SARFACT: BENCHMARKING TASK ANSWERS AND VISUAL EVIDENCE FOR SAR VISION–LANGUAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

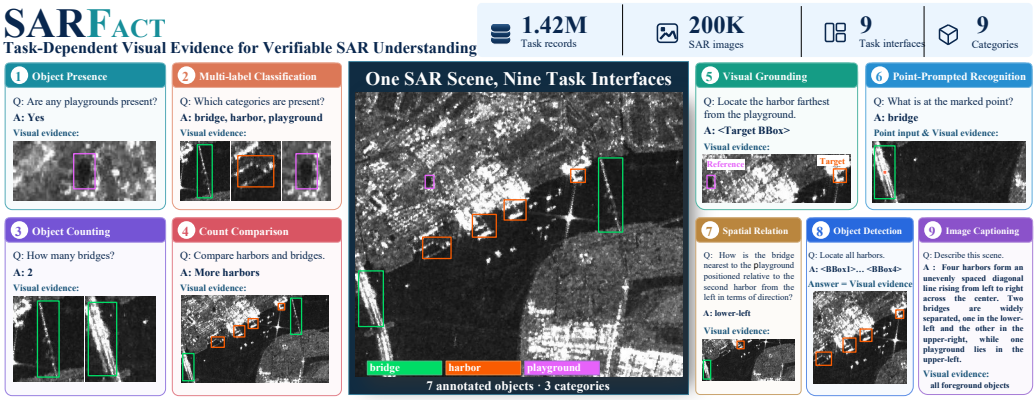


Figure 1: **Overview of SARFACT.** Built from four public SAR datasets, SARFACT provides 1.42M examples across nine task interfaces. The same SAR scene gives rise to different Task Answers and task-dependent Visual Evidence, with the required localized support determined by each task.

ABSTRACT

Evaluating SAR vision–language models only by the correctness of their task outputs leaves a verification gap: a model may produce the expected answer without identifying the visual support needed to inspect it. We introduce SARFACT, a large-scale object-centric SAR vision–language dataset and benchmark that preserves each task's native Task Answer and defines task-specific Visual Evidence: the localized object support needed to inspect the answer. Built from 200K SAR images and 1.11M foreground objects, SARFACT contains 1.42M examples across nine task interfaces. Task Answers and Visual Evidence are derived from shared localized object facts, while task-specific construction controls reduce shortcut-prone patterns and preserve informative evaluation cases. SARFACT separately evaluates Answer, Evidence, and Joint correctness. Experiments with six public VLMs reveal limited zero-shot performance and frequent cases in which correct Answers fail to satisfy the task-defined Visual Evidence requirements. Explicit Evidence supervision substantially improves joint Answer–Evidence performance. Annotations, task records, and evaluation code will be publicly released.

1 INTRODUCTION

Synthetic aperture radar (SAR) provides day-and-night, weather-resilient observations, but interpreting its backscatter patterns often requires domain expertise. Vision–language models (VLMs) offer a natural-language interface for querying and describing SAR scenes. Recent datasets have substantially expanded this capability: SARLANG-1M broadens geographic and semantic coverage through captioning and question answering, while SARChat-Bench-2M supports large-scale multi-task dialogue and incorporates object coordinates into counting and localization-related tasks (Wei

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

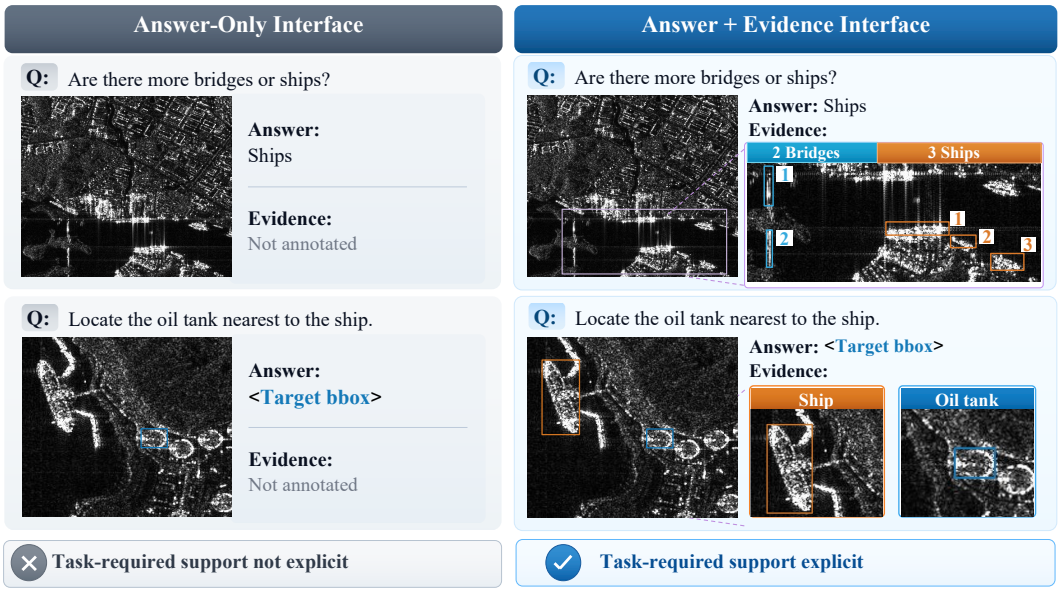


Figure 2: **Task Answers and task-required Visual Evidence serve different roles.** For Count Comparison, the answer identifies which category is more numerous, while Visual Evidence localizes the objects supporting both counts. For visual grounding, the target box is the answer, while Visual Evidence additionally identifies the reference object needed to inspect the queried relation.

et al., 2025; Ma et al., 2025). Yet benchmark success is still primarily determined by the requested task output, leaving an important question unresolved: can the model also identify the visual evidence needed to inspect its answer?

A task output does not necessarily provide such evidence, even when the output itself is localized. Figure 2 illustrates this distinction. For the comparison question "Are there more bridges or ships?", "More ships" is a complete semantic answer, but checking it requires the bridge and ship instances underlying the two counts. For "Locate the oil tank nearest to the ship," the target box is a complete grounding answer, yet inspecting the stated relation also requires the reference ship. Localization can therefore be part of the evidence without being sufficient by itself. What constitutes valid Visual Evidence depends on the task being answered.

We introduce SARFACT, a dataset and benchmark for evaluating Task Answers together with task-dependent Visual Evidence. For each task, SARFACT preserves the native Task Answer while specifying the localized object support required for inspection. Depending on the task, this support may be a valid witness, a complete object set, or objects with specific roles. Object Detection is the special case in which the localization answer itself satisfies the Evidence requirement.

SARFACT is built from a source-aware object pool containing 200,363 SAR images and 1,108,183 foreground objects from four public datasets, from which 1.42M examples across nine task interfaces are constructed. Source annotations are aligned into shared localized object facts while retaining annotation scope, providing a consistent factual basis for Task Answers and Visual Evidence across tasks. Task-specific construction then controls shortcut-prone answer, category, and composition patterns in held-out evaluation while preserving rare and informative cases. Image Captioning extends the same object facts to free-form scene descriptions.

Evaluation separates Answer, Evidence, and Joint correctness. Answers are assessed according to each task, while Evidence evaluates whether the required localized support is correctly provided; Joint correctness requires both on the same example. Across six public VLMs, zero-shot performance remains limited, and correct Answers frequently fail to satisfy the task-defined Visual Evidence requirements. Supervised experiments further show that explicit Evidence supervision substantially improves joint Answer–Evidence performance, while cross-benchmark evaluation examines whether learned task capabilities transfer beyond SARFACT.

Our contributions are threefold:

1. **A large-scale SAR vision–language benchmark for Task Answers and Visual Evidence.** SARFACT contains 1.42M examples across nine task interfaces, built from 200K SAR images and 1.1M localized foreground objects using shared object facts and task-specific construction controls.
2. **A task-dependent formulation of Visual Evidence.** SARFACT defines the localized support required by each task from shared object facts, including valid witnesses, complete object sets, and objects with specific roles, while preserving the native Task Answer.
3. **An empirical study of verifiable SAR understanding.** Experiments evaluate zero-shot performance, Evidence-aware supervision, and cross-benchmark transfer.

2 Related Work

2.1 SAR Vision–Language Datasets and Benchmarks

SAR vision–language resources differ in both semantic coverage and the prediction targets derived from their annotations. SAR-TEXT provides image–text pairs whose utility is studied in retrieval, captioning, and downstream VQA, while SARLANG-1M supplies captions and multi-task questions across geographically and semantically diverse imagery (He et al., 2025; Wei et al., 2025). SAREval complements these resources with SAR-specific assessment of perception, reasoning, and robustness (Wang et al., 2026).

Counting illustrates how source annotations can yield different supervision targets. SARLANG formulates instance counting with a numerical answer, whereas SARChat-Bench-2M retains instance coordinates alongside counts and evaluates both counting accuracy and localization (Wei et al., 2025; Ma et al., 2025). Both datasets use detection annotations to construct counting examples; SARChat additionally makes recovery of the counted instances an explicit target. For a dataset spanning multiple tasks, this raises the design question of how to specify supporting objects when the required coverage and object roles vary with the task.

2.2 Grounded Datasets and Evidence-Aware Evaluation

VizWiz-VQA-Grounding annotates the visual support for answers to questions asked by people with visual impairments (Chen et al., 2022). GQA generates compositional questions from scene graphs and uses functional programs to control conditional answer distributions and support diagnostic evaluation; for attention-based models, its grounding score measures attention over question-relevant regions (Hudson & Manning, 2019). CRIC combines scene and knowledge graphs to construct vision-and-commonsense questions and evaluates explicit object-grounding decisions jointly with answers, requiring both to be correct (Gao et al., 2023).

Remote-sensing benchmarks also associate task answers with spatial support. Chessboard derives question–answer pairs and question-relevant image cells from land-cover annotations, scoring answers and cells separately (Tosato et al., 2025). TerraScope-Bench derives geospatial answers from segmentation annotations and evaluates answer accuracy alongside mask quality, including under SAR inputs (Shu et al., 2026). Grounding can also be attached to explanation steps: Visual Reasoning Tracer (VRT) aligns intermediate and final objects with masks and measures required-object coverage and localization quality (Yuan et al., 2025), while VG-CoT provides bounding-box-linked reasoning annotations and evaluates explanation quality, grounding, and reasoning–answer alignment (Lim et al., 2026).

SARFACT brings task-specific evidence requirements and controls over example construction into a shared SAR resource spanning nine vision–language tasks. It derives native Task Answers and Visual Evidence from the same source-aware object facts, including for tasks whose answers are already localizations. Positive Object Presence admits a valid witness, Object Counting requires the complete counted-instance set, and Spatial Relation specifies subject and reference roles. Task-specific selection controls limit dominant answer and composition patterns while retaining variation in cardinality and spatial structure. Together, these design choices support Answer-only and Answer–Evidence studies that distinguish task-answer correctness from recovery of the objects required to support it.

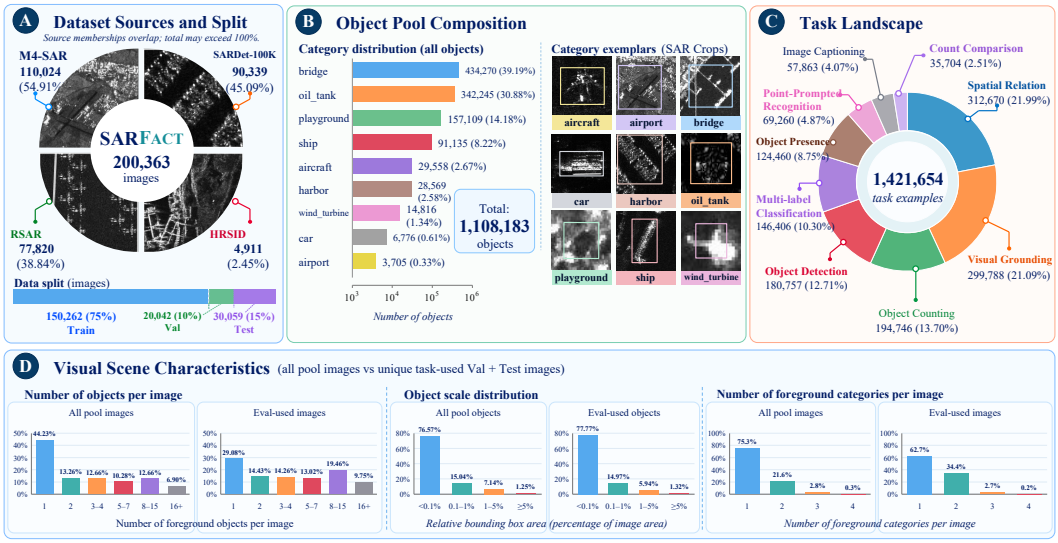


Figure 3: **Overview of SARFACT.** It shows (A) the dataset sources and train/validation/test split, (B) the category composition of the unified object pool with representative SAR crops, (C) the distribution of examples across the nine task interfaces, and (D) object-count, target-scale, and category-composition characteristics of the full dataset and held-out evaluation subset.

3 THE SARFACT DATASET

3.1 DATASET OVERVIEW AND COMPARISON

Dataset overview. SARFACT is a unified object-centric SAR vision–language benchmark built upon object annotations from four established SAR datasets: M4-SAR, SARDet-100K, RSAR, and HRSID (Wang et al., 2025a; Li et al., 2024; Zhang et al., 2025; Wei et al., 2020). It contains 200,363 SAR images and 1,108,183 foreground objects from nine categories, supporting nine vision–language tasks and 1,421,654 task examples. As summarized in Fig. 3, the benchmark spans diverse object categories, scene densities, target scales, and task interfaces.

Comparison with existing benchmarks. SARLANG-1M (Wei et al., 2025) emphasizes broad semantic coverage, while SARChat-Bench-2M (Ma et al., 2025) provides large-scale multi-task dialogue supervision. As summarized in Table 1, SARFACT takes a complementary direction toward verifiable and diagnostic object-centric evaluation, with three main differences.

1. SARFACT *makes Visual Evidence an explicit evaluation target*, separating the Task Answer from task-defined localized support and enabling separate Answer, Evidence, and Joint evaluation. Task Answer and Visual Evidence may coincide, as in Object Detection, or differ according to task semantics.

2. SARFACT *applies task-specific controls to improve the diagnostic value of tasks shared with existing benchmarks.* In the respective Test subsets, count-one cases comprise 20.3% of SARFACT Counting examples versus 64.0% in the evaluated SARLANG subset; zero-count cases account for 10.0% versus none; and cases with at least eight objects account for 26.4% versus 3.3%. These distributions indicate broader cardinality coverage and lower count-one concentration in SARFACT. Task-specific controls for the remaining tasks are summarized in Sec. 3.3.

3. *Captioning emphasizes fact-linked foreground coverage.* Whereas SARLANG provides broad free-form scene descriptions, SARFACT retains free-form language while requiring coverage of supported foreground content and using the corresponding objects as Visual Evidence. Illustrative reference captions and object-support visualizations are provided in Supplement Sec. S.4.1, Fig. 5.

Table 1: Comparison of SARFACT with representative large-scale SAR vision–language datasets in terms of scale, task coverage, output formulation, and evaluation protocol.

Dimension	SARLANG-1M	SARChat-Bench-2M	SARFACT
Images	118.3K	N/R	200.4K
Records	1.13M	2.06M	1.42M
Object Categories	1,696*	6	9
Common Task Families	Cnt.; Cls.; ImgCap.; BPR	Cnt.; Cls.; ImgCap.; BPR	Cnt.; ML-Cls.; ImgCap.; PPR
Additional Tasks	Pres.; OP	VG; SG	Pres.; VG; Det.; CC; SR
Output	Task Answer	Task Answer; boxes in selected tasks†	Task Answer and task-defined Visual Evidence
Evaluation	Task-specific metrics	Task-specific metrics; localization where applicable	Separate Answer / Evidence / Joint

SARFACT task abbreviations are defined in Table 2. Additional abbreviations: Cls., Classification; BPR, Box-Prompted Recognition; OP, Object Positioning; SG, Spatial Grounding; N/R, not reported.

* SARLANG-1M's 1,696 object categories follow its native vocabulary and are not directly comparable in ontology granularity with SARFACT's nine localized foreground categories.

† SARChat outputs boxes for Cnt., SG, and VG; box outputs do not necessarily constitute task-defined Visual Evidence.

Table 2: Overview of the nine tasks supported by SARFACT, with task answers and task-specific visual evidence.

Task	Task Answer	Visual Evidence
Object Detection (Det.)	All queried-category boxes	Same box set
Object Presence (Pres.)	Yes / No	Yes: ≥ 1 valid box; No: no positive box
Multi-label Classification (ML-Cls.)	All supported present categories	≥ 1 valid box per ground-truth positive category
Object Counting (Cnt.)	Integer count	All counted-object boxes; none if zero
Point-Prompted Recognition (PPR)	Target category	Target box
Spatial Relation (SR)	8-way relative direction	Ordered subject/reference boxes
Count Comparison (CC)	Category with the larger count, or `equal`	All boxes for both categories with first/second roles
Visual Grounding (VG)	Target box	Target + query-required supporting boxes
Image Captioning (ImgCap.)	Coverage-constrained caption	All supported foreground-object boxes

3.2 SUPPORTED TASKS AND VISUAL EVIDENCE

SARFACT supports nine tasks spanning object recognition and localization, quantitative and spatial reasoning, and scene-level description. For each task, it specifies the requested Task Answer and the localized object support required as Visual Evidence. These definitions rely on source-supported object facts rather than assumed physical completeness. Evidence requirements vary by task, from a valid witness to a complete object set or role-specific objects. For example, positive Object Presence requires one valid witness, whereas Object Counting requires all queried-category instances. Object Detection is the special case where Answer and Evidence coincide, while Spatial Relation and Visual Grounding may require additional supporting objects. Thus, a correct Task Answer does not necessarily satisfy the Visual Evidence requirement. Table 2 summarizes the task-wise definitions, with detailed rules in Supplement Sec. S.1.

Table 3: Principal task-specific controls applied to Validation and Test example construction to reduce shortcut-prone patterns while preserving informative cases.

Task	Principal Controls
Object Detection	Preserve variation in target count, object scale, and scene density, including small and crowded cases.
Object Presence	Certified absence; Yes/No balance within each source–query-category condition; challenging co-occurrence negatives.
Multi-label Classification	Suppress dominant label sets and strong category co-occurrences while retaining multi-category scenes and rare combinations.
Object Counting	Cover count ranges 1, 2–3, 4–7, 8–15, and 16+, with each range containing at most 30% of positive examples; certified zeros comprise about 10% of evaluation.
Point-Prompted Recognition	Require uniquely identifying points and same-image counterfactuals with at least two answer categories per retained image.
Spatial Relation	Retain unambiguous 8-way relations; limit dense-scene pair contributions while preserving pair diversity.
Count Comparison	Balance operand order; constrain pair-conditioned winner bias; retain equal cases while limiting trivial ties; control normalized count margins among non-equal cases.
Visual Grounding	Require at least two same-category candidates and a unique target; limit target reuse; vary directional, ordinal, size, and anchor-relative cues.
Image Captioning	Require multiple foreground categories or at least three objects from one category; retain scenes with meaningful spatial structure.

3.3 Dataset Construction

Reliable Object Fact Construction. Source annotations are aligned into a common representation by mapping category names to a unified namespace and converting object geometry to horizontal bounding boxes. Source-specific annotation scope is retained to distinguish certified absence from unannotated categories. For visual content shared across sources, duplicate annotations are identified and checked for consistency; conflicting cases are discarded, while consistent annotations are deduplicated and retained as canonical object facts.

Benchmark Splits. To prevent related visual content from leaking across splits, the object pool is partitioned before task generation. Images linked by duplicate content, shared parent scenes, crop lineage, or cross-source correspondence are grouped and assigned to the same split rather than inheriting source-native split labels. A 75%/10%/15% split yields 150,262 Train, 20,042 Validation, and 30,059 Test images.

Task-Specific Example Construction. Shared object facts provide the factual basis for Task Answers and Visual Evidence across tasks. Task candidates are first constructed under task-specific validity requirements so that the query, Answer, and required Visual Evidence are well defined. Validation and Test examples are then selected with task-specific controls that limit dominant answer, category, and composition patterns while retaining rare, difficult, and compositionally rich cases. Table 3 summarizes the principal controls, with further construction details provided in Supplement Sec. S.2.

Natural-Language Query Realization. Each task uses 10 AI-generated query templates that are manually reviewed for clear intent, consistent task semantics, and no cross-task ambiguity. All 90 templates are provided in Supplement Sec. S.4.

Caption Realization. Image captions are generated under object-fact constraints. GPT-5.6 produces the initial complete set from SAR images and structured object facts, while Astra regenerates a subset from annotated object geometry and the same facts. The Astra-regenerated Validation and Test captions serve as the references used in the reported caption experiments.

3.4 Quality Assurance

SARFact undergoes full-corpus programmatic checks for task semantics, Answer–Evidence consistency, cross-task fact consistency, and annotation conflicts. Following reviewer-level quality control, the human audit of Astra-regenerated captions across Train, Validation, and Test retains 500 judgments from 10 reviewers, covering 429 distinct captions. The judgment-level pass rate is 89.60% (448/500), with 87.32% pass/non-pass agreement (62/71) among the 71 captions with two retained reviews. These descriptive results are not estimates of corpus-wide quality. Detailed quality-assurance protocols are provided in Supplement Sec. S.3, with human caption audit details in Sec. S.3.4.

3.5 Evaluation Protocol

Answer evaluation. Object Detection uses GRIT Localization Accuracy (Gupta et al., 2022), which accounts for localization quality, missed objects, and extra predictions. Visual Grounding uses sample-level localization accuracy (Sun et al., 2023), with an IoU threshold of 0.25. Multi-label Classification is evaluated within each sample's source-supported category scope using example-level Jaccard (Godbole & Sarawagi, 2004; Tsoumakas & Katakis, 2007); predictions outside this scope are ignored, while Joint correctness requires an exact category-set match within the same scope. Image Captioning uses CIDEr-D (Vedantam et al., 2015), Core Fact F1, and Core FER. Core Fact F1 measures agreement with annotated category-presence and explicit count facts, while Core FER is the proportion of captions containing at least one explicit contradiction in these facts. These Core metrics evaluate category and count facts rather than all spatial descriptions in a caption. CIDEr-D uses the Astra Test references, while Core Fact F1 and Core FER are computed directly from the annotated category and count facts. The remaining structured tasks use exact accuracy. Macro-8 reports the equal-weight average across all eight structured tasks. Detailed metric definitions are provided in Supplement Sec. S.5.

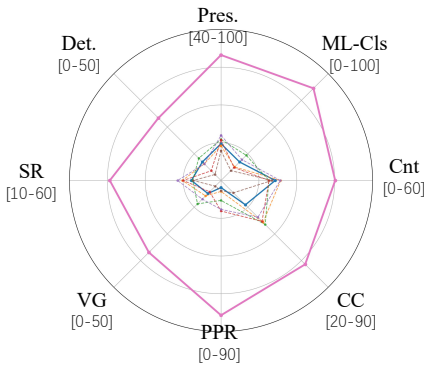
Evidence-aware evaluation. Evidence evaluation follows the task-specific requirements in Table 2. Positive Object Presence requires any valid witness, whereas Object Counting and Count Comparison require complete relevant object sets. Predicted and ground-truth evidence boxes are matched one-to-one by maximizing total IoU, with category and, where applicable, role consistency enforced; matches require IoU≥ 0.25. For Multi-label Classification, boxes outside the evaluable category scope are excluded before matching. Evidence F1 penalizes missing required support and unsupported predictions, while Evidence Exact requires the task-specific Evidence requirement to be fully satisfied with no unsupported predictions. Negative Object Presence and zero-count Object Counting are included in Answer evaluation but excluded from Evidence and Joint evaluation to avoid vacuous empty-set credit. Joint Accuracy requires both the task-specific Answer criterion and Evidence Exact on the same sample. For PPR, a valid single-box Evidence match also requires the predicted category to be correct, so Evidence F1 and Joint Accuracy coincide.

4 Experiments

4.1 Experimental Setup

Six publicly available VLMs are evaluated: Qwen3.5-4B, Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025b), and LLaVA-OneVision-2-8B-Instruct (An et al., 2026) as general-purpose models, and RSCoVLM-7B-2512 (Li et al., 2026) and GeoChat-7B (Kuckreja et al., 2024) as remote-sensing models. Results are reported on the SARFact Test split; Answer-only zero-shot Validation results are provided in Supplement Sec. S.7.1.

Supervised fine-tuning. Supervised experiments use Qwen3.5-4B with an independent LoRA adapter for each task and output setting, so the Answer-only and Answer–Evidence settings are trained separately. For paired examples, the images, base queries, and Answer targets are the same; the Answer–Evidence setting additionally includes Visual Evidence supervision and the corresponding output format. The backbone, visual encoder, and visual merger remain frozen. Checkpoints are selected on the Validation split and subsequently evaluated on Test. Training details are provided in Supplement Sec. S.6.



Model	Macro-8 ↑	Core F1 ↑	Core FER ↓	CIDEr-D ↑
Zero-shot setting:				
Qwen3.5-4B	21.02	0.12	0.12	0.55
Qwen3.5-9B	22.72	0.07	**0.09**	0.20
InternVL3.5-8B	25.03	0.88	1.15	1.29
LLaVA-OV2-8B	23.06	0.34	0.25	1.23
RSCoVLM-7B	**25.38**	4.45	3.55	**1.94**
GeoChat-7B	17.51	**5.69**	5.66	0.05
Supervised Finetuning:				
Qwen3.5-4B	60.77	60.80	65.22	43.47

Figure 4: **Answer-only performance on the SARFACT Test set.** Left: task-wise profiles over eight structured tasks, using the task-specific display ranges shown on the axes. Right: Macro-8 and Image Captioning results; CIDEr-D is scaled by 100, while other scores are percentages. Supervised results use task-specific adapters.

4.2 Answer-only Performance

Zero-shot benchmark. Figure 4 summarizes Answer-only performance on the SARFACT Test set. All six VLMs show limited zero-shot performance, with the best Macro-8 reaching only 25.38%. Among the evaluated models, neither increased model scale nor remote-sensing specialization yields consistent gains: Qwen3.5-9B improves only modestly over Qwen3.5-4B, while the two remote-sensing models span the highest and lowest Macro-8 scores. Zero-shot captioning is similarly weak, with Core Fact F1 reaching at most 5.69%.

Supervised learnability. Fine-tuning the same Qwen3.5-4B backbone substantially improves performance, increasing Macro-8 from 21.02% to 60.77%. Large gains occur across the task profile, including Point-Prompted Recognition from 4.22% to 80.30% and Multi-label Classification from 17.32% to 86.26%. These results show that the low zero-shot scores do not arise from an inherently unlearnable task formulation.

Captioning diagnostics. Supervision also increases Core Fact F1 from 0.12% to 60.80% and CIDEr-D from 0.55 to 43.47. Core FER, however, rises to 65.22%. The very low zero-shot FER should not be interpreted as strong factual captioning when Core Fact F1 is also near zero; after supervision, captions contain substantially more correct core facts, but explicit contradictions remain common.

4.3 Answer–Evidence Evaluation

Table 4 examines two questions: how requiring Visual Evidence affects task Answers, and whether explicit Evidence supervision improves joint Answer–Evidence correctness.

Zero-shot Answer–Evidence behavior. Requesting Visual Evidence changes Answer performance in both directions: Counting decreases from 21.46% to 10.84%, whereas Point-Prompted Recognition increases from 4.22% to 15.59%. More importantly, a correct Answer does not imply that the task-defined Visual Evidence requirement is satisfied. Among zero-shot Answer–Evidence outputs, only about 19.5% of correct PPR answers and 5.4% of correct Spatial Relation answers also satisfy Evidence Exact. This directly exposes the verification gap that Answer-only evaluation cannot capture.

Effect of Evidence supervision. Relative to Answer-only SFT, Answer scores increase on four of the five reported tasks, while Counting decreases modestly from 45.20% to 43.58%. Relative to zero-shot Answer–Evidence generation, Joint Accuracy increases substantially across all reported tasks, including from 1.24% to 53.16% for Presence and from 3.04% to 79.19% for PPR. Spatial Relation remains challenging: Answer accuracy reaches 50.70%, while Joint Accuracy reaches only 20.33%. Together, these results show that explicit Evidence supervision substantially improves joint Answer–Evidence performance while preserving strong Answer performance, although satisfying the task-defined Evidence requirements remains difficult for some interfaces.

Table 4: Zero-shot and supervised fine-tuning performance of Qwen3.5-4B on the SARFACT Test set (%). A: task-native Answer score; E: Evidence F1; J: Joint Accuracy. A+E denotes the Answer–Evidence setting. Evidence and Joint follow the task-specific evaluation rules in Sec. 3.5 and use IoU≥ 0.25.

Task	Zero-shot A	Zero-shot A+E	SFT A	SFT A+E
	A↑	A↑ / E↑ / J↑	A↑	A↑ / E↑ / J↑
Pres.	54.61	51.62 / 1.26 / 1.24	89.83	91.21 / 53.16 / 53.16
ML-Cls.	17.32	10.45 / 4.36 / 3.15	86.26	86.44 / 44.58 / 38.48
Cnt.	21.46	10.84 / 0.89 / 0.49	45.20	43.58 / 39.77 / 23.57
PPR	4.22	15.59 / 3.04 / 3.04	80.30	83.63 / 79.19 / 79.19
SR	19.77	23.79 / 4.79 / 1.29	46.70	50.70 / 30.86 / 20.33

Table 5: Accuracy (%) on task-aligned SARLANG subsets before and after SARFACT supervision; N denotes subset size and Δ the gain in percentage points.

SARLANG Task	N	Zero-shot	SARFACT-SFT	Δ
Presence	20,987	49.27	87.62	+38.35
Counting	10,585	51.43	80.29	+28.86
Classification	10,492	33.46	98.54	+65.08

4.4 Cross-Benchmark Evaluation

Cross-benchmark setting. Task-aligned subsets are extracted from the official SARLANG Test split based on question semantics and answer type. Qwen3.5-4B is evaluated on the same samples before and after loading the corresponding SARFACT-trained Answer-only adapter, under identical inference settings and with no SARLANG training or checkpoint selection. Presence and Counting use accuracy, while Classification uses exact matching of the normalized category set. Because SARFACT and SARLANG partially share upstream SARDet provenance, this experiment measures cross-benchmark transfer rather than strict unseen-image generalization.

Cross-benchmark results. Table 5 shows substantial gains across all three aligned tasks after loading the SARFACT-trained adapters, indicating that the learned task capabilities transfer to the corresponding SARLANG interfaces.

5 Limitations and Broader Impact

Limitations. SARFACT repurposes public SAR detection datasets and therefore inherits their annotation scope and boundaries. Its nine-category vocabulary covers only a subset of SAR applications; source-scoped annotations do not guarantee physical completeness, horizontal boxes discard object orientation, and a single reference caption cannot represent all valid scene descriptions. Public source images may also have appeared in foundation-model pretraining, which benchmark splitting cannot rule out.

Broader Impact. SARFACT evaluates not only whether a task answer is correct, but also whether it is accompanied by inspectable localized support. This can support failure analysis and model auditing, but Visual Evidence should not be interpreted as a causal explanation of model reasoning or a guarantee of deployment reliability. SARFACT is intended as an evaluation resource rather than a certification of model trustworthiness.

6 Conclusion

We present SARFACT, an object-centric SAR vision–language benchmark covering nine tasks built from shared localized object facts, with task-specific Visual Evidence defined for each task. Its task-wise construction further controls diversity, difficulty, and shortcut cues to improve the diagnostic value of evaluation examples. SARFACT enables systematic evaluation of both Task Answers and Visual Evidence, while experiments demonstrate its ability to expose model limitations and support supervised learning and transfer.

AI USE STATEMENT

In this work, generative AI tools were used to assist dataset construction, literature review, and manuscript preparation. Specifically, generative AI was used to generate candidate query templates for the benchmark tasks. The authors manually reviewed all candidate templates for semantic correctness, consistency with the predefined task specifications, and ambiguity, and only verified templates were retained for dataset construction. Generative AI was also used to generate image captions under author-defined prompts and object-fact constraints. The authors designed and controlled this process and manually audited a subset of the generated captions for factual correctness and adherence to the annotation criteria.

Generative AI tools also assisted code implementation. They further assisted literature identification and synthesis, drafting and revising manuscript text, and checking consistency between claims, task definitions, and reported results. The authors reviewed and revised the AI-assisted text and checked the cited sources. The authors take responsibility for the final dataset, code, experimental claims, and manuscript, including artifacts and text produced with the assistance of generative AI.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19098–19107, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Grounding_Answers_for_Visual_Questions_Asked_by_Visually_Impaired_People_CVPR_2022_paper.html.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. CRIC: A VQA dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5561–5578, 2023. doi: 10.1109/TPAMI.2022.3210780.

Shantanu Godbole and Sunita Sarawagi. *Discriminative Methods for Multi-labeled Classification*, pp. 22–30. Springer Berlin Heidelberg, 2004. ISBN 9783540247753. doi: 10.1007/978-3-540-24775-3_5. URL http://dx.doi.org/10.1007/978-3-540-24775-3_5.

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. GRIT: General robust image task benchmark, 2022. URL https://arxiv.org/abs/2204.13653.

Yiguo He, Xinjun Cheng, Junjie Zhu, Chunping Qiu, Jun Wang, Xichuan Zhang, Qiangjuan Huang, and Ke Yang. SAR-TEXT: A large-scale SAR image-text dataset built with SAR-Narrator and a progressive learning strategy for downstream tasks. *arXiv preprint arXiv:2507.18743*, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. GeoChat: Grounded large vision-language model for remote sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Qingyun Li, Shuran Ma, Junwei Luo, Yi Yu, Yue Zhou, Fengxiang Wang, Xudong Lu, Xiaoxing Wang, Xin He, Yushi Chen, and Xue Yang. Co-training vision-language models for remote sensing multi-task learning. *Remote Sensing*, 18(2), 2026. ISSN 2072-4292. doi: 10.3390/rs18020222. URL https://www.mdpi.com/2072-4292/18/2/222.

Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. SARDet-100K: Towards open-source benchmark and toolkit for large-scale SAR object detection. In *NeurIPS*, 2024.

Byeonggeuk Lim, Kyeonghyun Kim, JungMin Yun, and YoungBin Kim. VG-CoT: Towards trustworthy visual reasoning via grounded chain-of-thought. *arXiv preprint arXiv:2604.21396*, 2026.

Zhiming Ma, Xiayang Xiao, Sihao Dong, Peidong Wang, Haipeng Wang, and Qingyun Pan. SARChat-Bench-2M: A multi-task vision-language benchmark for SAR image interpretation. *arXiv preprint arXiv:2502.08168*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Yan Shu, Bin Ren, Zhitong Xiong, Xiao Xiang Zhu, Begüm Demir, Nicu Sebe, and Paolo Rota. TerraScope: Pixel-grounded visual reasoning for earth observation. *arXiv preprint arXiv:2603.19039*, 2026.

Jiamu Sun, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. RefTeacher: A strong baseline for semi-supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19144–19154, June 2023.

Lucrezia Tosato, Christel Tartini Chappuis, Syrielle Montariol, Flora Weissgerber, Sylvain Lobry, and Devis Tuia. Checkmate: Interpretable and explainable RSVQA is the endgame. *arXiv preprint arXiv:2508.13086*, 2025.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, July 2007. ISSN 1548-3932. doi: 10.4018/jdwm.2007070101. URL http://dx.doi.org/10.4018/jdwm.2007070101.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Chao Wang, Wei Lu, Xiang Li, Jian Yang, and Lei Luo. M4-SAR: A multi-resolution, multi-polarization, multi-scene, multi-source dataset and benchmark for optical-SAR fusion object detection. *arXiv preprint arXiv:2505.10931*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Ziyan Wang, Lei Liu, Gang Wan, Yuchen Lu, Fengjie Zheng, Guangde Sun, Yixiang Huang, Shihao Guo, Xinyi Li, and Liang Yuan. SAREval: A multi-dimensional and multi-task benchmark for evaluating visual language models on SAR image understanding. *Remote Sensing*, 18(1):82, 2026. doi: 10.3390/rs18010082.

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. doi: 10.1109/ACCESS.2020.3005861.

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. *arXiv preprint arXiv:2504.03254*, 2025.

Haobo Yuan, Yueyi Sun, Yanwei Li, Tao Zhang, Xueqing Deng, Henghui Ding, Lu Qi, Anran Wang, Xiangtai Li, and Ming-Hsuan Yang. Visual reasoning tracer: Object-level grounded reasoning benchmark. *arXiv preprint arXiv:2512.05091*, 2025.

Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, and Xiang Li. RSAR: Restricted state angle resolver and rotated SAR benchmark. In *CVPR*, pp. 7416–7426, 2025.