# SAR-TEXT: From Imperfect Multimodal Earth Observation to Large-Scale SAR–Language Supervision

Anonymous CVPR submission

Paper ID *****

## Abstract

*Multimodal Earth observation is rarely complete, synchronized, or uniformly informative in practice. Optical imagery may be unavailable or unreliable under cloud cover, nighttime conditions, or adverse weather, whereas synthetic aperture radar (SAR) remains observable but is substantially harder to interpret semantically. This modality gap limits the applicability of modern vision–language models in realistic remote sensing pipelines. In this paper, we study whether large-scale SAR–language supervision can serve as a practical bridge for multimodal representation learning under imperfect observations. We present SAR-TEXT, a 136,584-pair SAR image–text dataset built from heterogeneous SAR sources using a multi-stage caption generation pipeline, including annotation-to-caption conversion, segmentation-guided captioning, and rule-guided rewriting from optical descriptions. We further adopt a progressive transfer strategy that adapts vision–language foundation models from natural images to optical remote sensing and then to SAR. Experiments on cross-modal retrieval, caption generation, and downstream SAR visual question answering show that large-scale SAR–language supervision substantially improves performance over direct-transfer baselines. Human auditing further indicates that the automatically generated captions are generally usable at scale, while failure cases reveal the main bottlenecks under imperfect semantic supervision. Our results suggest that SAR–language alignment is a promising mechanism for robust multimodal remote sensing when observations are heterogeneous, incomplete, weakly paired, or only partially observable.*

## 1. Introduction

Earth observation increasingly relies on multiple sensing modalities, including optical satellites, synthetic aperture radar (SAR), infrared imagery, and geospatial metadata. In principle, combining these modalities enables richer scene understanding, improved retrieval, stronger foundation models, and more reliable downstream decision-making [4, 11, 18, 22, 25, 29]. In practice, however, multimodal monitoring is often imperfect. Optical observations may be corrupted by clouds, shadows, illumination changes, or acquisition gaps; sensor pairs may be only loosely aligned across space and time; and some modalities may be absent at inference time. These realities motivate a central question for realistic multimodal Earth observation: *how can semantic interpretability be preserved when the observation space is incomplete or heterogeneous?*

SAR is especially relevant in this setting. Because it is an active sensor and is robust to weather and illumination, SAR often remains available when optical imagery is degraded or missing [14, 31]. At the same time, SAR is notoriously difficult to interpret semantically. Its backscattering-based formation mechanism yields patterns that differ sharply from the appearance cues familiar from natural and optical remote sensing images. As a result, state-of-the-art vision–language models trained on natural or optical imagery do not directly transfer well to SAR, and the lack of large-scale SAR–text supervision remains a major bottleneck [6, 20, 30, 32].

In this paper, we argue that SAR–language supervision provides a useful bridge for multimodal Earth observation under imperfect conditions. Rather than viewing SAR as an isolated modality, we treat SAR–text alignment as a mechanism for preserving semantic accessibility when other modalities are unreliable, weakly paired, or unavailable. Building on this perspective, this study is centered on two components: (1) *SAR-TEXT*, a large-scale corpus of 136,584 SAR image–text pairs; and (2) a progressive transfer strategy that adapts vision–language models from natural images to optical remote sensing and finally to SAR.

Our starting point is the observation that manual SAR captioning is expensive, hard to scale, and often requires substantial domain expertise [30]. We therefore use a multi-stage caption generation pipeline, called *SAR-Narrator*, to transform object detection labels, semantic segmentation maps, and aligned optical descriptions into natural-

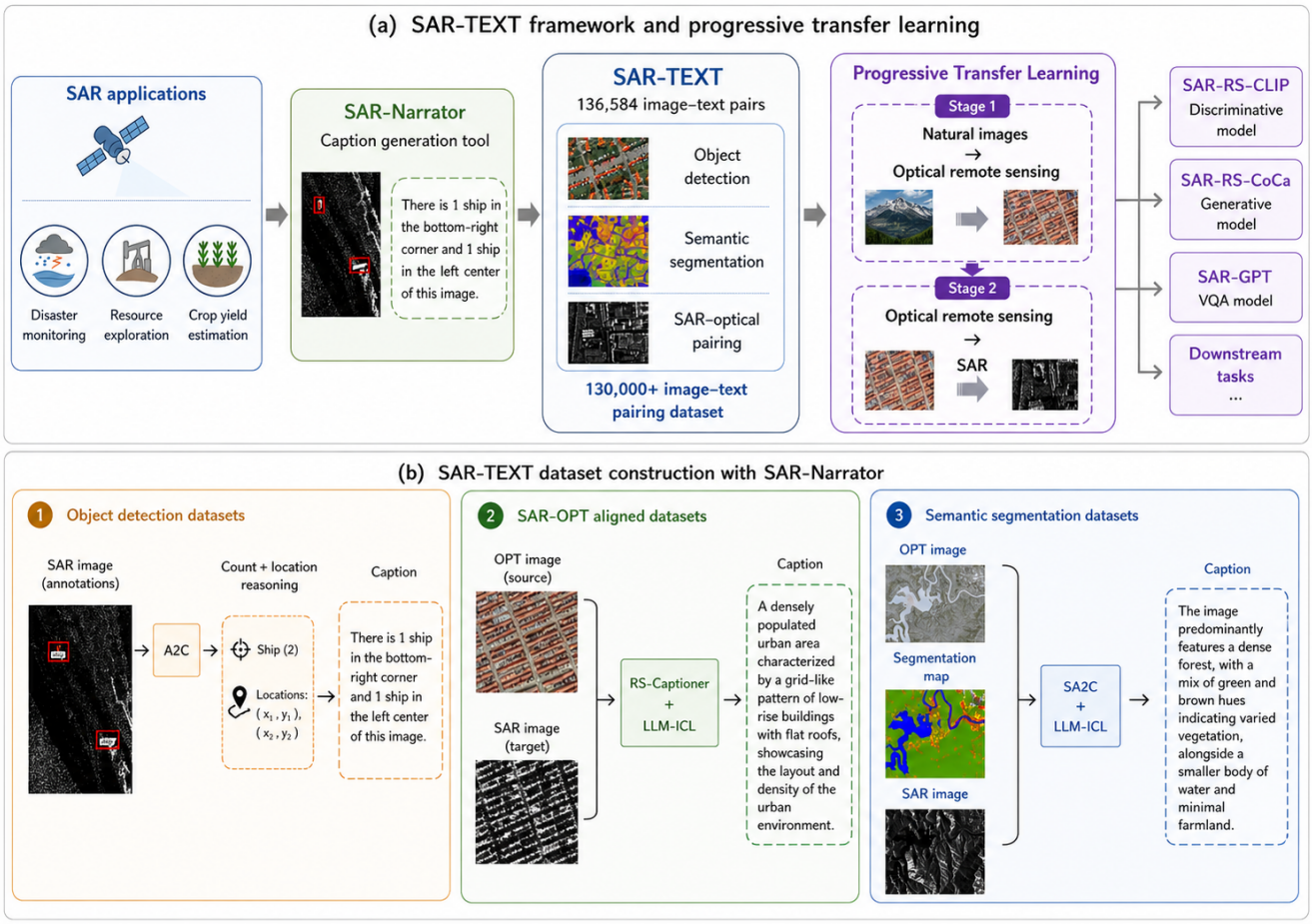


Figure 1. Overview of the proposed SAR–language supervision framework. SAR-Narrator converts heterogeneous SAR supervision sources into SAR-TEXT, which is then used to adapt vision–language models through progressive transfer from natural images to optical remote sensing and finally to SAR.

language supervision for SAR images. This yields broad coverage across object-centric scenes, natural landscapes, urban areas, and co-registered SAR–optical pairs. We then study whether such supervision improves retrieval and generation models in realistic settings where semantic alignment must survive strong modality mismatch.

Our contributions are fourfold. First, we reframe large-scale SAR–language supervision as a response to imperfect multimodal Earth observation. Second, we present SAR-Narrator, a practical pipeline for building large-scale SAR–text corpora from heterogeneous supervision sources. Third, we show that progressive transfer from natural images to optical remote sensing and then to SAR improves both retrieval and caption generation. Fourth, we analyze human verification and failure modes to characterize what kinds of semantic errors remain when supervision is generated automatically at scale, and we further examine SAR-GPT as a downstream VQA extension of SAR-TEXT.

## 2. Related Work

### 2.1. Vision–Language Learning for Remote Sensing

Vision–language learning has become an important direction in remote sensing, with recent work exploring image–text alignment, zero-shot recognition, scene understanding, change analysis, and multimodal question answering [4, 7, 9, 11, 18, 22, 25, 27, 29]. Existing remote sensing VLMs have primarily focused on optical imagery, where semantic categories are more accessible and large-scale image–text supervision is easier to collect. Foundation-style efforts have demonstrated promising results for optical remote sensing retrieval and transfer, but their assumptions do not directly carry over to SAR.

A key difficulty is that the gap between SAR and optical imagery is not merely a domain shift; it is a stronger modality shift rooted in distinct imaging physics. Models pretrained on natural images often rely on appearance statistics that do not exist in SAR. This makes naive transfer

Table 1. Source datasets used to construct SAR-TEXT.

| Dataset | Count | Type |
|---|---|---|
| MSAR-1.0 | 28,449 | Detection |
| SAR-Ship | 43,819 | Detection |
| OSdataset | 11,245 | SAR–Optical pair |
| HRSID-JPG | 3,642 | Detection |
| QXSLAB_SAROPT | 20,000 | SAR–Optical pair |
| SEN12 | 18,094 | SAR–Optical pair |
| optical_sar | 8,575 | SAR–Optical pair |
| whu-sar-opt | 1,600 | Segmentation |
| SSDD | 1,160 | Detection |
| Total | 136,584 | – |

unstable or ineffective, motivating intermediate adaptation through optical remote sensing before final SAR specialization [6, 32].

### 2.2. SAR Image–Text Data

Compared with optical remote sensing, SAR image–text resources remain limited in scale, diversity, and public accessibility [20, 30]. Early attempts at SAR captioning demonstrated feasibility but were typically constrained to narrow object categories or private datasets [30]. More recent multimodal remote sensing resources may include some SAR samples, but SAR usually represents only a small portion of the full corpus, and the resulting captions often lack SAR-specific expressiveness [17, 28, 29].

This lack of supervision is especially problematic under imperfect multimodal settings. When optical data are missing or unreliable, SAR may become the primary source of evidence. Without language supervision tailored to SAR, downstream multimodal models lose semantic interpretability precisely when robustness is most needed.

### 2.3. Imperfect Multimodal Observation

Real-world multimodal Earth observation often requires monitoring under incomplete, asynchronous, and uncertain observations. In remote sensing, such imperfections arise naturally from weather, seasonality, sensor availability, spatial misregistration, and temporal mismatch. While many multimodal pipelines assume clean paired data, practical systems must operate even when pairings are weak or one modality is absent. Existing work on SAR–optical fusion, registration, and joint modeling already highlights this challenge [5, 8, 13, 15, 21]. Our work addresses this setting by using SAR–text alignment as a robust semantic channel that can survive strong heterogeneity and partial pairing.

## 3. Method

Our approach consists of two parts: building a large-scale SAR–text corpus with heterogeneous supervision sources, and adapting vision–language models to SAR through progressive transfer.

### 3.1. Design Rationale

The goal of SAR-TEXT is not only to increase the number of SAR image–text pairs, but also to provide language supervision that is robust to heterogeneous and imperfect observation sources. Different SAR datasets expose different forms of supervision: object detection datasets provide reliable instance-level labels and coarse spatial locations; semantic segmentation datasets provide dense land-cover information; and SAR–optical aligned datasets provide richer scene-level context through paired optical observations. Instead of forcing all datasets into a single annotation format, SAR-Narrator uses source-specific conversion strategies and maps them into a shared natural-language space.

This design has two advantages. First, it preserves the strongest signal available from each source. Detection annotations are most reliable for object counts and positions, segmentation annotations are useful for dominant land-cover categories, and optical captions provide high-level scene descriptions when optical imagery is available. Second, converting all signals into text enables a unified training interface for vision–language models. As a result, the same SAR-TEXT corpus can support both discriminative alignment, such as SAR–text retrieval, and generative interpretation, such as SAR captioning.

### 3.2. SAR-Narrator

SAR-Narrator converts available structured or paired information into natural-language supervision for SAR images. It uses three complementary paths depending on the source dataset.

**Detection-to-caption (A2C).** For object detection datasets, we convert category labels, instance counts, and coarse spatial layout into text. The captioning logic uses object counts together with a $3 \times 3$ spatial partition of the image, producing simple yet effective captions such as "There is 1 ship in the bottom-right corner of this image" or "There are 14 oil tanks in this image." This branch is useful for ship, aircraft, bridge, and oil-tank scenes. The A2C branch intentionally uses simple templates rather than long free-form descriptions. This choice reduces hallucination because object detection annotations provide reliable category and location information but do not provide global scene context. For each annotated instance, the image is divided into coarse spatial regions, and the object category, count, and approximate position are verbalized. When multiple objects are present, the caption aggregates repeated categories and describes their relative locations. This produces captions that are concise but strongly grounded in the available annotation.

**Segmentation-to-caption (SA2C).** For semantic segmentation data, we summarize dominant land-cover categories from pixel-level labels and then rewrite the resulting description into a more natural caption. This branch captures scene-level semantics such as farmland, city, village, water, forest, and roads. Instead of exposing raw percentages directly, the final rewriting prioritizes salient scene elements and improves fluency. The SA2C branch is designed for scene-level interpretation. Since segmentation maps provide class proportions rather than natural sentence descriptions, we first identify dominant and secondary land-cover categories and then convert them into a short semantic summary. Very small categories are suppressed to avoid unstable captions caused by annotation noise. The resulting caption emphasizes salient classes and their approximate dominance, which is particularly useful for natural scenes such as farmland, water bodies, forests, and mixed rural regions.

**Rule-guided rewriting from optical descriptions.** For co-registered SAR–optical pairs, we generate optical captions with a remote sensing captioner and then rewrite them into SAR-suitable descriptions using an LLM under explicit rules. The rewriting removes color-centric language, avoids speculative phrasing, suppresses camera-related details, and preserves only visually grounded content that plausibly transfers to SAR. A small set of manually written in-context examples stabilizes the rewriting process [3, 10]. This rewriting branch is the most expressive but also the most vulnerable to semantic drift. To reduce this risk, we constrain the rewriting process with explicit rules. Color-related descriptions are removed because SAR does not encode visible color; camera-related phrases are suppressed because they are irrelevant to SAR interpretation; and uncertain visual details are avoided unless they are supported by scene structure. The goal is not to translate optical appearance directly into SAR appearance, but to retain high-level semantics that remain valid across the SAR–optical modality gap.

This design reflects the imperfect-observation setting directly. Rather than assuming one clean supervision source, SAR-Narrator accepts heterogeneous inputs—boxes, masks, or paired optical observations—and turns them into language signals tailored to SAR.

### 3.3. SAR-TEXT Dataset Construction

We build SAR-TEXT from multiple public SAR datasets spanning three source types: detection datasets, semantic segmentation datasets, and SAR–optical registration datasets [1, 5, 8, 13, 15, 16, 19, 26]. These sources cover ships, aircraft, bridges, oil tanks, farmland, water, forest, villages, urban structures, ports, and other scene categories. After cleaning corrupted samples and removing duplicates or near-duplicates using perceptual hashing [24], we obtain a final corpus of 136,584 SAR image–text pairs with an average caption length of 16.3 English words.

### 3.4. Quality Control

Because SAR-TEXT is constructed automatically, quality control is essential. We apply several filtering steps before training. First, corrupted images and unreadable files are removed. Second, near-duplicate samples are filtered using perceptual hashing to reduce over-representation of repeated scenes. Third, captions generated from each source are normalized to reduce formatting inconsistency, redundant punctuation, and repeated phrases. Finally, captions that are excessively short, excessively long, or structurally malformed are removed.

These filtering steps are particularly important for SAR–language learning. Unlike optical imagery, SAR images often contain speckle noise and ambiguous texture patterns, which can amplify the effect of noisy captions during contrastive learning. By removing duplicated or malformed samples, the final dataset provides more stable supervision for both retrieval and caption generation.

Unlike short class labels, many captions encode both scene semantics and coarse spatial structure. This is valuable for multimodal representation learning because it supports both discriminative retrieval and generative captioning.

### 3.5. Human Verification and Error Profile

Automatically generated supervision is useful only if its quality is high enough for learning. We therefore audit a random subset of 500 image–caption pairs. Two annotators with remote sensing experience independently score each pair on object correctness, count correctness, spatial correctness, scene correctness, and fluency using a 1–5 Likert scale, and also provide a binary accept/reject label.

The overall acceptance rate is 84.4% (422/500). Average scores are 4.24 for count correctness, 4.27 for fluency, 3.96 for spatial correctness, 3.88 for object correctness, and 3.76 for scene correctness. These results suggest that the captions are generally usable at scale, though scene-level interpretation remains the most challenging aspect.

Failure analysis shows that the dominant errors are wrong scene descriptions (35.9%), hallucination (19.2%), and style bias (17.9%), with fewer cases involving missed or wrong objects. This pattern is informative for imperfect multimodal learning: the easier parts of supervision are local counts and coarse positions, whereas the harder parts involve global scene interpretation and semantic grounding.

The error distribution also reveals that different supervision sources fail in different ways. Detection-derived captions are usually reliable for object categories and counts, but they may oversimplify complex scenes. Segmentation-

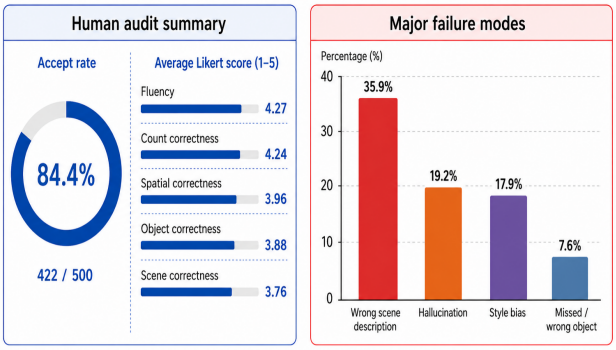


Figure 2. Human verification and failure-mode analysis of SAR-TEXT. The left panel reports the acceptance rate and dimension-wise Likert scores from human auditing, while the right panel summarizes the dominant error categories among rejected samples.

Table 2. Human verification results on 500 sampled SAR-TEXT pairs.

| Metric | Value |
|---|---|
| Sample size | 500 |
| Overall accept rate | 84.4% (422/500) |
| Object correctness | $3.88 \pm 0.66$ |
| Count correctness | $4.24 \pm 0.59$ |
| Spatial correctness | $3.96 \pm 0.55$ |
| Scene correctness | $3.76 \pm 0.74$ |
| Fluency | $4.27 \pm 0.53$ |

derived captions capture land-cover composition, but they may miss fine-grained structures and object-level semantics. SAR–optical rewriting provides richer descriptions, but it can introduce style bias or retain optical-centric assumptions if the rewriting constraints are insufficient. This suggests that caption noise in SAR-TEXT is structured rather than random, which partly explains why the dataset remains useful for downstream learning despite imperfect automatic generation.

### 3.6. Progressive Transfer for SAR–Language Alignment

A central difficulty in SAR–language learning is the large jump from natural-image pretraining to SAR imagery. Direct adaptation may abruptly change the visual embedding distribution and weaken cross-modal alignment with language. We therefore use progressive transfer with an intermediate optical remote sensing stage, following the broader intuition of transfer learning as staged distribution adaptation [6, 32]:

1. **Natural/optical adaptation:** start from standard vision–language backbones and continue pretraining on an optical remote sensing image–text corpus.
2. **SAR adaptation:** fine-tune the adapted models on SAR-TEXT.

We apply this strategy to both a CLIP-style retrieval model and a CoCa-style generative model. The intuition is that optical remote sensing shares semantic viewpoint and scene structure with SAR better than natural imagery does, while still remaining closer to the visual statistics learned during natural-image pretraining.

## 4. Experiments

### 4.1. Setup

We evaluate retrieval on HRSID-test [19] and OSdataset-512-test, representing object-centric and scene-diverse SAR scenarios, respectively. For retrieval, we report image-to-text Recall@K, text-to-image Recall@K, and mean recall. For caption generation, we report SPICE, BLEU-1/2/3/4, METEOR, ROUGE-L, and CIDEr.

For the retrieval backbone, we use CLIP ViT-L/14 [12]. For the generative backbone, we use CoCa ViT-L/14 [23]. Training follows a full fine-tuning setup on a single RTX 4090 GPU using the OpenCLIP codebase [2].

### 4.2. Implementation Details

For retrieval experiments, we fine-tune a CLIP ViT-L/14 backbone using image–text contrastive learning. Each training batch contains paired SAR images and captions from SAR-TEXT, and the model is optimized to align corresponding visual and textual embeddings while separating non-matching pairs. For caption generation, we fine-tune a CoCa ViT-L/14 backbone to jointly preserve image–text alignment and autoregressive captioning ability. The same SAR-TEXT training split is used for both discriminative and generative models.

In the progressive transfer setting, the models are first adapted to optical remote sensing image–text data and then further fine-tuned on SAR-TEXT. This two-stage protocol is kept consistent across CLIP and CoCa to isolate the effect of progressive transfer. During evaluation, no optical images are provided to the SAR models; all retrieval and caption generation results are obtained from SAR images alone.

For the downstream VQA extension, we construct SAR-VQA instructions from SAR-TEXT and fine-tune a compact multimodal model to obtain SAR-GPT. This evaluation is not intended to replace the retrieval and captioning benchmarks, but to test whether SAR-TEXT provides supervision that can support interactive semantic reasoning over SAR images.

### 4.3. Cross-Modal Retrieval Results

Across both test sets, direct transfer from generic CLIP performs poorly on SAR. Fine-tuning on SAR-TEXT im-

Table 3. Retrieval results on HRSID-test.

| Model | i2t-R@1 | i2t-R@5 | i2t-R@10 | t2i-R@1 | t2i-R@5 | t2i-R@10 | Mean Recall |
|---|---|---|---|---|---|---|---|
| CLIP | 0.15 | 0.56 | 1.02 | 0.05 | 0.51 | 0.92 | 0.54 |
| HQRS-CLIP | 0.00 | 0.36 | 0.61 | 0.05 | 0.46 | 1.07 | 0.42 |
| SAR-CLIP | 2.09 | 9.64 | 18.71 | 2.29 | 10.91 | 19.58 | 10.54 |
| SAR-RS-CLIP | **2.65** | **10.91** | **20.50** | **2.55** | **11.78** | **20.75** | **11.52** |

Table 4. Retrieval results on OSdataset-512-test.

| Model | i2t-R@1 | i2t-R@5 | i2t-R@10 | t2i-R@1 | t2i-R@5 | t2i-R@10 | Mean Recall |
|---|---|---|---|---|---|---|---|
| CLIP | 0.71 | 3.07 | 4.48 | 0.71 | 3.77 | 8.02 | 3.46 |
| HQRS-CLIP | 1.89 | 5.90 | 8.96 | 3.54 | 8.02 | 12.26 | 6.76 |
| SAR-CLIP | 4.48 | 14.86 | 23.82 | 4.72 | 20.99 | 29.72 | 16.43 |
| SAR-RS-CLIP | **5.66** | **16.04** | 23.11 | **5.19** | 20.75 | 28.77 | **16.59** |

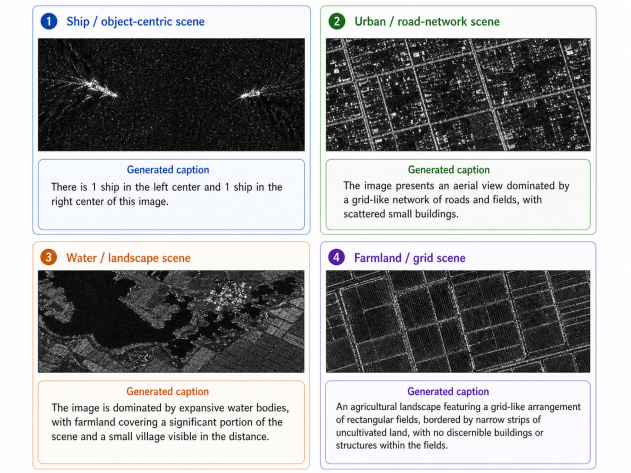


Figure 3. Qualitative captioning examples generated by SAR-RS-CoCa on representative SAR scenes, including object-centric, urban, water-dominated, and farmland-grid scenarios.

proves retrieval substantially, and the progressive strategy improves it further. On HRSID-test, SAR-RS-CLIP reaches a mean recall of 11.52, compared with 0.54 for CLIP and 10.54 for single-stage SAR fine-tuning. On OSdataset-512-test, SAR-RS-CLIP reaches 16.59 mean recall, compared with 3.46 for CLIP and 16.43 for single-stage SAR fine-tuning.

These results support two claims. First, large-scale SAR–language supervision provides effective cross-modal alignment for SAR images. Second, progressive transfer offers a modest but consistent advantage over one-step adaptation, especially on more diverse scenes.

### 4.4. Caption Generation Results

Generic CoCa performs poorly on SAR, indicating that caption generation does not emerge from natural-image pre-training alone. Fine-tuning on SAR-TEXT yields large gains, and progressive transfer improves results further. On HRSID-test, SAR-RS-CoCa increases CIDEr from 0.005 for CoCa to 3.186 while also reaching the best BLEU-4, ROUGE-L, and METEOR scores. On OSdataset-512-test, SAR-RS-CoCa similarly yields the strongest overall performance.

These findings show that SAR–language supervision is useful not only for discriminative alignment but also for generative semantic interpretation.

Figure 3 provides qualitative examples from different scene types. In object-centric scenes, the model can identify sparse bright targets and generate captions involving both object counts and relative positions. In urban scenes, the model describes grid-like road structures and scattered buildings, indicating that it has learned scene-level layout cues rather than only isolated bright scatterers. In natural scenes, such as water-dominated and farmland regions, the model captures dominant land-cover patterns and expresses them in coherent language.

At the same time, these examples also illustrate the limitations of SAR captioning. The generated descriptions are generally correct at the coarse semantic level, but they remain less detailed than optical captions and may inherit common phrasing patterns from the training corpus. This observation is consistent with the human verification results, where fluency and count correctness are stronger than fine-grained scene correctness.

The progressive strategy consistently outperforms single-stage mixed training. This supports the idea that adapting first to optical remote sensing and then to SAR provides a smoother transition than jumping directly from natural-image statistics to SAR.

Table 5. Caption generation results on HRSID-test.

| Model | SPICE | B1 | B2 | B3 | B4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|---|
| CoCa | 0.043 | 0.092 | 0.036 | 0.009 | 0.003 | 0.077 | 0.093 | 0.005 |
| RS-CoCa | 0.059 | 0.081 | 0.039 | 0.014 | 0.007 | 0.088 | 0.121 | 0.001 |
| SAR-CoCa | **0.689** | 0.680 | 0.624 | 0.571 | 0.519 | 0.522 | 0.781 | 2.513 |
| SAR-RS-CoCa | 0.688 | **0.694** | **0.637** | **0.583** | **0.530** | **0.523** | **0.792** | **3.186** |

Table 6. Caption generation results on OSdataset-512-test.

| Model | SPICE | B1 | B2 | B3 | B4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|---|
| CoCa | 0.022 | 0.118 | 0.026 | 0.004 | 0.000 | 0.040 | 0.109 | 0.019 |
| RS-CoCa | 0.054 | 0.095 | 0.187 | 0.091 | 0.053 | 0.035 | 0.132 | 0.205 |
| SAR-CoCa | 0.282 | 0.409 | 0.301 | 0.243 | 0.198 | 0.211 | 0.391 | 0.631 |
| SAR-RS-CoCa | **0.298** | **0.420** | **0.319** | **0.263** | **0.219** | **0.228** | **0.410** | **0.665** |

Table 7. Comparison of training strategies on HRSID-test.

| Model | i2t1 | i2t5 | i2t10 | t2i1 | t2i5 | t2i10 | MR |
|---|---|---|---|---|---|---|---|
| Single-SAR-RS-CLIP | 2.24 | 8.92 | 16.57 | 1.99 | 9.74 | 17.19 | 9.44 |
| SAR-RS-CLIP | **2.65** | **10.91** | **20.50** | **2.55** | **11.78** | **20.75** | **11.52** |

Table 8. Comparison of training strategies on OSdataset-512-test.

| Model | i2t1 | i2t5 | i2t10 | t2i1 | t2i5 | t2i10 | MR |
|---|---|---|---|---|---|---|---|
| Single-SAR-RS-CLIP | 2.83 | 11.32 | 19.34 | 4.01 | 15.80 | 25.71 | 13.17 |
| SAR-RS-CLIP | **5.66** | **16.04** | **23.11** | **5.19** | **20.75** | **28.77** | **16.59** |

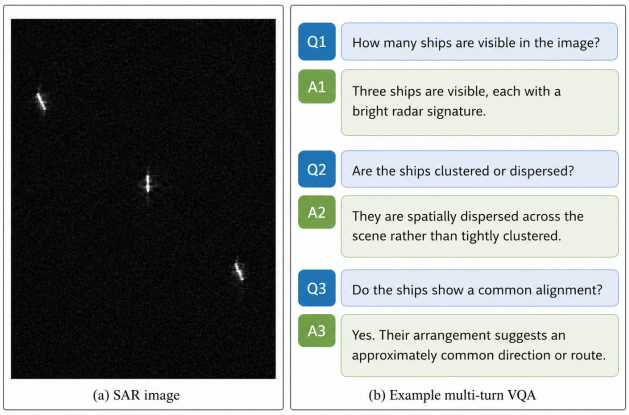


Figure 4. Qualitative example of multi-turn SAR visual question answering with SAR-GPT. Given a SAR image containing several ships, SAR-GPT answers questions about object count, spatial distribution, and approximate alignment, illustrating its ability to support interactive semantic reasoning over SAR imagery.

### 4.5. Generalization Beyond In-Domain Evaluation

To test whether SAR-TEXT transfers beyond benchmarks built with similar annotation procedures, we evaluate on SARLANG-1M-Cap [20]. Because this benchmark does not provide a dedicated retrieval split, the original paper samples 100-image and 500-image subsets for retrieval evaluation and uses the original benchmark for captioning. Fine-tuning on SAR-TEXT substantially improves both retrieval and generation relative to unfine-tuned baselines. The trend is consistent across both retrieval subset sizes, and caption generation metrics improve by roughly twofold after SAR-TEXT fine-tuning. This suggests that SAR-TEXT provides useful supervision that generalizes beyond its own construction pipeline.

This evaluation is important because SAR-TEXT contains captions generated by a specific multi-stage pipeline. A model could in principle overfit to the linguistic style of this pipeline rather than learning transferable SAR–language alignment. Evaluation on SARLANG-1M-Cap provides an additional check against this possibility. Although the benchmark differs in scale and caption style, models fine-tuned on SAR-TEXT still improve over unfine-tuned baselines, suggesting that the learned representations are not limited to the exact templates used during SAR-TEXT construction.

The generalization results also indicate that SAR–language supervision can benefit datasets with different annotation conventions. This is especially relevant for SAR, where existing datasets vary substantially in sensor type, resolution, geographic region, and task definition. A language-aligned representation can provide a common semantic interface across these heterogeneous sources.

### 4.6. Downstream Extension: SAR-GPT for SAR VQA

Beyond retrieval and caption generation, we also examine whether SAR-TEXT can support instruction-style SAR image understanding. To this end, SAR-VQA is constructed from SAR-TEXT and used to fine-tune a TinyGPT-V-based

Table 9. Downstream SAR-VQA results of SAR-GPT on representative test sets.

| Split | Model | SPICE | B4 | METEOR | R-L | CIDEr |
|---|---|---|---|---|---|---|
| HRSID | TinyGPT-V | 0.1789 | 0.0000 | 0.1296 | 0.1975 | 0.2146 |
| | S234-SAR | 0.2959 | 0.1328 | 0.2239 | 0.2975 | 0.9667 |
| | SAR-GPT | **0.3211** | **0.1716** | **0.2350** | **0.3341** | **1.4732** |
| OS-256 | TinyGPT-V | 0.1269 | 0.0215 | 0.1093 | 0.1332 | 0.0483 |
| | S234-SAR | 0.1669 | 0.0730 | 0.1554 | 0.1942 | 0.5513 |
| | SAR-GPT | **0.2974** | **0.2057** | **0.1999** | **0.3276** | **1.7976** |

multimodal model, resulting in SAR-GPT. This evaluation is useful because VQA requires a model to map SAR visual evidence into language responses in a more interactive form than standard captioning.

As shown in Table 9, SAR-GPT consistently outperforms TinyGPT-V and S234-SAR on both HRSID and OSdataset-256 SAR-VQA subsets. The gains are particularly clear in CIDEr and BLEU-4, suggesting that SAR-TEXT-derived instruction tuning improves both semantic grounding and response specificity. These results further indicate that SAR-TEXT is not limited to image–text matching or caption generation, but can also serve as a foundation for downstream SAR-language interaction.

## 5. Discussion Through the Lens of Imperfect Observation

The results highlight several implications for robust multimodal Earth observation under imperfect sensing conditions.

**Missing or unreliable optical observations increase the value of SAR–text alignment.** When optical imagery is unavailable or less reliable, language-aligned SAR representations can preserve semantic access to the scene.

This is particularly relevant in operational settings where optical imagery is not guaranteed. In such cases, SAR is often the only available source, but its interpretation requires domain expertise. By aligning SAR imagery with language, the model can expose SAR content through a semantic interface that is more accessible to downstream retrieval, summarization, and human-in-the-loop analysis.

**Weak pairing can still be useful if semantic transformation is controlled.** The optical-to-SAR rewriting branch shows that imperfect paired supervision can still generate useful language targets when grounded by explicit rules and in-context examples.

This observation is important because perfectly synchronized, manually described SAR–optical pairs are difficult to obtain at scale. The proposed construction process does not require optical captions to be directly copied into SAR descriptions. Instead, it uses optical information as a semantic reference and then filters the description through SAR-aware rewriting rules. This makes weak cross-modal pairing a practical source of supervision rather than a strict requirement.

**Noise matters, but structured noise is learnable.** Human verification reveals that scene-level semantics remain the main failure point, while counts and spatial phrases are more reliable. This indicates that automatic SAR–language supervision is imperfect but not arbitrary; its errors are structured enough to support effective training.

The structured nature of the noise also suggests concrete directions for improvement. Detection-derived captions could be enriched with contextual scene cues, segmentation-derived captions could incorporate object-level priors, and SAR–optical rewriting could be made more conservative when modality-specific evidence is weak. These improvements would further reduce hallucination while retaining the scalability of automatic annotation.

## 6. Limitations and Future Work

This study has several limitations. First, the generated captions are not fully human-authored and may contain style bias, hallucinations, or optical-centric assumptions introduced during rewriting. Second, scene-level semantics remain more error-prone than object counts and coarse spatial descriptions, as also reflected by the human audit. Third, although retrieval, caption generation, and SAR-VQA directly evaluate SAR–language alignment, they do not fully cover all imperfect-observation scenarios. Future work should therefore examine missing-modality reasoning, asynchronous SAR–optical pairing, temporal mismatch, caption-noise robustness, disaster monitoring under cloud cover, and uncertainty-aware multimodal fusion. Finally, while progressive transfer is empirically effective, a more formal analysis of why staged adaptation mitigates the natural-to-SAR modality gap remains an open direction.

## 7. Conclusion

We presented a study of large-scale SAR–language supervision for multimodal Earth observation under imperfect observations. By constructing SAR-TEXT with SAR-Narrator and adapting vision–language models through progressive transfer, we show that SAR–text alignment improves retrieval, caption generation, and downstream SAR visual question answering. Human auditing further shows that the generated supervision is generally usable at scale while revealing remaining bottlenecks in scene-level semantics and grounding. Overall, our findings suggest that language-aligned SAR representations can serve as a practical semantic bridge when multimodal Earth observation is incomplete, heterogeneous, or weakly paired.

# References

[1] Jie Chen, Zhixiang Huang, Runfan Xia, Bocai Wu, Lei Sheng, Long Sun, and Baidong Yao. Large-scale multi-class sar image target detection dataset-1.0. *Journal of Radars*, 14: 1488, 2022. 4

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 5

[3] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 4

[4] Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. 1, 2

[5] Meiyu Huang, Yao Xu, Lixin Qian, Weili Shi, Yaqin Zhang, Wei Bao, Nan Wang, Xuejiao Liu, and Xueshuang Xiang. The qxs-saropt dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:2103.08259*, 2021. 3, 4

[6] Zhongling Huang, Zongxu Pan, and Bin Lei. What, where, and how to transfer in sar target recognition based on deep cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2324–2336, 2019. 1, 3, 5

[7] Pallavi Jain, Dino Ienco, Roberto Interdonato, Tristan Berchoux, and Diego Marcos. Senclip: Enhancing zero-shot land-use mapping for sentinel-2 with ground-level prompting. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5656–5665. IEEE, 2025. 2

[8] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102638, 2022. 3, 4

[9] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023. 2

[10] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 4

[11] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 1, 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[13] Michael Schmitt, Lloyd Haydn Hughes, and Xiao Xiang Zhu. The sen1-2 dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:1807.01569*, 2018. 3, 4

[14] Arsenios Tsokas, Maciej Rysz, Panos M Pardalos, and Kathleen Dipple. Sar data applications in earth observation: An overview. *Expert Systems with Applications*, 205:117342, 2022. 1

[15] Yuanyuan Wang and Xiao Xiang Zhu. The sarptical dataset for joint analysis of sar and optical image in dense urban area. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6840–6843. IEEE, 2018. 3, 4

[16] Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. A sar dataset of ship detection for deep learning under complex backgrounds. *remote sensing*, 11(7):765, 2019. 4

[17] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 3

[18] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 1, 2

[19] Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8: 120234–120254, 2020. 4, 5

[20] Yimin Wei, Aoran Xiao, Yexian Ren, Yuting Zhu, Hongruixuan Chen, Junshi Xia, and Naoto Yokoya. Sarlang-1m: A benchmark for vision-language modeling in sar image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 2026. 1, 3, 7

[21] Yuming Xiang, Rongshu Tao, Feng Wang, Hongjian You, and Bing Han. Automatic registration of optical and sar images via improved phase congruency model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5847–5861, 2020. 3

[22] Kelu Yao, Nuo Xu, Rong Yang, Yingying Xu, Zhuoyan Gao, Titinunt Kitrungrotsakul, Yi Ren, Pu Zhang, Jin Wang, Ning Wei, et al. Falcon: A remote sensing vision-language foundation model. *arXiv preprint arXiv:2503.11070*, 2025. 1, 2

[23] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 5

[24] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions.(2010). *URL http://www.phash. org/docs/pubs/thesis_zauner. pdf*, 2010. 4

[25] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction

tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025. 1, 2

[26] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021. 4

[27] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. 2

[28] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2(8), 2023. 3

[29] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1, 2, 3

[30] Kai Zhao and Wei Xiong. Exploring data and models in sar ship image captioning. *IEEE Access*, 10:91150–91159, 2022. 1, 3

[31] Xiaobing Zhou, Ni-Bin Chang, and Shusun Li. Applications of sar interferometry in earth and environmental science research. *Sensors*, 9(3):1876–1912, 2009. 1

[32] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 1, 3, 5