset Preparation: DRISHTI-GCV

DRISHTI-GCV was constructed to address the fragmentation of remote-sensing (RS) vision-language datasets, which typically focus on a single task, operate at fixed resolutions, or fail to model difficulty. Our dataset unifies captioning, grounding, and VQA across RGB, SAR, and infrared (IR) modalities, paired with explicit resolution metadata, difficulty annotations, and multi-scale object statistics. The resulting multi-resolution, difficulty-aware suite (Table I) supports training DRISHTI to handle tiny dense objects, heterogeneous sensor modalities, multi-resolution scenes, and numeric or spatially compositional queries.

TABLE I
DRISHTI-GCV DATASET COMPOSITION.

Task	Samples	Resolution Range
Captioning (Stage I)	∼22k	256^2-512^2
Captioning (Stage II)	∼20k	up to 2048^2
Grounding	∼60k	224^2-2048^2
VQA (Generalized)	∼30k	256^2-512^2
VQA (Specialized)	∼20k	512^2-2048^2
SAR (SARLANG)	∼15k	512^2
Infrared (GeoAI)	∼13k	1024^2
Total	**∼180k**	224^2-2048^2

A. Captioning Dataset Preparation

Caption annotations are sourced from VRSBench and Git-10M [13]. We normalize captions through lightweight text preprocessing (tokenization, lemmatization) and filter out ultra-short or off-topic descriptions.

1) Generalized Captioning Split (Stage I): The Stage I split targets broad RS semantics at moderate resolution. We select $\sim$22k image-caption pairs (12k from VRSBench, 10k from Git-10M) spanning 256^2-512^2. Qwen3-VL-32B evaluates each sample using three criteria: **(i)** object-caption alignment, **(ii)** semantic density, and **(iii)** n-gram repetition: yielding a difficulty score from 1 to 4. We retain samples rated 1-3 based on object size, clutter, and spatial-constraint complexity.

2) Specialized Captioning Split (Stage II): Stage II focuses on complex, high-density scenes. We oversample images with many small objects, strong spatial relations, and high visual clutter. Most retained captions fall within difficulty levels 3-4, and we additionally include upsampled high-resolution tiles ($\geq 1024^2$).

B. Grounding Dataset Preparation

Our grounding suite ($\sim$30k samples) integrates VRSBench (512$\times$512 tiles with oriented bounding boxes (OBBs) and referring expressions), OPT-RSVG (224$\times$224 and 800$\times$800 crops) [14], RSVG/RSVG-HR [15], and DIOR/DOTA-v2 [16], [17], spanning resolutions from 224^2 to 2048^2.

DIOR and DOTA provide only a few object classes and lack referring expressions; we therefore generate natural-language descriptions using GPT-5-mini to unify grounding supervision across datasets. To standardize localization quality, we construct consistent OBBs where originals are absent or axis-aligned. A hybrid pipeline combining SAM-3 box-prompted segmentation with classical CV methods (contours and `minAreaRect`) yields rotation-aware bounding boxes for all samples, ensuring coherent spatial annotations suitable for multi-scale grounding.

C. VQA Dataset Preparation

For VQA, we integrate VRSBench with RSVQA-LR/HR to build a difficulty-aware, multi-resolution QA corpus. VRS-Bench provides rich question-answer structures that capture spatial context and reasoning depth, while RSVQA adds both low- and high-resolution variants. Difficulty scores (1-5) assigned by Qwen3-VL-32B reflect constraint count, spatial-relational complexity, and required domain knowledge.

1) Generalized VQA Split: The generalized split comprises $\sim$30k QAs (9k numeric, 10.5k classification, 10.5k semantic), retaining only difficulty 1-3 samples for broad-coverage training.

2) Specialized VQA Split: The specialized split ($\sim$20k QAs) focuses on difficulty 4-5 questions. These samples come from high-resolution aerial tiles ($\geq 512^2$) where tiny structures remain resolvable. This subset targets fine-grained relational reasoning under clutter.

D. SAR and IR Dataset Preparation

We initially explored cross-modal translation methods such as Pix2Pix [18], ThermalGAN, and ControlNet-style diffusion [19] to convert SAR and IR/thermal imagery into RGB for reuse of natural-image encoders. However, these transformations proved unstable: they frequently hallucinated textures or over-smoothed gradients, degrading modality-specific structure critical for RS understanding.

1) SAR Preparation: To preserve SAR characteristics, we directly subsample 15k captioning and 15k VQA samples from SARLANG-1M [20], converting them into instruction-format pairs for multimodal alignment without altering radiometric properties.

2) Infrared (IR) Preparation: For IR, we prioritize principled dataset construction. Using Microsoft Planetary Computer's GeoAI interface, we retrieve geographically anchored scenes from the National Agriculture Imagery Program (NAIP), covering diverse U.S. regions at 0.3-1 m GSD with RGB+NIR bands. We extract all spectral channels and generate true-color, NIR-only, and false-color composites, followed by tiling into 1024×1024 crops to capture both global context and fine structures. From over 300 scenes, we assemble a large-scale IR corpus and annotate it with a GPT-5-mini prompting pipeline tailored to RS semantics, followed by selective human verification. This yields 12k VQA instances and 2k captions. Unlike cross-modal translation, this modality-specific approach preserves radiometric integrity and significantly improves captioning and VQA performance across IR modalities.

IV. Training and Inference Pipeline

We adopt a two-stage, curriculum-style pipeline built on Qwen-3-VL-8B-Instruct, chosen for its native support for high-resolution inputs via dynamic tiling and resolution-aware visual encoding, which matches the multi-scale nature of RS imagery. Stage I adapts the model to remote sensing using DRISHTI-GCV generalized mixed-task dataset (captioning + VQA), while Stage II performs task-specific specialization for captioning and VQA, including DPO-based alignment and a SAM3-powered, tool-augmented numerical Grounding module. Figure below summarizes the overall training pipeline.

A. Stage I: General Remote-Sensing Adaptation

The pretrained model **Qwen3-VL-8B-Instruct** is adopted as the base vision–language backbone, chosen for its support of high-resolution inputs and multimodal positional embeddings. A mixed remote-sensing dataset comprising image–caption pairs and image–question–answer triples from DRISTI-GCV's generalized split is used for supervised fine-tuning. The adaptation is performed via **LoRA**, updating only low-rank parameter matrices in attention and MLP layers, while freezing the base weights. At the end of this stage, a domain-adapted checkpoint is obtained, which serves as the foundation for all subsequent task-specific specialization.

B. Stage II-A: Caption-Specialized Curriculum

Task-specific SFT. From the generalized checkpoint, we specialize a captioning branch using the DRISHTI-GCV specialized caption split, which emphasizes semantically dense descriptions, multiple object types and relations, and multi-scale scenes (tiny objects to city-scale context) up to $\sim 2048^2$ resolution. We reuse the Stage I LoRA setup and optimizer, minimizing autoregressive cross-entropy over caption tokens.

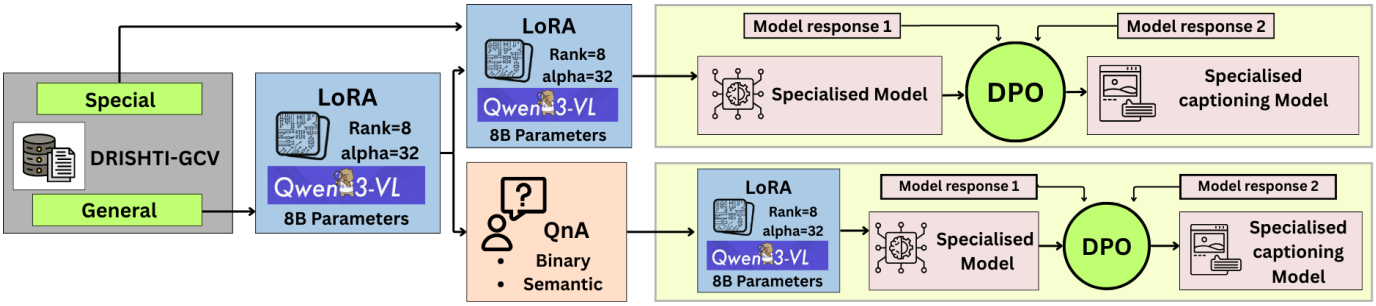


Fig. 1. Training Pipeline

C. Stage II-B: VQA-Specialized Curriculum

The VQA module proceeds in three steps:

1) **Specialized fine-tuning.** The RS-adapted backbone is fine-tuned on specialised subset of DRISTI-GCV covering semantic, yes/no, and numeric (counting/area) questions.
2) **Query routing.** At inference, a lightweight classifier (Qwen3-VL-30B) routes each question into one of three branches: *Numeric* $\rightarrow$ segmentation-based counting/area module, *Yes/No* $\rightarrow$ VQA head, *Semantic* $\rightarrow$ VQA head with optional preference-based refinement.
3) **Numeric reasoning.** For numeric queries: (i) generate a short segmentation prompt from the question, (ii) apply multi-scale segmentation via a tiling module + promptable segmenter (e.g. SAM3), (iii) fuse overlapping masks, (iv) compute object count and physical area (via GSD $\times$ pixel count), (v) return numeric answer in natural language.

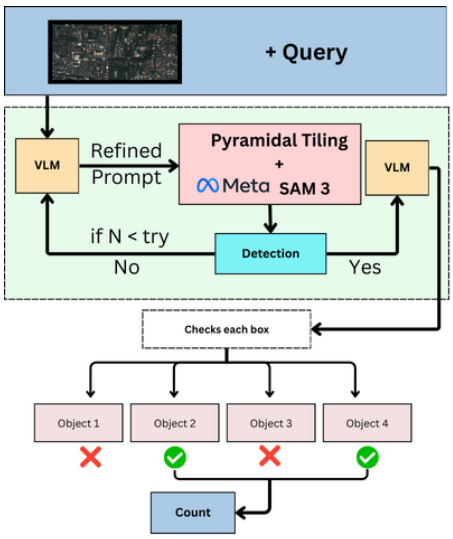


Fig. 2. Counting/Area Pipeline

D. Grounding via the Adaptive Hierarchical Grounding Network (AHG-Net)

We introduce the *Adaptive Hierarchical Grounding Network* (AHG-Net), a specialized architecture designed to bridge the gap between open-vocabulary segmentation and nuanced spatial reasoning. Addressing the challenges of scale variation and high-density mask outputs, the AHG-Net integrates a centroid-based optimization mechanism with a stochastic refinement policy. The grounding pipeline proceeds through five distinct stages:

1) **Subject-Reference Query Decomposition:** Complex queries often involve relative positioning that conflates target objects with their environment. To disambiguate these roles, we first decompose the input query q into two semantic constituents:
   - **Primary Subject (ϕ_{pri}):** The specific entity to be located (e.g., "the car to the left of a bridge").
   - **Secondary Subject (ϕ_{sec}):** The reference object establishing the spatial context (e.g., "car to the left of **a bridge**").

   This decomposition allows the system to treat the secondary subject purely as a geometric anchor, focusing the computationally expensive refinement solely on the primary target.
2) **Pyramidal SAM3 Segmentation:** Standard segmentation frequently misses small objects or fragments large ones. We address this via a pyramidal tiling strategy using the SAM3 encoder. The image is processed at multiple discrete resolution levels to generate a comprehensive candidate pool for both subjects:

$$\mathcal{M}_{\text{raw}} = \{(m_j, s_j)\}_{j=1}^{J}, \quad s_j \in [0, 1].$$

   This ensures that the initial proposal set captures both dominant structural elements and fine-grained details.
3) **Centroid-based Clustering for Dense Segmentation Masks:** In high-resolution settings, the pyramidal encoder often generates an excessive number of redundant candidates ($J \gg T_{\text{max}}$), creating a bottleneck for downstream processing. To optimize this, we employ a density-aware clustering strategy:
   - **Greedy box clustering:** Candidate bounding boxes are clustered in the 2D coordinate space using a greedy, centroid-based procedure. Starting from a randomly sampled box as the initial centroid, each remaining box is assigned to the nearest existing centroid (in terms of center-distance or $1 - \text{IoU}$);

if this distance exceeds a threshold, a new cluster is created. This process yields clusters $\{C_k\}$ of geometrically similar candidates.

- **Centroid selection:** For each cluster C_k, a representative *centroid mask* m_k^* is chosen by minimizing the distance to the cluster mean:

$$m_k^* = \arg\min_{m \in C_k} d\big(c(m), \bar{c}_k\big),$$

where $c(m)$ denotes the box center of m and $\bar{c}_k = \frac{1}{|C_k|}\sum_{m \in C_k} c(m)$ is the mean center of cluster C_k. This greedy centroid selection effectively selects closer candidates within each cluster.

This step splits the large number of input proposals into smaller optimized sets of potential Subjects that can then be classified by the VLM.

4) **Subject Confidence Estimation:** Within each cluster, boxes are viewed in the context of the full image and are classified into three bands — *High*, *Medium*, or *Low* confidence of being query-relevant subjects. High-confidence boxes whose structure and colour attributes closely match the described subject class are immediately added to a provisional *selected mask set*, while the remaining Medium- and Low-confidence boxes are retained for further refinement.
5) **Chunked VLM refinement over ambiguous proposals:** The union of Medium- and Low-confidence boxes is partitioned into manageable chunks of size K. For each chunk, the VLM is provided only the corresponding bounding-box crops and the textual query. It re-evaluates each candidate as High, Medium, or Low confidence subject. Low-confidence boxes are discarded. Medium-confidence boxes are then examined at finer scale by cropping around each box and querying the VLM again, allowing it to focus on local structure, colour, and other object attributes. Newly promoted High-confidence boxes from this stage are added to the selected mask set, progressively filtering out erroneous detections while retaining plausible subjects (both primary and secondary).
6) **Final spatial reasoning over primary and secondary subjects:** After the refinement loop, the selected mask set contains only subject-consistent candidates. The VLM orchestrator then performs spatial reasoning to determine which of these masks correspond to the primary and secondary objects specified in the query. Primary objects may be defined globally (e.g., "the northernmost stadium") or relationally with respect to secondary objects (e.g., "the ship to the left of the pier near the refinery"). Using these directional and relational cues, conflicting masks are eliminated and the subset of masks that best satisfies the query semantics is retained as the final grounded solution.

Since evaluating all ambiguous masks via a Vision-Language Model (VLM) is computationally prohibitive, we employ a **stochastic batching strategy**:

$$\mathcal{B}_k \sim \text{Uniform}(\mathcal{M}_{\text{amb}}).$$

Random subsets $\mathcal{B}_k$ are sampled and fed to the multimodal controller, **Qwen3-VL-30B-A3B-Instruct** for classification. The VLM functions as a semantic discriminator, analyzing the visual context, the spatial anchor $\tilde{m}_{\text{sec}}$, and the batch simultaneously. It identifies the subset of masks within $\mathcal{B}_k$ that optimally align with the primary query ϕ_{pri}, effectively resolving ambiguity through contextual reasoning.

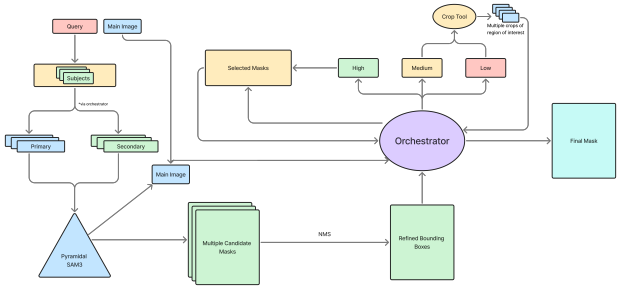


Fig. 3. SAM3 based Grounding pipeline with Qwen3-VL Orchestrator

E. FCC-to-RGB Reconstruction

We start from four-band NAIP imagery (R, G, B, NIR) and generate synthetic false-color composites (FCCs) by randomly sampling any 3-band subset that includes NIR, then permuting them.

A dual-path classifier predicts the spectral identity of each FCC channel, assigning each channel to one of {R, G, B, NIR}. Let the predicted assignments for the three channels be $\hat{y}_1, \hat{y}_2, \hat{y}_3$. The missing RGB band is then identified as:

$$B_{\text{miss}} = \{R, G, B\} \setminus \{\hat{y}_1, \hat{y}_2, \hat{y}_3\}.$$

If red or green is missing, we reconstruct it using a root-polynomial color correction (RPCC) mapping. For a pixel with input color vector $x = [x_1, x_2, x_3]^T$, the corrected channel output is:

$$C_{\text{out}} = \psi(x)^T v,$$

$$\psi(x) = [x_1,\ x_2,\ x_3,\ \sqrt{x_1 x_2},\ \sqrt{x_1 x_3},\ \sqrt{x_2 x_3},\ 1]^T.$$

Here, v denotes the learned coefficient vector (from calibration). RPCC is known to better preserve color under varying exposure compared to conventional polynomial regression. If the blue channel is missing instead, we use a pretrained 3-D lookup table (LUT) with trilinear interpolation to estimate the missing band:

$$B_{\text{out}} = \sum_{i=1}^{8} w_i\, T_i,$$

where T_i are the eight surrounding LUT vertices and w_i are the interpolation weights.

The final reconstructed RGB image is

$$I_{\text{RGB}} = \{R_{\text{out}},\ G_{\text{out}},\ B_{\text{out}}\},$$

which is then fed to downstream vision–language pipelines.

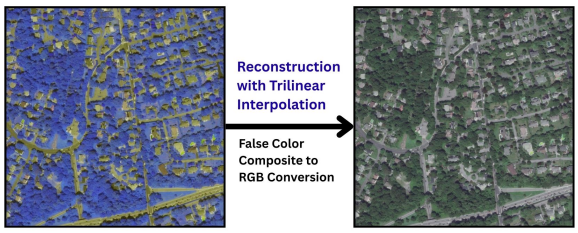


Fig. 4. FCC to RGB Conversion

V. Experiments

This section documents our systematic exploration, including failed approaches that motivated final design choices. Unless otherwise specified, **DRISHTI** denotes the complete system: Qwen3-VL-8B-Instruct with two-stage curriculum training, DPO alignment, question-type routing, and SAM3-based numeric module.

A. Preliminary Explorations and Design Rationale

Direct Box Regression. Initial attempts treated grounding as direct bounding box regression with L2 loss. While effective for single salient objects, multi-object queries collapsed into few coarse boxes with inconsistent coordinates.

Specialized Grounding Models. LiSAT [21], VLM-FO1-3B, Rex-Omni, and Rex-Thinker improved single-object localization but failed on: dense small objects (merged/missed), compositional queries (multi-hop reasoning), and domain shift (RS textures).

Segmentation-First Pipelines. Grounded SAM2 [22], RemoteSAM, and GeoSAM produced clean masks but required manual prompts per instance and yielded misaligned blobs in dense scenes.

Reasoning Model Distillation. We explored distilling reasoning traces from Gemini 2.5 Pro using a random sample of 9,500 VQA problems using LORA. This approach yielded *unexpected findings*: while the model appeared to internalize reasoning capabilities as intended, it exhibited a tendency to over-complicate simple tasks, resulting in performance degradation. While this method showed promise for complex, multi-hop queries, the performance degradation on simpler tasks—stemming from excessive reasoning overhead—led us to prioritize the proposed curriculum fine-tuning strategy.

These failures motivated DRISHTI's design:
(i) semantic learning via captioning/VQA rather than box regression,
(ii) Grounding via the Adaptive Hierarchical Grounding Network (AHG-Net), and
(iii) lightweight routing over heavy reasoning.

B. Backbone Selection

TABLE II
Performance Comparison of Models on the Generalized Dataset Captioning

Model	BERT-BLEU (Mean)	BLEU (Mean)	BERT-F1 (Mean)
Qwen2-VL-7B-Instruct	0.6818	0.0588	0.9434
Qwen3-VL-8B-Instruct	0.7066	0.0500	0.9405
MiMo-VL-7B-RL-2508	0.6640	0.0402	0.9362
Ovis2.5-9B	0.6742	0.0410	0.9309
InternVL3_5-8B	0.7876	0.0556	0.9405
DRISHTI Caption	**0.8216**	**0.0930**	**0.9466**

TABLE III
Performance Comparison of Models on the Specialized Dataset Captioning

Model	BERT-BLEU (Mean)	BLEU (Mean)	BERT-F1 (Mean)
Qwen3-VL-8B Instruct	0.7199	0.0521	0.9389
Qwen3-VL-32B Instruct	0.6017	0.0630	0.9040
MiMo-VL-7B-SFT-2508	0.6201	0.0287	0.9327
Qwen3-VL-8B Thinking	0.6390	0.0342	0.9322
InternVL 3.5-8B	0.7347	0.0871	0.9486
Ovis2.5-9B	0.5715	0.0327	0.8915
GPT-4o	0.8174	0.1002	0.9504
GPT-5-mini	0.8095	0.0412	0.9370
DRISHTI Caption	**0.8493**	**0.1342**	**0.9544**

Qwen3-VL-8B-Instruct is selected for:
(i) competitive performance despite smaller size than 32B alternatives,
(ii) Native dynamic resolution via M-RoPE,
(iii) LoRA compatibility, and
(iv) Tool-calling integration for SAM3
Notably, the 32B model underperforms its 8B counterpart, suggesting scale alone does not address RS challenges without domain adaptation.

C. VQA Results

Table **??** reports specialized VQA performance. DRISHTI achieves best semantic, binary metrics (+13.1%, +7.2% over Qwen3-VL-32B-Instruct).

TABLE IV
Performance Comparison of Models on the Generalized VQA

Model	Binary (Exact Match)	Numeric Score	Semantic (BERT-BLEU)
Qwen3-VL-8B-Instruct	0.6995	0.4586	0.6808
Ovis2.5-9B	0.7513	0.4904	0.5616
InternVL3.5-8B	0.7264	**0.5007**	0.7497
DRISHTI VQA	**0.7709**	0.4392	**0.8469**

Fig. 5. Question :"Time of the game played",Ans:"4 Quarters", Ground Truth:"48 mins"

On real-world RGB VQA (Table VII), DRISHTI surpasses all baselines on binary and semantic tasks. The reasoning-distilled variant (Qwen3-VL-8B-Thinking) exhibits substantially degraded performance, validating our design decision against overthinking.

D. Model Sensitivity at multiple resolutions

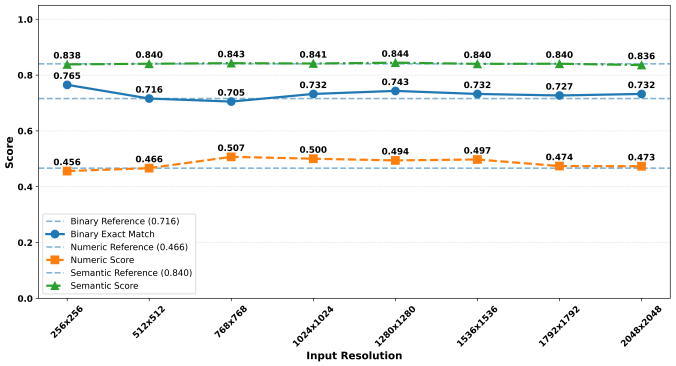


Fig. 7. Model Sensitivity Comparision on Different Resolutions

The plot demonstrates the model's robustness to input resolution changes, as performance across Binary, Numeric, and Semantic tasks remains stable with minimal variance relative to reference baselines (dashed lines)

VI. Experimental Results Tables

A. Captioning Results

1) FIT-RSFG and Real-World Captioning: Comparing DRISHTI to LLaVA-1.5-7B and GeoChat on FIT-RSFG captioning.

On real RGB scenes (Table V), DRISHTI clearly improves over its zero-shot version, indicating good transfer to the actual deployment domain.

TABLE V
PERFORMANCE COMPARISON OF MODELS ON THE REAL-WORLD RGB CAPTIONING

Model	BERT-BLEU	BLEU	BERT-F1
InternVL 3.5-8B	0.7482	0.0341	0.9368
MiMo-VL-7B-SFT-2508	0.6627	0.0106	0.9283
Ovis2.5-9B	0.6888	0.0152	0.9237
Qwen3-VL-8B-Instruct	0.6850	0.0170	0.9314
DRISHTI Caption	**0.7778**	**0.0397**	**0.9371**

B. VQA Results

1) FIT-RSFG VQA: Table VI shows results on FIT-RSFG VQA. DRISHTI already outperforms GeoChat substantially, even before adding routing and SAM3-based counting.

TABLE VI
VQA PERFORMANCE ON FIT-RSFG VQA (BINARY + NUMERIC). DRISHTI OUTPERFORMS GEOCHAT BY **+13.2%**.

Model	Overall Acc.	Count Acc.	Binary Acc.
GeoChat	0.5347	0.3564	0.7641
DRISHTI	**0.6667**	0.4115	0.7659

2) Real-World RGB VQA: On the real-world RGB set (Table VII), DRISHTI with full curriculum + routing + SAM3 pipeline improves binary and semantic performance over all baselines while achieving competitive numeric scores.

TABLE VII
PERFORMANCE COMPARISON OF MODELS ON THE REAL-WORLD RGB PROCESSED IMAGES VQA

Model	Binary (Exact Match)	Numeric Score	Semantic (BERT-BLEU)
InternVL 3.5-8B	0.8357	0.5032	0.8456
MiMo-VL-7B-SFT-2508	0.9019	0.6351	0.8480
Ovis2.5-9B	0.9477	**0.7163**	0.7851
Qwen3-VL-8B-Instruct	0.9479	0.6653	0.8325
DRISHTI VQA	**0.9480**	0.7020	**0.9370**

3) Specialized RS VQA: Table **??** reports results on the specialized high-difficulty VQA split. DRISHTI achieves the best semantic BERT-BLEU and best binary accuracy, supporting the choice of Qwen3-VL-8B-Instruct as the curriculum backbone.

C. Routing and SAM3-Based Counting

As discussed in Section IV-C, numeric questions are routed by Qwen3-VL-4B-Instruct to a SAM3 + pyramidal tiling module, rather than being answered purely by the VLM. Comparing the reasoning-traces Qwen3-VL-8B-Thinking row in Table VII to DRISHTI shows that curriculum + routing + SAM3 substantially improve numeric, binary, and semantic performance simultaneously-something no single off-the-shelf VLM achieves in this RS setting.

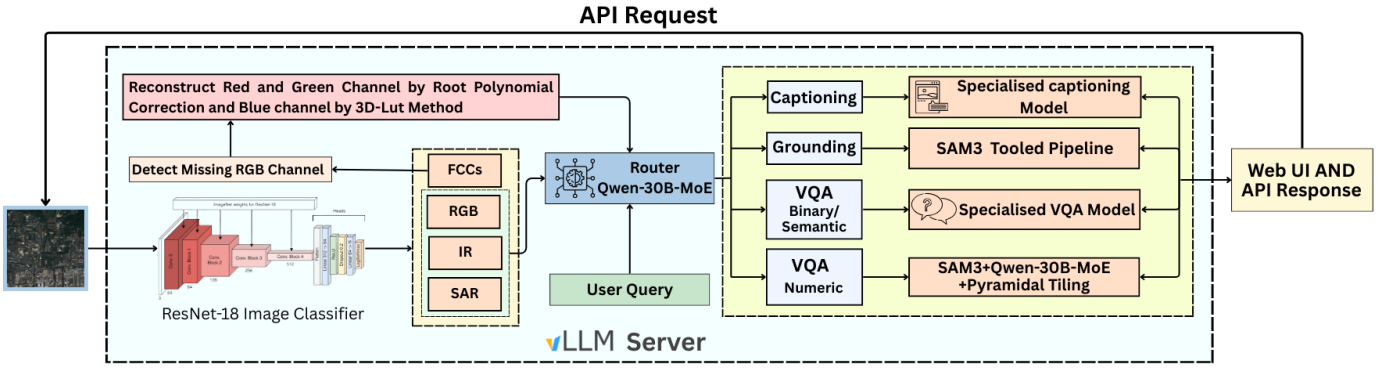


Fig. 6. DRISHTI System Architecture: A microservices-based orchestration layer dispatches multimodal queries to specialized inference engines via an intelligent routing fabric.

VII. System Design and User Interface

A. Architectural Overview

The DRISHTI system employs a modular inference architecture designed to handle the high variability of remote sensing queries. The core pipeline (Figure 6) orchestrates interactions between a lightweight routing layer, specialized vision-language models (VLMs), and geometric segmentation tools.

Routing Logic. To optimize computational resources, we implement a dual-routing strategy.

- **Visual Routing:** A lightweight ResNet classifier categorizes input imagery (e.g., dense urban, maritime, or SAR) to pre-filter eligible downstream tools.
- **Query Routing:** A frozen Qwen3-VL-30B-A3B serves as an "LLM-as-a-judge," analyzing user intent to dispatch requests to one of three specialized pipelines: *Captioning*, *Grounding*, or *VQA*. Within VQA, the router further discriminates between semantic queries (routed to the VLM) and numeric/spatial queries (routed to the segmentation engine).

Inference Engine. The system leverages **Qwen3-VL-8B** as the primary reasoning backbone, fine-tuned via the DRISHTI curriculum with LoRA adapters. To support high-precision object counting and area estimation, we integrate **SAM3** (Segment Anything Model 3). SAM3 operates in a pyramidal tiling mode, allowing it to detect small objects across large gigapixel satellite scenes. These geometric outputs (masks, oriented bounding boxes) are injected back into the VLM context for grounded reasoning.

B. Implementation and Interface

The backend is built on a scalable microservices pattern. All VLMs are served via a vLLM-based inference server, enabling optimized KV-cache reuse and batching across different tools. The frontend, implemented in React, provides a unified multimodal chat interface (Figure 8) where users can upload RGB, SAR, or IR imagery.

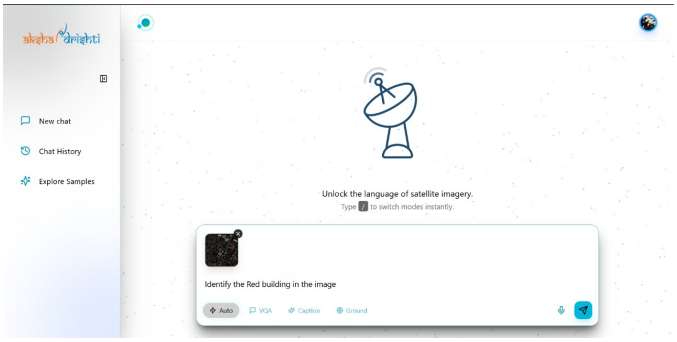


Fig. 8. The DRISHTI user interface featuring a unified chat panel for multimodal interaction and visualization of model outputs.

Crucially, the system maintains a "contextual memory" of intermediate artifacts. If a user asks to "count the ships" (triggering SAM3) and subsequently asks "are they moving?", the system retrieves the previously computed bounding boxes and masks from cache rather than re-running segmentation, ensuring low-latency conversational interactivity.

VIII. Conclusion and Future Work

This paper presented DRISHTI, a unified remote sensing vision–language framework combining a curriculum-trained Qwen3-VL backbone with SAM3-based segmentation. By decoupling numeric reasoning from semantic generation, the system achieves robust performance across diverse resolutions and modalities. Future work will focus on integrating **Complex-Valued CNNs** into the VLM encoder to natively process SAR phase and amplitude data, a novel architectural shift that unlocks physical scattering properties currently ignored by standard real-valued transformers. We also plan to extend the framework to multi-date temporal reasoning.

References

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

[2] D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, "Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model," *arXiv preprint*, 2024.

[3] Z. Yang *et al.*, "Skyeyegpt: Unifying remote sensing vision–language tasks via instruction tuning with large multimodal models," *arXiv preprint*, 2024.
[4] Z. Li *et al.*, "Vrsbench: A versatile vision–language benchmark dataset for remote sensing image understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
[5] X. Lu *et al.*, "Rsicd: A remote sensing image captioning dataset," *arXiv preprint*, 2017.
[6] S. Lobry *et al.*, "Rsvqa: Visual question answering for remote sensing data," *arXiv preprint*, 2020.
[7] J. Bai *et al.*, "Qwen3-vl technical report," *arXiv preprint*, 2025.
[8] Anonymous, "Direct preference optimization and bert-bleu alignment," *Technical report*, 2025.
[9] ——, "Detect anything via next point prediction: Rex-omni," *arXiv preprint*, 2025.
[10] ——, "Vlm-fo1: Vision–language models for open vocabulary detection," *Technical report*, 2025.
[11] ——, "Lisat: Language-guided spatial analysis tool," *arXiv preprint*, 2025.
[12] S. Liu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
[13] C. Liu, K. Chen, R. Zhao, Z. Zou, and Z. Shi, "Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model," *IEEE Geoscience and Remote Sensing Magazine*, 2025.
[14] K. Li, D. Wang, H. Xu, H. Zhong, and C. Wang, "Language-guided progressive attention for visual grounding in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.
[15] Y. Zhan, Z. Xiong, and Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.
[16] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.
[17] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.
[18] J. Henry, T. Natalie, and D. Madsen, "Pix2pix gan for image-to-image translation," *Research Gate Publication*, pp. 1–5, 2021.
[19] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.
[20] Y. Wei, A. Xiao, Y. Ren, Y. Zhu, H. Chen, J. Xia, and N. Yokoya, "Sarlang-1m: A benchmark for vision-language modeling in sar image understanding," *arXiv preprint arXiv:2504.03254*, 2025.
[21] J. Quenum, W.-H. Hsieh, T.-H. Wu, R. Gupta, T. Darrell, and D. M. Chan, "Lisat: Language-instructed segmentation assistant for satellite imagery," *arXiv preprint arXiv:2505.02829*, 2025.
[22] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

APPENDIX

*We have used GPT-5 for structuring queries to build grounding dataset and Grammarly AI for rephrasing the report

a) Rex-Omni.: Rex-Omni is a 3B-parameter multimodal model that frames visual perception as auto-regressive next-point prediction. It uses a Qwen-based vision–language backbone in which a ViT encoder and language decoder share a unified token space over both text and discretized spatial coordinates. A dedicated portion of the vocabulary is reserved for quantized coordinate tokens, allowing bounding boxes, points, polygons, and keypoints to be emitted as standard sequences. This yields a single generative interface that can natively support detection, grounding, OCR, keypoint estimation, and GUI/layout understanding. Training is performed on large-scale supervised data spanning diverse grounding and detection tasks, with additional geometry-aware objectives that promote precise, non-redundant spatial predictions. As a result, Rex-Omni exhibits strong zero-shot performance across common and dense detection benchmarks, as well as document and interface grounding scenarios.

b) VLM-FO1: VLM-FO1 is designed as a modular perception component for vision–language models that converts fine-grained localization into a retrieval-style operation over region tokens rather than direct coordinate generation, aligning spatial reasoning with the tokenized pipeline of large language models. A hybrid region encoder produces tokens that jointly encode semantics and geometry, allowing the language model to attend to specific regions as discrete units while retaining global scene understanding. Training focuses on strengthening region-level perception and its interaction with the base language model so that improved localization and counting do not compromise general multimodal reasoning, leading to consistent gains on grounding and region-centric benchmarks.

c) Qwen-3-8B-Instruct: Qwen 3-VL 8B is an 8B-parameter vision–language model that pairs a transformer-based visual encoder with a decoder-only language backbone, linked through a projection layer so that images are represented as visual tokens within the same sequence as text. Pretraining on large-scale image–text corpora and subsequent instruction tuning for multimodal dialogue and reasoning equip the model to handle tasks such as captioning, open-ended visual question answering, OCR-style reading, and general instruction following in multiple languages. The design emphasizes strong alignment between visual understanding and language generation while maintaining a parameter budget suitable for practical deployment, making Qwen 3-VL 8B a versatile base model for downstream multimodal applications.

d) Intern-VL-3.5-8B: InternVL-3.5-8B follows a ViT–MLP–LLM design, coupling an InternViT vision backbone with a lightweight projection module and an 8B-parameter language model to support high-resolution, fine-grained image understanding under a unified multimodal token interface. Its training pipeline combines multimodal pretraining with large-scale instruction tuning that includes reasoning-focused and tool-oriented data, followed by alignment methods that refine step-by-step visual–textual reasoning and response quality. The model incorporates efficient high-resolution handling and decoupled vision–language deployment, enabling long-context multimodal interaction and strong performance on diverse reasoning, agentic, and GUI-interaction tasks while remaining relatively compact.

e) Grounded SAM2: Grounded SAM2 combines an open-vocabulary grounding model with Segment Anything 2 (SAM2) to map text queries directly to object masks. A text prompt (e.g., "cargo ships", "red rooftops") is first processed by the grounding module, which predicts bounding boxes and confidence scores for all matching objects. These

TABLE VIII
LIST OF APIS, LICENSING, AND CONNECTIVITY REQUIREMENTS

Component	Endpoint URL	Base Model/Lib	License	Connectivity
VQA Specialized	maximuspookus–qwen3-vlm-vqa-serve	Qwen2.5-VL	Apache 2.0	Online (Hosted on Modal)
Captioning Specialized	maximuspookus–qwen3-vlm-captioning-serve	Qwen2.5-VL	Apache 2.0	Online (Hosted on Modal)
IR Model	maximuspookus–merge-qwen3-vl-ir-serve	Qwen2-VL (Fine-tuned)	Apache 2.0	Online (Hosted on Modal)
SAR Model	maximuspookus–merge-qwen3-vl-sar-serve	Qwen2-VL (Fine-tuned)	Apache 2.0	Online (Hosted on Modal)
ResNet Classifier	maximuspookus–resnet-modality-classifier-serve	ResNet (Torchvision)	BSD 3-Clause	Online (Hosted on Modal)
FCC RGB Synthesis	rockstar4119–fcc-rgb-synthesis-inference-predict	Custom / NumPy	MIT / BSD	Online (Hosted on Modal)
FCC Health Check	rockstar4119–fcc-rgb-synthesis-inference-health	Custom / FastAPI	MIT	Online (Hosted on Modal)
SAM3 Agent (Pyramidal)	rockstar4119–sam3-agent-pyramidal-v2-fastapi-app	SAM 2	Apache 2.0	Online (Hosted on Modal)
Qwen3VL 30B	rockstar4119–qwen3-vl-vllm-server-30b-vllm-server	Qwen2.5-VL-32B	Apache 2.0	Online (Hosted on Modal)

Note: All endpoints are hosted on Modal.com and require an active internet connection. The models themselves are open-weights and can be deployed offline if hardware permits.

TABLE IX
DRISHTI TRAINING HYPERPARAMETERS.

Parameter	Value
Base Model	Qwen3-VL-8B-Instruct
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Quantization	4-bit QLoRA
Effective Batch Size	256
Per-Device Batch	16
Gradient Accumulation	16
Learning Rate	5×10^{-5}
Optimizer	AdamW-8bit
Weight Decay	10^{-4}
Warmup Steps	100
Schedule	Cosine decay
Epochs (Stage I)	3
Epochs (Stage II)	2
DPO β	0.1
Random Seed	3407

TABLE X
DRISHTI BERT-F1 BY SCENE CATEGORY.

Category	DRISHTI	Qwen3-VL-8B
Urban	0.921	0.876
Industrial	0.918	0.862
Agricultural	0.935	0.901
Coastal/Harbor	0.912	0.847
Airport	0.927	0.889

TABLE XI
COUNTING ACCURACY BY OBJECT DENSITY.

Object Count	SAM3 + Pyramidal	VLM Direct
1-10	0.89	0.72
11-50	0.74	0.41
51-100	0.61	0.22
100+	0.48	0.11

TABLE XII
ABLATION STUDY ON DRISHTI-GCV SPECIALIZED VQA SPLIT.

Configuration	Binary	Numeric	Semantic
Base Qwen3-VL-8B	0.684	0.381	0.703
+ Stage I SFT	0.721	0.392	0.756
+ Stage II SFT	0.754	0.401	0.812
+ Question Router	0.768	0.518	0.835
+ SAM3 Numeric	0.768	0.642	0.835
+ DPO Alignment	**0.779**	**0.642**	**0.860**

boxes are then used as prompts for SAM2, which refines them into high-quality instance masks, leveraging SAM2's strong generalization and, for video, its temporal memory. This yields a flexible open-vocabulary segmentation pipeline where free-form text input produces pixel-accurate masks suitable for counting, region-level reasoning, and downstream VLM conditioning.

f) Pyramidal tiling: To make SAM-style segmentation tractable on high-resolution RS scenes with tiny objects, we adopt a pyramidal tiling strategy. For each image I, we extract overlapping crops at multiple tile sizes $T = \{t_1, \ldots, t_L\}$ *without* rescaling the pixels, run SAM (or SAM2/3) independently on each crop, and map all masks back to global coordinates. Smaller tiles act as zoomed-in windows that recover small or low-contrast instances, while larger tiles preserve scene context for large structures. Finally, we apply IoU-based non-maximum suppression across all scales to merge duplicates and retain a compact set of high-quality masks.

g) Supervised Fine-Tuning (SFT): In the SFT stage, the vision-language model is trained on instruction-formatted image-text pairs. Each example is composed of an image I, a textual prompt x (instruction and/or question), and a target sequence $y = (y_1, \ldots, y_T)$ (caption or answer). An autoregressive conditional distribution is defined as

$$p_\theta(y \mid I, x) = \prod_{t=1}^{T} p_\theta\big(y_t \mid I, x, y_{<t}\big),$$

and the token-level negative log-likelihood over the training set $\mathcal{D}$ is minimized:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(I,x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T} \log p_\theta\big(y_t \mid I, x, y_{<t}\big)\right].$$

TABLE XIII
DRISHTI-GCV SOURCE DATASET BREAKDOWN.

Source	Samples	Tasks
VRSBench	29,614	Caption, VQA, Ground
Git-10M (filtered)	10,000	Caption
RSVQA-LR	8,500	VQA
RSVQA-HR	7,200	VQA
OPT-RSVG	12,000	Grounding
RSVG/RSVG-HR	4,800	Grounding
DIOR/DOTA-v2	6,200	Grounding

TABLE XIV
OBJECT CATEGORY DISTRIBUTION IN DRISHTI-GCV.

Category	Percentage
Vehicles (cars, trucks, planes)	28.3%
Buildings (residential, commercial)	24.1%
Infrastructure (roads, bridges)	18.7%
Water bodies & ships	12.4%
Agricultural land	9.2%
Other (parks, forests, etc.)	7.3%

Through this objective, the model is aligned to follow multi-modal instructions and to produce coherent, task-appropriate outputs conditioned on both the image and the text prompt.

SYSTEM PROMPT (BRIEF VERSION)

You are a remote-sensing segmentation and mask-verification agent operating under a strict one-shot workflow. You get exactly one call to `segment_phrase_batch` and must include 3–5 true synonyms of the target feature (add color only if explicitly stated). SAM3 returns unique masks via NMS; if zero masks are produced, immediately call `report_no_mask()`.

After segmentation, classify each mask by quick-scan confidence: HIGH (clear match), MEDIUM (likely but uncertain), LOW (ambiguous or possibly wrong). HIGH confidence masks may be accepted immediately. MEDIUM or LOW confidence masks must be examined with `examine_each_mask()` before selection. Determine whether masks represent the primary target, not reference or intermediate objects; if all masks are the wrong feature type, you must output `report_no_mask()` because the one-shot chance is already spent.

Perform all spatial reasoning using mask bounding-box centroids, interpreting terms like "leftmost," "in the left," "below," "near," or compounds such as "below and rightmost of X." Apply directional filtering first when needed, then superlative selection. Always ensure that selected masks satisfy both feature type and spatial constraints.

For dark masks, apply mandatory shadow vs. real-object discrimination. Shadows show elongation, uniform darkness, adjacency to tall objects, and consistent sun-angle direction; real objects have structure, edges, and shape consistent with the target. After confidence, target, spatial, and shadow checks, call either `select_masks_and_return`, `examine_each_mask`, or `report_no_mask` with no extra text after the tool call. ‘ [1]

% Decrements the counter if you removed it, to avoid skipping a number (optional)

[1]**Query Refactoring and Grounding:** Assisted by **ChatGPT GPT-5 Mini** (OpenAI). The model was utilized for its ability to refactor initial queries into precise, grounded search terms for enhanced information retrieval. Usage is subject to OpenAI's applicable Terms of Use and pricing policies.

Rephrasing and Clarity: Assisted by **Grammarly Students (Premium)** for advanced rephrasing and improved clarity of the final text content. Note: This tool was used on the text content itself, not for native command-level editing within the LaTeX source files.

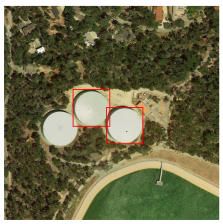
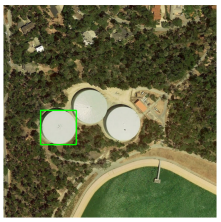
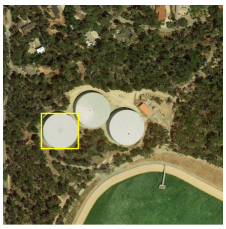
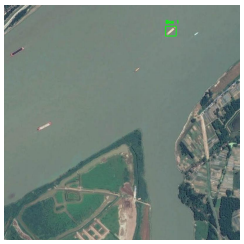
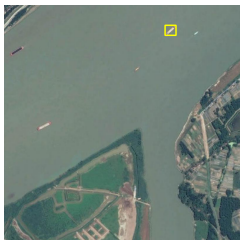
Queries	REX-Omni	LISAT	Ours	Ground Truth
Leftmost Storage Tank				
Largest ship in the right half				

TABLE XV

VISUAL COMPARISON OF REX-OMNI, LISAT, AND THE PROPOSED METHOD AGAINST THE GROUND TRUTH FOR THREE DIFFERENT VISUAL GROUNDING QUERIES

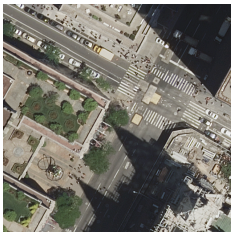
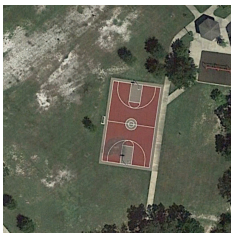
Query	Image	GPT-4o-mini	Qwen3-VL-32B-Instruct	Ours	Ground Truth
In which direction is the sun facing?		south west	top right	north west	north west
How long does the game go on for in the image?		60 minutes	30 minutes	48 minutes	48 minutes

TABLE XVI
COMPARISON OF RESULTS OF GPT-40-MINI, QWEN3-VL-32B-INSTRUCT, AND THE PROPOSED METHOD AGAINST THE GROUND TRUTH FOR THREE DIFFERENT VISUAL QUESTION ANSWERING QUERIES