SARChat-2M: A Multi-Task Vision-Language Dataset and Benchmark for SAR Image Interpretation

Anonymous ACL submission

Abstract

Synthetic Aperture Radar (SAR) image interpretation remains challenging for general vision-language models because SAR appearance differs substantially from natural and optical remote-sensing imagery. We introduce SARChat-2M, a large-scale SAR instruction-following dataset automatically derived from unified SAR detection annotations, and SARChat-Bench, a multi-task benchmark built on its held-out test split for evaluating SAR-oriented vision-language understanding. SARChat-2M converts detection annotations into six task families, including classification, fine-grained description, instance counting, spatial grounding, region-conditioned identification, and category-guided referring. Instead of reporting a single overall score, SARChat-Bench emphasizes task-specific metrics and deterministic parsing rules. Experiments with 16 fine-tuned open-source VLMs, together with a before-and-after study on InternVL2.5-8B, show that SAR-domain instruction tuning is important for in-domain recognition and region-conditioned understanding, while counting with localization, multi-target grounding, and referring remain difficult. We also discuss limitations arising from template-generated language, class/task imbalance, and inherited annotation incompleteness. We will release the dataset, evaluation scripts, and trained weights upon publication.

1 Introduction

Synthetic Aperture Radar (SAR) is an active microwave imaging modality that can observe the Earth in cloudy, dark, and adverse-weather conditions. Unlike optical imagery, a SAR image is formed by coherent radar backscatter. Targets often appear as bright scattering centers with speckle noise, layover, shadow, blurred boundaries, and strong orientation dependence. These properties make SAR valuable for all-weather Earth observation, maritime and traffic monitoring, infrastructure inspection, environmental monitoring, and disaster response, but they also make SAR interpretation difficult for models trained mainly on natural or optical remote-sensing images.

In recent years, deep neural networks, notably CNNs (LeCun et al., 1998) and ViTs (Dosovitskiy, 2020), have achieved remarkable progress in remote sensing data analysis, enhancing both processing efficiency and analytical accuracy. However, existing research mainly focuses on visual feature extraction, while lacking deep semantic parsing and reasoning capabilities (Li et al., 2024a), limiting model applicability in complex scenarios.

With the advancement of Large-Language Models (LLMs), Vision-Language Models (VLMs), through integrating pre-training and instruction tuning, have demonstrated robust zero-shot learning and generalization in multimodal tasks (Dai et al., 2023). This has inspired researchers to explore the deep integration of visual models with LLMs.

Although models designed for optical remote sensing images, like RSGPT (Hu et al., 2023) and GeoChat (Kuckreja et al., 2024), have shown preliminary achievements, they struggle to perform well in SAR applications. SAR images inherently pose significant interpretation challenges due to their scattering imaging mechanisms, characterized by blurred target edges, dispersed speckles, and orientation sensitivity. Meanwhile, existing SAR datasets primarily focus on visual recognition tasks (Kuckreja et al., 2024; Cheng et al., 2022; Zhang et al., 2023), leaving a critical shortage of large-scale, high-quality image-text alignment datasets. Both these intrinsic characteristics and data limitations impede the advancement of VLMs in the SAR domain.

Current VLMs are primarily trained on conventional natural images without extensive fine-tuning for the SAR vertical domain. Despite their strong visual capabilities for natural images, these models still have significant room for improvement in SAR

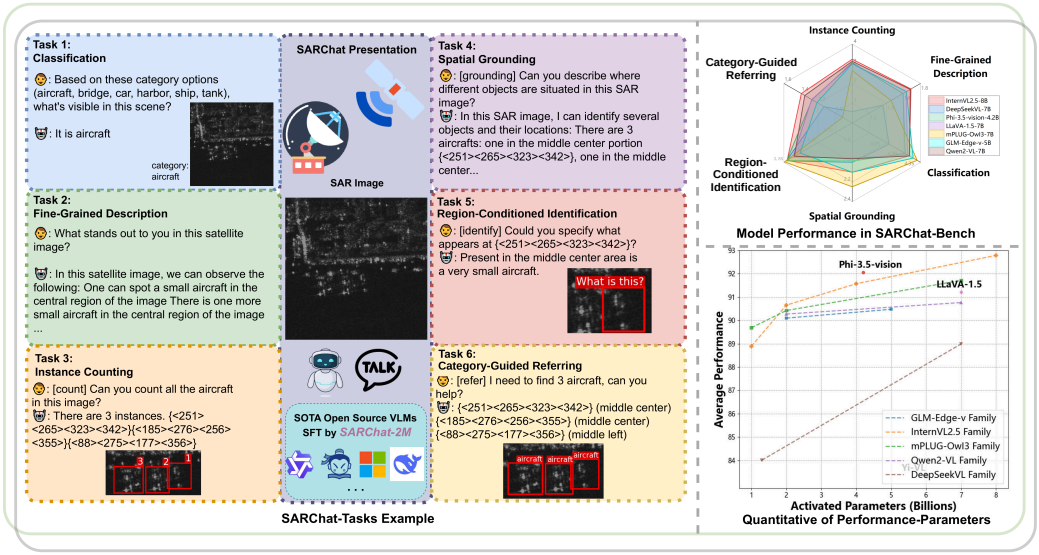

Figure 1: **An overview of SARChat-2M.** The left figure demonstrates the representative tasks realized with the SAR image-text dataset, SARChat-2M, constructed in this paper, and validates the dataset's effectiveness in supporting multi-task applications. The right figure presents the correlation radar charts and quantitative line graphs derived from the performance evaluation of 16 VLMs based on this dataset, establishing the SARChat-Bench evaluation suite within this domain.

image interpretation. Building upon the SARDet-100K dataset (Li et al., 2024b) with its rich SAR imagery and detection annotations, we construct SARChat-2M, a task-oriented SAR-specific image-text pair dataset, to address the insufficient SAR image interpretation capabilities of existing VLMs.

As shown in Figure 1, we present SARChat-2M, a large-scale multimodal conversational dataset for SAR images, and use its held-out test split to construct SARChat-Bench, a multi-task evaluation suite. The SARChat-2M dataset contains approximately 2 million SAR instruction samples across maritime, land-cover, and urban infrastructure scenarios, featuring standardized target categories, bounding boxes, grid-based spatial descriptions, and task identifiers. The dataset supports major vision-language tasks such as image captioning, VQA, visual localization, and object detection. To systematically evaluate model performance in these domains, we design six specific benchmark tasks: **classification**, **fine-grained description**, **instance counting**, **spatial grounding**, **region-conditioned identification**, and **category-guided referring**. To validate the usefulness of the dataset and benchmark, we evaluate 16 state-of-the-art VLMs of varying parameter scales, including InternVL2.5, DeepSeekVL, GLM-Edge-V, and the mPLUG-Owl family. We report in-domain fine-tuned results separately from non-fine-tuned settings to make the effect of SARChat-2M explicit.

Our contributions are threefold.

1. We construct SARChat-2M, a large-scale SAR instruction-following dataset automatically derived from SARDet-100K detection annotations, with standardized category names, coordinate serialization, grid-based spatial descriptions, and task-specific prompts.

2. We introduce SARChat-Bench, a deterministic multi-task evaluation suite built on SARChat-2M's held-out split, for SAR VLMs, covering classification, fine-grained description, counting, spatial grounding, region-conditioned identification, and category-guided localization.

3. We benchmark 16 open-source VLMs under SAR-domain LoRA fine-tuning and provide a task-level analysis showing that coarse recognition and region-conditioned identification are relatively easy, while multi-target localization, category-guided referring, and fine-grained description remain challenging.

2 Related Work

2.1 VLMs for Remote Sensing

VLMs are capable of converting images into natural language descriptions and parsing the relationships between objects, demonstrating remarkable

performance in tasks such as text-image retrieval, image captioning, and visual question answering. Recently, models like RemoteClip (Liu et al., 2024) have been applied to the field of remote sensing images, primarily focusing on cross-modal retrieval and zero-shot classification. However, these models have not addressed tasks such as image description generation and visual grounding. The RSGPT model has achieved text description and visual question answering for remote sensing images, but it has not expanded to tasks such as classification and detection. The GeoChat model has advanced multi-task conversational processing of high-resolution remote sensing imagery, including scene classification, visual question answering, multi-turn dialogue, visual grounding, and reference object detection. However, these models, including GeoChat, predominantly rely on optical remote sensing training data, leading to suboptimal performance in SAR-specific interpretation tasks. EarthGPT (Zhang et al., 2024) has extended the application of multimodal large language models to the remote sensing field through instruction tuning, but its performance in SAR multi-task processing still needs improvement. Compared with natural images, the interpretation of SAR images is more challenging, which poses higher demands on the model's processing capabilities and adaptability.

2.2 Remote Sensing Vision-Language Datasets

Remote sensing datasets are essential for models that interpret remote sensing imagery. Existing datasets such as UCM Captions (Qu et al., 2016a), Sydney Captions (Qu et al., 2016b), RSICD (Lu et al., 2017), RSITMD (Yuan et al., 2022), and RSVG (Zhan et al., 2023) provide preliminary resources for studying the correlation between remote sensing images and text. However, these datasets are limited not only in scale but also in modality, containing only optical images without SAR data, leaving SAR interpretation capabilities largely unexplored. Although large-scale datasets like MillionAID (Long et al., 2021), FMoW (Christie et al., 2018), and BigEarthNet (Sumbul et al., 2019) exist, they lack text-image pairs. The RS5M dataset (Zhang et al., 2023), containing 5 million image-text pairs, is still limited to optical images. The MMRS-1M dataset (Zhang et al., 2024), which covers optical, infrared, and SAR modes, has a very low proportion of SAR image-text data. Therefore, this paper constructs the SARChat-2M dataset, which focuses on SAR images and contains over 2 million instruction samples, covering tasks such as classification, detection, caption generation, VQA, and grounding.

3 Data Construction and Description

3.1 The Procedure of Data Construction

3.1.1 Dataset Overall

As shown in Figure 2, we propose SARChat-2M, a multi-task dataset for SAR images, comprising 2 million multimodal dialogue samples (1,836,912 train and 226,636 test samples) to support model training and evaluation. We strictly follow the official SARDet-100K image-level train/val/test split; all instruction samples are generated separately within each official split rather than randomly split after generation, ensuring no image- or bounding-box-level leakage between training and test sets.

Based on the SARDet-100K dataset (Li et al., 2024b), it incorporates multimodal adaptations and enhanced language annotations from ten established SAR detection benchmarks such as AIR-SARShip(1.0&2.0) (Wang et al., 2019), HRSID (Wei et al., 2020), MSAR (Chen et al., 2022), SADD (Zhang et al., 2021a), SAR-AIRcraft (Zhirui et al., 2023), ShipDataset (Wang et al., 2019), SSDD (Zhang et al., 2021b), OGSOD (Wang et al., 2023), and SIVED (Lin et al., 2023). The SARChat-2M covers six semantic categories (Ship, Aircraft, Car, Tank, Bridge, and Harbor) and supports six core SAR image analysis tasks: **classification**, **fine-grained description**, **instance counting**, **spatial grounding**, **region-conditioned identification**, and **category-guided referring**. These diverse tasks are designed to enhance VLMs' capabilities in SAR image interpretation, with approximately 2 million automatically generated instruction samples derived from the detection annotations.

Difference from SARDet-100K. SARDet-100K provides unified SAR images and detection boxes, while SARChat-2M converts these detection annotations into language-grounded instruction data and benchmark tasks. The added components include: (i) canonical category names and synonym normalization for language evaluation; (ii) spatial verbalization based on bounding boxes and a 3×3 grid; (iii) task identifiers and prompt templates for six task families; (iv) multi-instance serialization rules for object lists; and (v) deterministic parsing and task-specific scoring protocols. Therefore, SARChat-2M is not a new SAR image collection

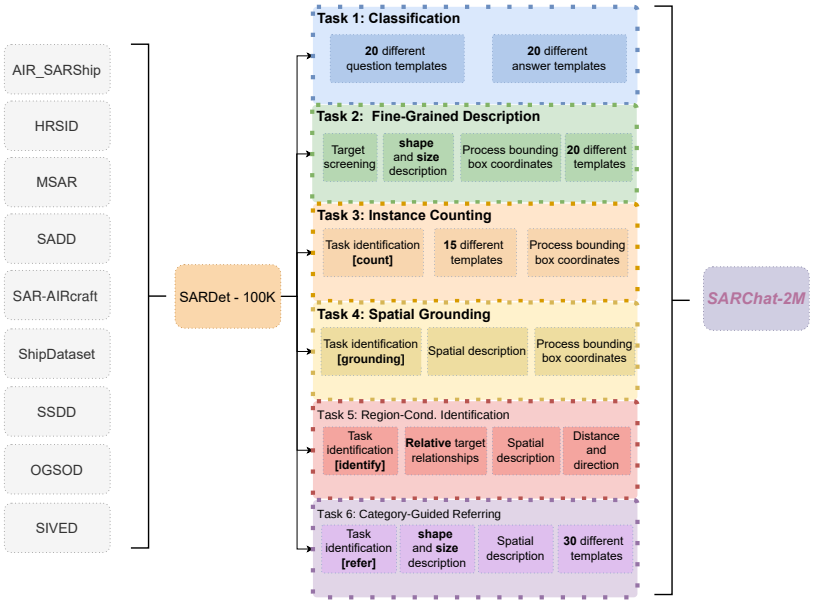

Figure 2: **Construction of SARChat-2M dataset.** On the left, ten existing SAR detection datasets. The middle part is the SARDet-100K dataset, formed by integrating the ten datasets on the left. On the right, six core tasks constructed based on the dataset are presented, with each task corresponding to different task identifiers, operation steps, and relevant templates.

independent of SARDet-100K; its contribution is a standardized SAR vision-language layer and evaluation suite built on top of existing SAR resources.

Quality control. We applied automatic validation to remove invalid boxes, out-of-bound coordinates, extreme aspect ratios, and inconsistent category names. Template outputs are validated by schema constraints so that every generated answer contains the required fields for the corresponding task. We also performed sampled manual inspection to verify category consistency and spatial verbalization. Since the inspection was not exhaustive, residual missing objects and imprecise boxes may remain; we discuss this limitation in Section 6.

3.1.2 Task Definition

Based on SAR image characteristics and VLM capabilities, SARChat-Bench contains six tasks: **classification** evaluates category recognition; **fine-grained description** requires category, morphology, size, and coarse location descriptions; **instance counting** tests whether a model can count all targets of a queried category and return their coordinates; **spatial grounding** evaluates multi-object spatial relation understanding and box prediction; **region-conditioned identification** asks the model to identify the object inside a provided bounding box, so it measures recognition rather than localization; and **category-guided referring** asks the model to localize one or more instances of a queried category. The current referring task uses category-level prompts rather than complex attribute- or relation-based referring expressions.

3.1.3 Task-Oriented Data Generation

Based on the six tasks, we generate multimodal instruction-response pairs from SARDet-100K annotations. Detailed templates are provided in Appendix D.4.

(0) Dataset Definitions

All tasks use the same coordinate format, $\{<x_1><y_1><x_2><y_2>\}$, where (x_1,y_1) and (x_2,y_2) are the top-left and bottom-right corners. Coarse spatial descriptions are derived from a 3×3 grid. Task identifiers specify the expected output type: $[count]$ for counting, $[grounding]$ for spatial grounding, $[identify]$ for region-conditioned identification, and $[refer]$ for category-guided referring.

(1) Classification Task

We use 20 question-answer templates for category recognition, with standardized category names for single- and multi-target scenes.

(2) Fine-Grained Description Task

For fine-grained description, we filter images below 224×224 pixels, remove targets with area ratio $R_{\text{pct}} < 1\%$, aspect ratio above 10:1, or out-of-bound coordinates, and verbalize category, size, morphology, and grid location. We use 40 templates for single- and multi-target scenes. The area

ratio is:

$$R_{\text{pct}} = \frac{w_{box} \times h_{box}}{W_{img} \times H_{img}} \times 100\% \quad (1)$$

where w_{box} and h_{box} are box width and height, and W_{img} and H_{img} are image width and height.

(3) Instance Counting Task

We create 15 $[count]$ templates. Responses include the object count and serialized boxes for all matched instances.

(4) Spatial Grounding Task

We generate 15 $[grounding]$ templates that describe relative distance and direction among targets. Proximity is defined by a threshold of $(W_{img} + H_{img})/8$, and directional relations cover horizontal, vertical, and diagonal layouts.

(5) Region-Conditioned Identification

Given a ground-truth box in the prompt, the $[identify]$ task asks the model to identify the enclosed object and describe size, morphology, and grid location. Size uses five area-ratio bins ($\geq$0.4: very large; $\geq$0.25: large; $\geq$0.1: medium; $\geq$0.03: small; $<$0.03: very small), and morphology uses aspect-ratio bins ($>$1.5: tall-body; 0.67–1.5: approximately square; $<$0.67: wide-body). We use 20 response templates.

(6) Category-Guided Referring Task

The $[refer]$ task uses category prompts such as "Where is the {category}?" and outputs all matching boxes with grid-based orientation descriptions.

3.2 Quantitative Analysis of Datasets

The quantitative analysis in this study focuses on two key dimensions: category distribution and object morphological patterns.

Category	Training	Test
Ship	93,373 (46.98%)	10,741 (44.38%)
Aircraft	40,705 (20.48%)	6,779 (28.01%)
Car	9,561 (4.81%)	1,230 (5.08%)
Tank	24,187 (12.17%)	1,773 (7.33%)
Bridge	27,615 (13.89%)	3,281 (13.56%)
Harbor	3,306 (1.66%)	399 (1.65%)

Table 1: Category Distribution Statistics

(1) Category Distribution Characteristics

As shown in Table 1, the Ship category dominates both training and test sets (46.98% and 44.38% respectively), while the Harbor category represents less than 2%. A significant distribution shift is observed in the Aircraft category, with a 7.53% increase in the test set compared to the training set. Categories such as Car, Tank, and Bridge maintain moderate and stable proportions across both sets. This class distribution aligns with real-world SAR imagery characteristics, where certain target types naturally appear more frequently than others due to the inherent nature of SAR remote sensing applications and operational scenarios.

(2) Object Morphology Analysis

This study quantifies geometric characteristics using aspect ratio (AR):

$$AR = \frac{h_{box}}{w_{box}} \quad (2)$$

Metric	Training Set	Test Set	Diff-Rate
Mean	1.28	1.26	-0.02
Median	1.062	1.05	-0.017
SD	1.18	0.91	-0.22

Table 2: Aspect Ratio Distribution Comparison

As shown in Table 2, the differences in central tendency between training and test sets are minimal (mean: -0.02, median: -0.017). The test set exhibits a 0.22 lower standard deviation, indicating a more concentrated distribution. The key morphological distribution intervals of targets are illustrated in Appendix D.2.

The dataset exhibits three distinct morphological categories based on aspect ratio (AR): wide-body (AR $\leq$ 0.67), nearly square-shaped (0.67 $<$ AR $\leq$ 1.5), and tall-body (AR $>$ 1.5). Detailed distribution analysis can be found in Appendix D.2.

4 SARChat-Bench Evaluation Method and Settings

This section details the evaluation methodology of SARChat-Bench, a standardized benchmark suite built on SARChat-2M's test split for multi-task assessment of VLMs in SAR interpretation.

4.1 Evaluation Metrics

(1) Classification Accuracy and Exact Match: For classification, we use standard accuracy:

$$\text{Acc}_{cls} = \frac{N_{correct}}{N_{total}} \times 100\%. \quad (3)$$

For instance counting, we report exact-match accuracy and mean absolute error (MAE):

$$\text{Acc}_{count} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{c}_i = c_i], \quad (4)$$

$$\text{MAE}_{count} = \frac{1}{N} \sum_{i=1}^{N} |\hat{c}_i - c_i|. \quad (5)$$

Model	Param	Count EM	Class Acc	Fine-Grained	Reg.-Cond. ID (Multi)	Reg.-Cond. ID (Single)
InternVL2.5	8B	**74.14**	97.25	**63.43**	98.84	98.98
Qwen2-VL	7B	72.79	97.30	63.11	97.54	99.18
GLM-Edge-V	5B	73.44	98.02	61.45	96.69	95.96
Phi-3.5-vision	4.2B	72.69	96.42	59.95	98.93	98.59
LLaVA-1.5	7B	71.89	96.90	61.35	97.84	98.42
mPLUG-Owl3	7B	71.00	**98.80**	54.65	**99.27**	**99.51**
DeepSeekVL	7B	20.66	93.23	51.08	98.97	99.05
Yi-VL	6B	32.62	95.32	38.15	93.63	97.95

Table 3: Recognition and description results on SARChat-Bench (best variant per model family). Reg.-Cond. ID reports normalized category identification accuracy under a provided region prompt; it does not evaluate localization. Full results for all 16 models are in Appendix B.

Model	Param	Count IoU@0.5	Ground (Multi)	Ground (Single)	Ref (Multi) IoU@0.5	Ref (Single) IoU@0.5
Qwen2-VL	7B	**50.24**	64.17	83.87	**26.29**	**57.04**
InternVL2.5	8B	52.17	62.25	87.91	23.46	60.13
Phi-3.5-vision	4.2B	47.60	58.85	87.29	17.16	55.70
LLaVA-1.5	7B	46.80	62.70	85.79	15.48	56.70
GLM-Edge-V	5B	44.56	61.38	**89.96**	15.41	51.81
mPLUG-Owl3	7B	35.27	56.37	93.32	7.66	38.00
DeepSeekVL	7B	4.19	**65.32**	85.78	13.66	48.84
Yi-VL	6B	9.44	53.68	72.38	2.69	16.63

Table 4: Localization and category-guided referring results on SARChat-Bench (best variant per model family). Ground = Spatial Grounding; Ref = Category-Guided Referring.

(2) Set Matching F1: For tasks that output a set of objects or attributes, we compute precision, recall, and F1 after deterministic parsing and label normalization:

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN}. \tag{6}$$

We report this metric as set-level F1 rather than treating it as accuracy.

(3) Intersection over Union (IoU): In tasks involving localization, spatial grounding, and category-guided referring, IoU measures the overlap between predicted and ground-truth bounding boxes. We report IoU@0.5 as the primary localization metric and IoU@0.25 as a diagnostic metric for small or elongated SAR targets whose boundaries are noisy or weakly annotated. IoU is not used for Region-Conditioned Identification because the target region is already provided in the prompt.

(4) Overall Score Calculation (supplementary):

$$S_m = \sum_{t \in T} \frac{n_t}{\sum_{j \in T} n_j} a_{m,t}. \tag{7}$$

Here, n_t denotes the number of samples in task t, $a_{m,t}$ denotes the normalized score of model m on task t, and T is the set of all tasks. **Important caveat:** because Region-Conditioned Identification accounts for approximately 77% of the total samples, the weighted average score S_m is dominated by this single task. We report the weighted overall score only as a compact summary index; it should not be interpreted as the primary leaderboard metric. We strongly recommend using per-task metrics (reported in Table 3) as the primary evidence for model comparison. A macro-average over tasks would give equal weight to each task regardless of sample size, but such a metric would also be misleading because different tasks use different measurement scales (accuracy, IoU, F1, set matching). The detailed calculation of $a_{m,t}$ for each task can be found in Appendix F.

4.2 Assessment Methods

This section elaborates on the specific evaluation method processes for six types of tasks.

Parsing and normalization. We parse free-form model outputs with task-specific regular expressions. Category names are normalized with a fixed synonym table (e.g., *car/automobile*, *port/harbor*), coordinates are extracted from the canonical {<x_1><y_1><x_2><y_2>} format, and grid labels are mapped by the center point of the predicted box with deterministic boundary tie-breaking. Invalid or unparsable outputs are counted as incorrect for the corresponding field.

(1) **Instance Counting:** Compare predicted and label object counts for single-class evaluation. We report count exact match and MAE; when coordinates are also required, localization quality is reported separately with IoU to avoid conflating count and box errors.

(2) Spatial Grounding: Evaluate spatial accuracy through IoU-based bbox matching and abstract po-

sition analysis (e.g., "top", "bottom") from natural language descriptions.

(3) Region-Conditioned Identification: Since the target region is provided in the prompt, this task is evaluated as region-conditioned recognition rather than localization. We parse the model response, normalize category names with the synonym table, and compare the predicted category with the ground-truth category. When attributes are available, we evaluate them through set-level matching. IoU is not used as the primary metric for this task because the target box is already given in the input prompt. Invalid or unparsable responses are counted as incorrect.

(4) Category-Guided Referring: Assess category-guided localization accuracy through IoU metrics in both single-target and multi-target contexts. Given a category prompt, the model must output bounding boxes for all instances of that category.

(5) Fine-Grained Description: Segment predictions and ground-truth into short phrases, extract category and position information, and compare content sets for detailed description evaluation.

(6) Classification: Compare predicted and ground-truth categories to assess accuracy Acc.

5 Experiments and Analysis

Evaluation regimes. Unless otherwise stated, Table 3 reports in-domain LoRA fine-tuned open-source VLMs. Zero-shot or API-only models are evaluated with the same prompts and decoding settings but are reported separately to avoid mixing training regimes. For decoding, we use deterministic generation with temperature 0 and a fixed maximum output length for each task.

5.1 Benchmark Evaluation

To verify the effectiveness and practicality of the SARChat-2M dataset, we conducted extensive experiments on SAR image interpretation tasks using sixteen mainstream large vision-language models. Our preliminary analysis reveals that these models, despite their strong performance on natural images, struggle significantly with SAR image interpretation without domain-specific fine-tuning, demonstrating the critical importance of SAR-domain adaptation. Specifically, we conducted a before-and-after analysis on InternVL2-8B to quantify the impact of SAR fine-tuning. The results are presented in Appendix H.

As shown in Tables 3 and 4, the evaluated models include recent advances such as Qwen2-VL(Wang et al., 2024), InternVL2.5(Chen et al., 2024), DeepSeekVL(Lu et al., 2024), Phi-3.5-vision(Abdin et al., 2024), GLM-Edge-V(GLM et al., 2024), mPLUG-Owl3(Ye et al., 2024), Yi-VL(Young et al., 2024) and LLaVA-1.5(Liu et al., 2023). We additionally evaluate three newer model families—InternVL3.5 (1B–8B), Qwen3 (2B–8B), and the SAR-specific RSCOVLM-7B—under the same fine-tuning protocol; the extended results are reported in Appendix C. Rather than ranking models by overall score, we organize the analysis around four key observations:

Observation 1: Classification is nearly saturated. Most fine-tuned VLMs exceed 96% classification accuracy, with mPLUG-Owl3-7B reaching 98.80%. This indicates that coarse SAR target recognition under the in-domain split is relatively easy for modern VLMs after SAR-domain tuning. Classification alone is not sufficient to characterize SAR interpretation ability; the more diagnostic tasks are grounding, category-guided referring, and fine-grained description.

Observation 2: Region-conditioned identification is recognition, not localization. Given a bounding box as input, most models achieve 93–99% identification accuracy. This high score should be interpreted as evidence of region-conditioned recognition: the model only needs to identify and describe the object inside a given region, which is substantially easier than localizing targets from scratch. The region prompt drastically reduces the search space, so this task does not measure scene understanding or localization ability.

Observation 3: Category-guided referring and multi-target localization expose the real difficulty. Category-guided referring remains the hardest localization task, especially in multi-target scenes where IoU@0.5 drops below 27% for all models. The large asymmetry between region-conditioned identification (>93%) and multi-target referring (<27% at IoU@0.5) indicates that recognizing an object inside a given box is much easier than searching for the correct box from a category prompt. Multi-target grounding also shows a substantial drop compared to single-target grounding, suggesting that current VLMs struggle with target density and maintaining object identities across multiple boxes.

Observation 4: Fine-grained description is limited by both task difficulty and data scarcity. Fine-grained description scores are consistently

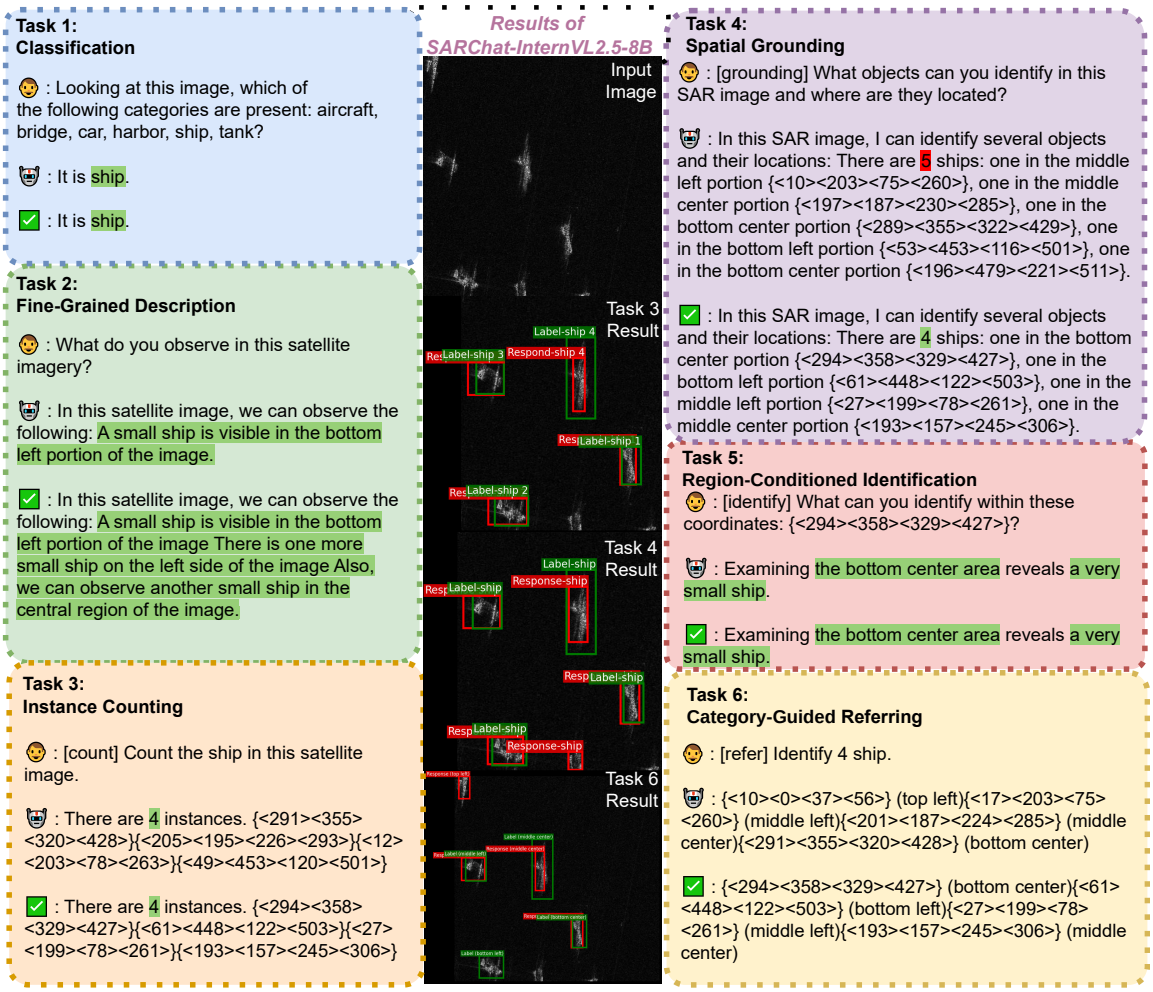

Figure 3: **Evaluation examples on SARChat-Bench.** VLM predictions are shown in green/red for correct/incorrect descriptions, with the ground truth in green and the predictions in red boxes. And [Human], [Bot], and [Check] icons denote user input, VLMs' response, and standard output, respectively.

lower than classification and region-conditioned identification across all models. This is partly because the task requires multiple attributes (category, size, morphology, location) to be correct simultaneously, and partly because this task has the smallest training subset (∼2.5% of total samples). The data scarcity issue should be addressed in future benchmark versions.

5.2 Dialogue Visualization

Figure 3 presents examples across six tasks from SARChat-Bench. The model completes these tasks with reasonable performance, and its coordinate predictions largely align with the ground truth annotations. In the spatial grounding task, the model identifies an additional ship that was not included in the original annotations, which may indicate potential for detecting previously unmarked targets in SAR imagery, though these could also be false positives requiring expert verification.

6 Conclusion

This research introduces SARChat-2M, a large-scale SAR instruction-following dataset automatically derived from SARDet-100K detection annotations, and SARChat-Bench, a deterministic multi-task benchmark built on its held-out split for evaluating SAR-oriented vision-language understanding. Rather than proposing a new model architecture, we provide a standardized SAR vision-language layer and evaluation suite that converts detection annotations into six task families. Experiments with 16 open-source VLMs show that SAR-domain tuning substantially improves coarse recognition and region-conditioned understanding, while multi-target localization, category-guided referring, counting with localization, and fine-grained description remain difficult. We publicly release the dataset, evaluation scripts, and trained weights to support future SAR VLM research.

Limitations

(1) Template bias. Most language in SARChat-2M is generated from templates. Models may learn to match format patterns rather than acquiring genuine natural-language understanding of SAR imagery. This limits the generalizability of observed improvements to free-form, human-written queries.
(2) Inherited annotation noise. The underlying visual annotations come from SARDet-100K and its integrated data sources. Residual annotation inconsistencies, missing targets, and imprecise bounding boxes are inherited by SARChat-2M. Notably, there exist cases where VLMs successfully identify valid targets that were not originally annotated, highlighting the annotation completeness challenge.
(3) Task imbalance. Region-Conditioned Identification accounts for approximately 77% of total samples, while Fine-Grained Description accounts for only $\sim$2.5%. This imbalance affects both training (models see far more identification samples) and evaluation (the weighted overall score is dominated by a single task). Per-task metrics should be the primary comparison criterion.
(4) In-domain evaluation only. All evaluation is conducted within the SARDet-100K train/test split. Results do not necessarily generalize to cross-sensor, cross-region, or cross-dataset scenarios. Future work should include out-of-distribution evaluation on independent SAR datasets.
(5) Category-level referring. The current category-guided referring task uses category prompts ("Where is the {category}?") rather than complex attribute- or relation-based referring expressions. It evaluates category-guided localization, not full referring expression comprehension.
(6) No human-written language benchmark. All prompts and responses in the current benchmark are template-generated. A human-written query test set with natural, diverse language would provide a stronger test of language generalization.

Ethics Statement

In this study, all SAR datasets and methodologies are used for academic research and benchmark development under their respective licenses and data usage agreements. SARChat-2M is constructed from publicly available remote-sensing resources and does not contain personally identifiable information. We encourage users of the dataset and trained models to follow applicable data policies, report limitations clearly, and avoid using benchmark scores as evidence of reliability in safety-critical deployments without additional domain validation.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Jie Chen, Zhixiang Huang, Runfan Xia, Bocai Wu, Lei Sheng, Long Sun, and Baidong Yao. 2022. Large-scale multi-class sar image target detection dataset-1.0 [ol]. *Journal of Radars*, (1).

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

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

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. 2018. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. 2024a. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*.

Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. 2024b. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Xin Lin, Bo Zhang, Fan Wu, Chao Wang, Yali Yang, and Huiqin Chen. 2023. Sived: A sar image dataset for vehicle detection based on rotatable bounding box. *Remote Sensing*, 15(11):2825.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

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

Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016a. Deep semantic understanding of high resolution remote sensing image. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5.

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

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE.

Chao Wang, Rui Ruan, Zhicheng Zhao, Chenglong Li, and Jin Tang. 2023. Category-oriented localization distillation for sar object detection and a unified benchmark. *IEEE Transactions on Geoscience and Remote Sensing*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. 2019. A SAR dataset of ship detection for deep learning under complex backgrounds. *Remote. Sens.*, 11(7):765.

Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. 2020. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *IEEE Access*, 8:120234–120254.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. 2022. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868*.

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

Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, Israr Ahmad, Dece Pan, Chang Liu, Yue Zhou, Jun Shi, and Shunjun Wei. 2021a. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18).

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

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

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. 2023. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

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

Appendix

A Implementation Details

During the fine-tuning stage, each open-source model is trained for 1 epoch with a batch size of 4 (effective batch size 32 with gradient accumulation steps of 4). This study employs LoRA (Hu et al., 2022) with rank=8 and alpha=32 on all linear layers. The learning rate is initialized at 1e-4 with a 0.1 warm-up ratio. All experiments are conducted on 2 NVIDIA A100 GPUs using bfloat16 precision. Training is implemented using the MS-SWIFT framework (Zhao et al., 2024) for efficient distributed training. In the most time-consuming case, the training process for a single model took up to 192 hours to complete.

B Full Benchmark Results

Table 5 reports the complete per-task scores for all 16 fine-tuned VLMs, including the weighted average score S_m. As discussed in Section 4.1, the weighted overall score is dominated by Region-Conditioned Identification ($\sim$77% of samples) and should not be used as the primary metric for model comparison.

C Extended Evaluation with Newer Models

To examine whether SARChat-Bench generalizes to newer model families, we additionally evaluate InternVL3.5 (1B–8B), Qwen3 (2B–8B), and RSCOVLM-7B, a SAR-domain vision-language model. All models are fine-tuned with the same LoRA protocol described in Section 5.1. Results are shown in Table 6.

Analysis. InternVL3.5-8B achieves the highest performance across most tasks, confirming continued improvement of general-purpose VLMs. Notably, the SAR-specific RSCOVLM-7B attains the best single-target spatial grounding score (94.83), outperforming all general-purpose models—suggesting that domain-specific pre-training benefits localization in SAR imagery. However, its Fine-Grained Description score (57.67) remains lower than InternVL3.5-8B (67.80), indicating room for improvement in semantic generation. Consistent with the main evaluation (Section 5.2), Classification and Region-Conditioned Identification remain near-saturated (>96% and >97%, respectively), while Category-Guided Referring continues to be the most challenging task across all model families.

D Data structure analysis

D.1 Word Frequency Analysis of SARChat-2M

As shown in Figure 4 , location words (such as center, middle, top) and target object words (such as ship, aircraft, tank) have the highest occurrence frequencies in SAR image descriptions, and the adjective "small" is the most frequently used descriptive word.

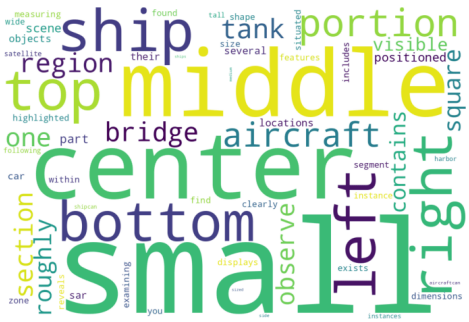

Figure 4: Cloud Map of Word-frequency Distribution

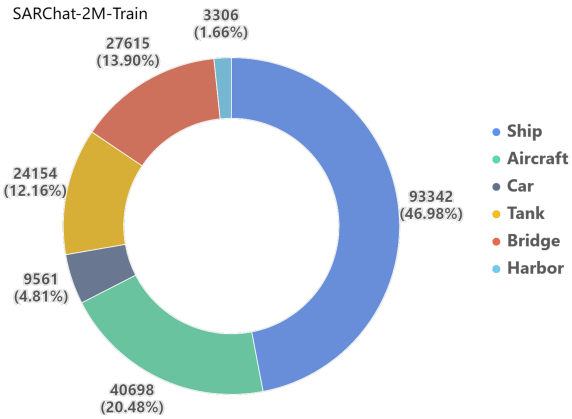

Figure 5: The Proportion Distribution of Samples in the Training Set

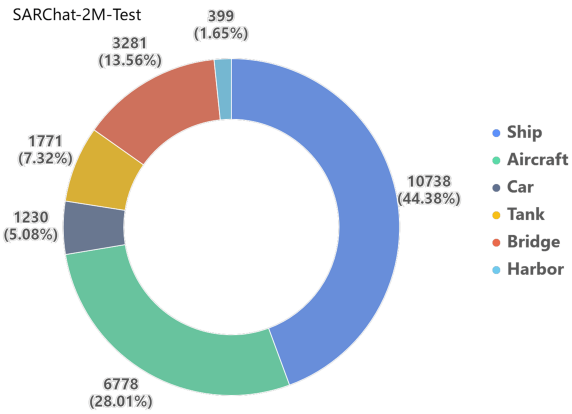

Figure 6: The Proportion Distribution of Samples in the Testing Set

Model	Param	Avg score	Tasks													
			Only count	Instance Count		Abstract position	Spatial Ground		Reg.-Cond. ID		Multi-target Ref		Single-target Ref		Descript	Class
				IoU=.25	IoU=.5		Multi	Single	Multi	Single	IoU=.25	IoU=.5	IoU=.25	IoU=.5		
InternVL2.5	**8B**	92.79	**74.14**	61.37	52.17	81.25	62.25	87.91	98.84	98.98	37.49	23.46	74.86	60.13	63.43	97.25
	4B	91.57	72.68	57.54	47.35	83.33	60.89	85.90	98.01	98.76	34.05	18.86	69.92	55.29	58.84	97.27
	2B	90.55	71.52	54.11	44.22	50.00	60.81	81.92	97.79	98.63	27.05	13.91	68.50	52.16	56.36	96.69
	1B	88.89	69.87	50.18	39.35	0.00	56.30	82.24	96.98	98.60	22.13	9.94	62.33	44.99	53.30	96.65
DeepSeekVL	**7B**	88.99	20.66	8.49	4.19	64.29	65.32	85.78	98.97	99.05	28.75	13.66	64.34	48.84	51.08	93.23
	1.3B	84.01	19.61	4.00	1.32	75.00	60.38	82.00	96.40	97.45	16.11	6.23	53.58	34.28	44.44	47.37
Phi-3.5-vision	**4.2B**	92.06	72.69	57.48	47.60	62.50	58.85	87.29	98.93	98.59	31.65	17.16	70.95	55.70	59.95	96.42
GLM-Edge-V	**2B**	90.20	71.59	51.97	40.37	42.86	59.15	86.33	97.54	98.60	24.15	10.66	65.57	46.46	57.86	97.39
	5B	90.48	73.44	56.30	44.56	75.00	61.38	89.96	96.69	95.96	30.68	15.41	69.36	51.81	61.45	98.02
mPLUG-Owl3	**7B**	91.71	71.00	48.07	35.27	**100.00**	56.37	93.32	**99.27**	**99.51**	19.72	7.66	57.27	38.00	54.65	98.80
	2B	90.32	67.56	41.56	28.83	75.00	45.65	97.58	98.95	99.42	14.91	5.42	50.46	30.16	41.76	98.31
	1B	89.68	67.03	38.64	24.98	75.00	44.07	97.19	98.72	98.87	11.86	4.12	44.34	24.02	40.16	98.06
Qwen2-VL	**7B**	90.76	**72.79**	58.51	50.24	0.00	64.17	83.87	97.54	99.18	39.11	26.29	70.55	57.04	63.11	97.30
	2B	90.27	69.63	53.62	45.47	50.00	59.04	78.49	97.55	99.26	32.60	20.12	65.31	51.53	55.20	96.88
LLaVA-1.5	**7B**	91.21	71.89	56.89	46.80	57.14	62.70	85.79	97.84	98.42	30.81	15.48	71.89	56.70	61.35	96.90
Yi-VL	**6B**	84.35	32.62	14.35	9.44	75.00	53.68	72.38	93.63	97.95	7.76	2.69	32.95	16.63	38.15	95.32

Table 5: Full performance comparison of all 16 fine-tuned VLMs on SARChat-Bench. The weighted average score (Avg) is dominated by Region-Conditioned Identification and should not be used as the primary comparison metric.

Model	Param	Count EM	Ground (Multi)	Ground (Single)	Reg.-Cond. ID (Multi)	Reg.-Cond. ID (Single)	Ref (Multi)	Ref (Single)	Descript	Class
InternVL3.5	8B	67.31	78.02	87.24	99.23	99.55	43.86	74.23	67.80	98.12
	4B	65.62	72.64	86.43	98.10	99.45	42.07	72.28	66.54	98.05
	2B	60.21	55.88	83.27	98.32	98.31	30.72	64.29	59.38	96.84
	1B	53.94	60.06	80.75	98.81	98.31	19.28	55.40	51.48	96.78
Qwen3	8B	62.10	79.73	81.16	98.70	99.34	39.74	65.91	63.39	97.84
	4B	62.77	67.77	81.35	99.87	99.40	40.33	67.80	65.95	97.86
	2B	56.95	54.84	75.15	97.62	98.42	30.99	61.90	57.02	97.19
RSCOVLM	7B	62.64	78.21	**94.83**	98.35	99.01	40.04	66.85	57.67	97.54

Table 6: Extended evaluation on SARChat-Bench with InternVL3.5, Qwen3, and RSCOVLM. Ground = Spatial Grounding accuracy; Ref = Category-Guided Referring accuracy; EM = Exact Match. Metric definitions follow Section 4.1.

D.2 Analysis of Datasets Composition

Training Set Category Distribution

As shown in Figure 5, in the SARChat-2M training set, the category distribution shows significant differences. Among them, the "Ship" category has the largest proportion, reaching 46.98%, followed by the "Aircraft" category, with a proportion of 20.48%. These two categories account for the majority of the samples in the training set. It can be seen that the sample distribution in the training set is imbalanced, and the "Ship" and "Aircraft" categories dominate. This may enable the model to learn the features of these two categories more comprehensively during the training process. However, since the samples of other categories are relatively few, the model's ability to learn and generalize their features may be affected to a certain extent.

Test Set Category Distribution

As shown in Figure 6, in the SARChat-2M test set, the "Ship" category has the largest proportion among all categories, reaching 44.38%, and the "Aircraft" category ranks second, with a proportion of 28.01%. The distribution trends of these two major categories in the test set are similar to those in the training set. This indicates that the test set has a certain similarity to the training set in terms of the overall category distribution and can be used to test the model's generalization ability on data with a similar distribution. However, the slight differences in the proportions also remind us to comprehensively consider various factors when evaluating the model's performance.

Region-Conditioned Identification Shape Distribution Analysis

As shown in Figure 7, in the region-conditioned identification shape distribution, the "Roughly Square" shape has the largest proportion, with a quantity of 179,652. This shape has an absolute advantage among all shape categories. This means that in the region-conditioned identification task, the number of samples of the "Roughly Square" shape is much larger than that of other shapes. The model may be more sensitive to this shape and tend to identify the target as the "Roughly Square" shape during the recognition process. Therefore, when training and optimizing the model, attention should be paid to improving the recognition ability of other shapes to achieve a more balanced recognition effect.

Distribution Analysis of Morphological Categories

As shown in Table 7, which is computed over valid object-level bounding boxes after morphology-related filtering (rather than over image-level samples), nearly square-shaped mor-

Category	Dataset	Valid Bboxes	Mean AR	Median AR	Std Dev	AR Distribution (%)		
						AR≤0.67	0.67<AR≤1.5	AR>1.5
Ship	Train	93,342	1.34	1.07	1.24	28.37	39.67	31.96
	Test	10,738	1.308	1.026	1.10	29.34	39.82	30.84
Aircraft	Train	40,698	1.074	1.047	0.32	5.85	87.67	6.48
	Test	6,778	1.08	1.041	0.31	4.56	87.36	8.08
Car	Train	9,561	1.23	1.08	0.56	13.18	60.07	26.75
	Test	1,230	1.21	1.07	0.53	12.28	62.11	25.61
Tank	Train	24,154	1.10	1.00	0.84	1.58	94.29	4.13
	Test	1,771	1.09	1.00	0.29	1.41	94.36	4.23
Bridge	Train	27,615	1.56	1.18	1.92	18.38	44.83	36.79
	Test	3,281	1.568	1.2	1.24	18.01	44.59	37.4
Harbor	Train	3,306	1.20	1.01	0.72	14.19	67.93	17.88
	Test	399	1.23	1.01	0.81	15.04	68.42	16.54

Table 7: Aspect Ratio Analysis of Valid Object Bounding Boxes (after filtering invalid coordinates, extreme aspect ratios, and out-of-bound regions)

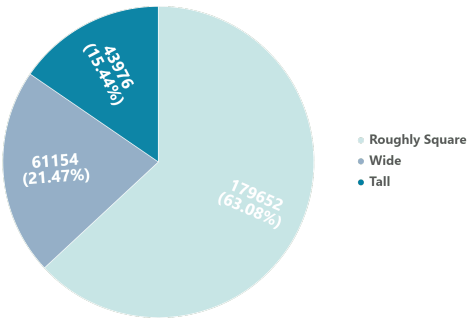

Figure 7: Morphological distribution

phology dominates both datasets, accounting for 58.31% in training and 59.38% in test sets (category-weighted averages), indicating its prevalence in target morphologies. Note that the totals in Table 7 may differ slightly from the category counts in Table 1 due to the removal of invalid bounding boxes. Wide-body shapes maintain stable distributions (18.14% training, 17.72% test), while tall-body shapes show a moderate decrease from training (23.54%) to test (22.90%) sets. This distribution diversity provides varied morphology patterns for training, and the consistent morphology distributions between training and test sets support comparable in-domain evaluation.

Category-Specific Morphological Patterns

As shown in Table 7, each category displays distinctive morphological characteristics. Bridge exhibits the highest average aspect ratio (1.56) with balanced distribution across all morphologies (18% wide-body, 45% square, 37% tall-body). Ship demonstrates diverse shapes (28% wide-body, 40% square, 32% tall-body), reflecting their real-world variability. Tank and Aircraft show highly concentrated distributions, with nearly square shapes dominating at 94% and 87% respectively, facilitating efficient model learning. Car and Harbor maintain moderate distributions with 60-68% nearly square shapes and balanced remaining proportions.

D.3 Analysis of Task based on Datasets Composition

Our dataset is designed to support unified multi-modal tasks through a comprehensive task taxonomy, as shown in Table 8. It encompasses six primary tasks: Classification, Fine-Grained Description, Instance Counting, Spatial Grounding, Region-Conditioned Identification, and Category-Guided Referring. Among these, the first three tasks are target-quantity independent, while Spatial Grounding, Region-Conditioned Identification, and Category-Guided Referring are further categorized into single-object and multi-object variants. This systematic organization enables diverse training scenarios and supports more fine-grained in-domain evaluation.

Task Type	Train	Test
Instance Counting	95493 (5.2%)	11794 (5.2%)
Spatial Grounding	94456 (5.1%)	11608 (5.1%)
Region-Conditioned Identification	1423548 (77.5%)	175565 (77.4%)
Category-Guided Referring	95486 (5.2%)	11703 (5.2%)
Fine - Grained Description	46141 (2.5%)	6032 (2.7%)
Classification	81788 (4.5%)	10024 (4.4%)

Table 8: Task type distribution in training and test sets

As illustrated in Figure 8, the training set comprises 1,836,912 entries. Region-Conditioned Identification dominates with 1,423,548 entries (77.50%), enabling extensive region-conditioned feature learning. Instance Counting and Category-

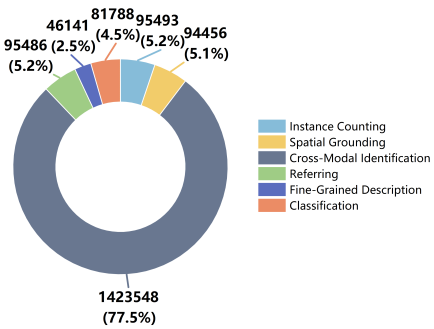

Figure 8: Train Task Distribution

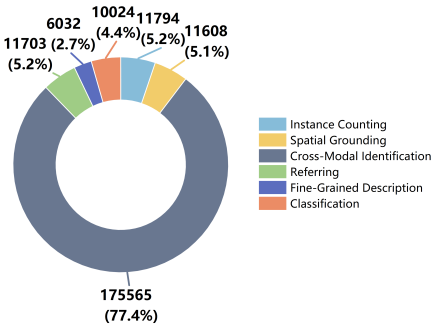

Figure 9: Test Task Distribution

Guided Referring tasks contain 95,493 (5.20%) and 95,486 (5.20%) entries respectively, while Spatial Grounding accounts for 94,456 entries (5.14%). Fine-Grained Description includes 46,141 entries (2.51%), with its relatively limited data volume potentially affecting model performance. The Classification task contains 81,788 entries (4.45%).

The test set (Figure 9) maintains a parallel distribution across its 226,636 entries. Region-Conditioned Identification remains dominant with 175,565 entries (77.47%), followed by Instance Counting (11,704, 5.16%), Category-Guided Referring (11,703, 5.16%), Spatial Grounding (11,608, 5.12%), Classification (10,024, 4.42%), and Fine-Grained Description (6,032, 2.66%). This consistent distribution ensures comparable evaluation conditions between training and test sets.

The train and test splits maintain consistent task distributions, but the distribution is intentionally not balanced. Region-Conditioned Identification dominates the dataset because each annotated region can naturally generate a region-conditioned identification sample, whereas fine-grained image-level descriptions are generated more conservatively. While the consistency between training and test sets ensures comparable evaluation conditions, two aspects warrant attention: the relatively limited data in Fine-Grained Description tasks may constrain model performance and could benefit from expansion, while the dominant proportion of Region-Conditioned Identification data necessitates careful consideration of overfitting during training through appropriate regularization methods.

D.4 Dataset Examples

Classification Task:

User: "Please identify the target categories in the figure (candidate categories: {total category list})"
AI: "The {specific category} targets can be identified in the figure"

Spatial Grounding Task:

User: "[grounding] Can you describe the spatial arrangement of various features in this SAR image?"
AI: "In this SAR image, I can identify several objects and their locations: There is a {category-1} in the top right portion {bbox}. There are 2 {category-2} in the bottom left: one in the bottom left portion {bbox}, another in the middle left portion {bbox}. The {category-1} is far from the {category-2}."

Fine-Grained Description Task:

User: "Please describe the content of the satellite image in detail"
AI: "In the figure, {size}{category} can be seen located at {orientation},"+"and there is also {size}{category} located at {orientation}"

Instance Counting Task:

User: "[count] How many {category} are there in the image?"
AI: "There are 3 instances. {bbox}{bbox}{bbox}"

Region-Conditioned Identification Task:

User: "[identify] What object is located in {bbox}?"
AI: "This is a medium-sized {category} located in the top right portion of the image. It appears to be wide in shape."

Category-Guided Referring Task:

User: "[refer] Where is the {category}?"
AI: "{bbox} (top right)."

E Motivation for the Dual-benchmark Setup

For spatial grounding and category-guided referring, we construct single-target and multi-target localization settings. For region-conditioned identification, we distinguish single-region and multi-region recognition settings. The core research motivations are as follows:

(1) Evaluation of the Model's Target Discrimination Ability

The single-target scenario aims to test the model's basic recognition ability for independent targets. In contrast, the multi-target scenario focuses on examining the model's ability to separate and select targets in complex environments, especially when multiple targets exhibit similar features. This dual-benchmark design can effectively diagnose the performance differences of the model in scenarios with varying degrees of complexity.

(2) Revelation of Target Association Understanding Issues

In multi-target scenarios, the model usually needs to understand the spatial and semantic relationships between targets. By comparing the performance differences of the model in single-target and multi-target scenarios, it is possible to evaluate whether the model truly understands the descriptions of the positional relationships between targets. This helps to identify the limitations of the model when dealing with relative position descriptions such as "the vehicle on the left" and "the tank in the middle".

(3) Exposure of Attention Mechanism Defects

In multi-target scenarios, the model is highly prone to problems such as attention divergence or overlap. When there are multiple similar targets in an image, the model may have difficulty accurately locating the specific target described by the user. Through the comparison between single-target and multi-target scenarios, the deficiencies of the model in attention allocation can be clearly demonstrated.

(4) Simulation of Real-world Application Scenarios

Real-world applications cover both simple single-target scenarios and complex multi-target environments. The establishment of the dual-benchmark is more in line with real-world usage requirements, providing a more comprehensive dimension for model evaluation and helping to improve the applicability and reliability of the model in actual deployments.

F Task-specific Performance Scoring

To evaluate model performance on each task, the task-specific accuracy is calculated by Formula 8. For each task t, the accuracy score $a_{m,t}$ of model m is computed by averaging the accuracy scores across all subtasks:

$$a_{m,t} = \frac{1}{k} \sum_{i=1}^{k} a_{m,t,i} \tag{8}$$

where $a_{m,t}$ is the average accuracy of model m on task t, k is the number of subtasks, and $a_{m,t,i}$ is the accuracy of model m on the i-th subtask of task t. This approach ensures that each subtask contributes equally to the overall task score.

G The Analysis of Model Size

Based on the data analysis in Table 9, it can be concluded that for most task-related metrics, there is a trend of performance improvement as the model size increases from 1B to 8B. For example, the instance-counting accuracy rises from 69.87% to 74.14%, the spatial-grounding accuracy increases from 56.30% to 62.25%, the fine-grained description accuracy goes up from 53.30% to 63.43%, and the classification accuracy climbs from 96.65% to 97.25%. This indicates that an increase in model size is beneficial to enhancing the performance of these tasks. However, the accuracy of abstract locations in the spatial-grounding task shows a unique trend of change. This metric increases from 0.00% for the 1B model to 83.33% for the 4B model, but then decreases to 81.25% for the 8B model, not increasing monotonically with the model size. Evidently, the influence of model size on some specific tasks follows complex patterns. Therefore, when selecting a model, it is necessary to comprehensively consider the task type and model size to achieve optimal performance.

H Comparison before and after training

To evaluate the impact of the SARChat-2M training dataset, we conducted a comparative analysis on InternVL-2.5-8B, the best-performing model on SARChat-Bench —before and after fine-tuning. Our evaluation metrics reveal that without fine-tuning on SARChat-2M, the model fails to comprehend most SAR-specific targets, exhibiting near-zero performance on tasks involving target interpretation and description. The only exception is the Instance Counting Task, where InternVL2.5-8B achieves a baseline accuracy of 63.71%. This relatively high performance can be attributed to the prevalence of Ship-on-sea samples, where SAR imaging exhibits distinctive characteristics: the smooth sea surface creates specular reflection, causing most electromagnetic waves to scatter away

Performance	8B	4B	2B	1B
Instance Counting Accuracy	74.14	72.68	71.52	69.87
Instance Counting Accuracy (IoU = 0.25)	61.37	57.54	54.11	50.18
Instance Counting Accuracy (IoU = 0.5)	52.17	47.35	44.22	39.35
Spatial Grounding Accuracy	62.25	60.89	60.81	56.30
Abstract Location in Spatial Grounding Accuracy	81.25	83.33	50.00	0.00
Spatial Grounding Single Accuracy	87.91	85.90	81.92	82.24
Region-Conditioned Identification (Multi) Accuracy	98.84	98.01	97.79	96.98
Region-Conditioned Identification (Single) Accuracy	98.98	98.76	98.63	98.60
Category-Guided Referring (Multi) Accuracy (IoU = 0.25)	37.49	34.05	27.05	22.13
Category-Guided Referring (Multi) Accuracy (IoU = 0.5)	23.46	18.86	13.91	9.94
Category-Guided Referring (Single) Accuracy (IoU = 0.25)	74.86	69.92	68.50	62.33
Category-Guided Referring (Single) Accuracy (IoU = 0.5)	60.13	55.29	52.16	44.99
Fine-Grained Description Accuracy	63.43	58.84	56.36	53.30
Classification Accuracy	97.25	97.27	96.69	96.65

Table 9: Performance comparison across different model sizes

Model	Only	Spatial Ground		Reg.-Cond. ID		Multi-target Ref		Single-target Ref		Descript	Class
	count	Multi	Single	Multi	Single	IoU=.25	IoU=.5	IoU=.25	IoU=.5		
InternVL2.5(8B)	63.71	5.07	0.38	0.12	0.17	0	0	0	0	0.04	17.8
SARChat-InternVL2.5(8B)	74.14	62.25	87.91	98.84	98.98	37.49	23.46	74.86	60.13	63.43	97.25

Table 10: Performance comparison of InternVL2.5 before and after fine-tuning

from the sensor direction, resulting in weak returns that appear as dark areas. Meanwhile, the metallic structures and geometric features of ships (such as dihedral and trihedral corner reflectors) generate strong backscattering, concentrating radar waves back to the sensor, thus appearing as bright spots. As shown in Figure 10, these samples present relatively straightforward recognition scenarios, leading to higher counting accuracy scores. The comparative results between the base model and its fine-tuned version on SARChat-Bench are presented in Table 10. Overall, fine-tuning with the SARChat-2M dataset proves essential for enabling VLMs to interpret SAR imagery effectively.

H.1 Edge-side models for SAR Applications

This study has multiple edge-side models ($\leq$5B parameters) trained on SARChat-2M and evaluates their performances. According to Tables 3 and 4, these models exhibit task-specific performance variations, achieving strong accuracy in region-conditioned identification, while showing potential for improvement in category-guided referring tasks. These models support domain-specific fine-tuning for rapid task adaptation. After optimization, they can operate efficiently on satellite or ground-edge devices, enabling real-time SAR data processing while reducing dependence on cloud infrastructure and minimizing data transmission costs.

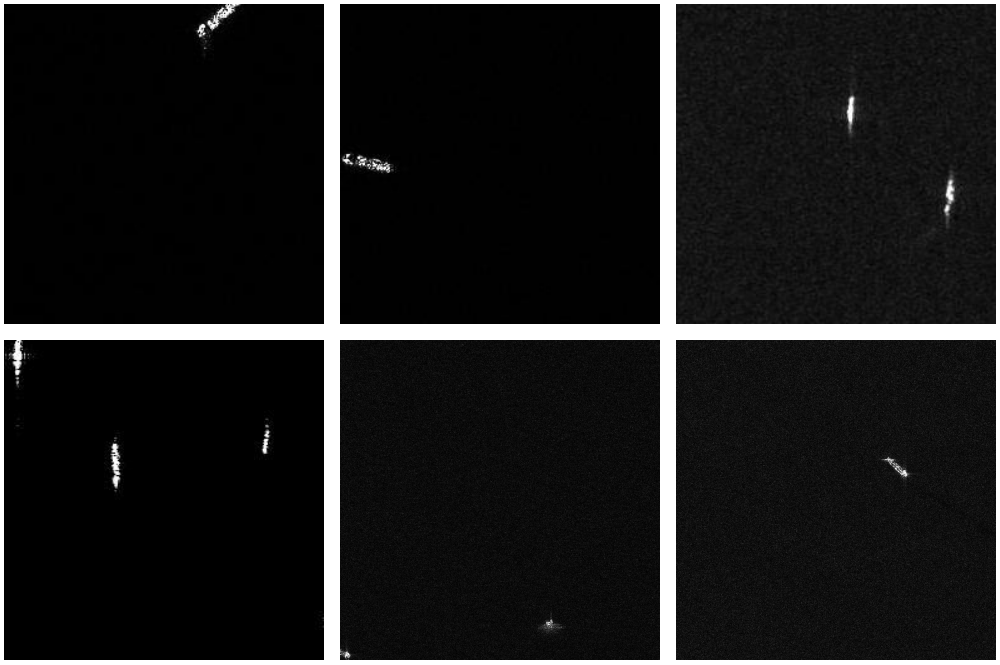

Figure 10: **Simple Examples of Ship detection in SAR images.** Ships appear as distinct bright spots in these SAR images, making them relatively easy to count even for VLMs without SAR-specific training.