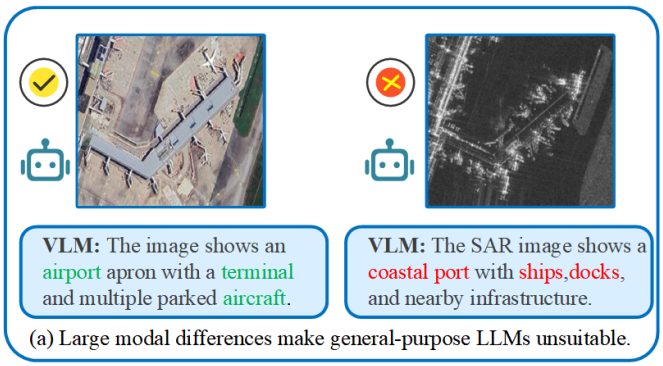
FUSAR-GPT : A Spatiotemporal Feature-Embedded and Two-Stage Decoupled Visual Language Model for SAR Imagery

Supplementary Material

7. The specificity of SAR images

As illustrated in Fig. 6, SAR imagery exhibits inherent limitations that constrain visual–semantic understanding. First, the large modality gap between optical and SAR images leads to systematic misinterpretation: optical models rely on color, texture, and shading, whereas SAR captures microwave backscatter dominated by geometric and dielectric responses. Consequently, general-purpose VLMs often map SAR scattering patterns to incorrect optical concepts, such as hallucinating roads or vehicles. Second, SAR imagery is intrinsically sparse—strong scatterers form a small number of saturated bright points, while extensive dark regions contain weak but semantically relevant signals that lack explicit structural cues. This highly polarized intensity distribution causes model attention to collapse onto a few bright pixels, suppressing contextual information embedded in the dark background. Together, these modality discrepancies and sparsity characteristics highlight the fundamental limitations of applying standard VLMs to SAR imagery.



(a) Large modal differences make general-purpose LLMs unsuitable.

(b) Target features in SAR images are sparse.

Figure 6. Challenges in developing SAR visual language models.

8. Additional Experiments

Table 5 presents an extended comparison on three SAR understanding tasks with additional baselines. FUSAR-GPT consistently outperforms all competing methods by a significant margin across counting, grid-based localization, and classification tasks. In particular, it achieves substantial improvements in counting accuracy and grid prediction , highlighting its strong spatial reasoning capability. Compared with recent remote sensing–oriented models such as SAR-CLIP, SAR-JEPA, and SkyCLIP, our method demonstrates superior cross-task generalization, indicating that the combination of SFT1 pre-alignment and TLM-based feature fusion enables a more unified and robust representation for diverse SAR understanding tasks.

Table 5. Performance comparison on SAR understanding tasks.B and L denote ViT-Base and ViT-Large backbones, respectively.

Model	Count	Grid			Classification
	@Acc	Acc@100	Acc@50	Top1	@Acc
BAN-L[16]	21.21	12.63	23.74	37.37	57.83
ChangeCLIP-B[5]	27.78	14.14	30.81	43.43	72.59
GeoRSCLIP-B[47]	30.30	10.61	20.71	37.88	68.98
Prithvi-B[3]	21.21	8.08	17.17	28.79	40.66
RemoteCLIP-L[20]	31.31	11.62	21.71	37.88	67.92
SAR-CLIP-B[10]	31.82	29.80	44.95	65.66	69.43
SAR-JEPA-B[18]	28.28	31.82	57.58	68.69	58.89
SkyCLIP-L[37]	21.21	27.78	49.50	60.61	64.16
SARLANG-1M[38]	46.97	36.36	67.17	83.33	64.61
BITA-L[39]	21.21	10.61	18.18	35.35	60.99
GeoChat-L[13]	31.82	9.60	16.67	39.90	69.73
VHM-L[26]	17.17	7.58	14.65	30.81	68.83
FUSAR-GPT	**52.53**	**52.02**	**79.29**	**91.41**	**74.04**

We further conduct comparisons on the Target Counting task with proprietary large models, including ChatGPT-5.2 and Gemini-3. As shown in Fig.7, we explore different strategies for integrating AEF features.Table 6 show that our method consistently outperforms these strong baselines, while the proposed TLM-based fusion achieves the best performance among all integration schemes. These findings further validate the effectiveness of the TLM module in leveraging AEF priors for improved SAR understanding.

Table 6. models evaluated on Target Counting task.

Model	Base	TLM	Sum	Concat	ChatGPT-5.2	Gemini-3
Accuracy	34.85	52.53	36.36	37.88	18.18	30.3

At IoU = 0.5, we conduct a preliminary comparison with a representative remote sensing detector, R3Det, on a SAR

Table 7. Corpus-level captioning performance.

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	METEOR	SPICE
LLaVA-1.5-7B[21]	73.54	67.20	61.34	55.59	81.97	42.13	84.14
LLaVA-1.6-7B[22]	75.03	69.08	63.35	57.83	93.63	43.13	84.94
InternVL-3.5-4B[36]	76.76	70.76	64.99	59.36	116.21	42.34	84.36
Qwen2-VL-3B[35]	72.11	65.80	59.80	54.11	91.34	41.20	83.07
Qwen2-VL-7B[35]	76.11	70.20	64.57	59.13	109.11	42.52	84.62
Qwen2.5-VL-3B[2]	61.56	56.27	51.17	46.24	81.68	35.88	81.22
Qwen2.5-VL-7B[2]	66.71	61.39	56.22	51.19	91.68	37.95	83.02
Qwen3-VL-4B[1]	77.64	71.61	65.87	60.31	116.20	43.13	85.45
Qwen3-VL-8B[1]	77.41	71.49	65.84	60.36	110.79	43.09	85.71
SARLANG-1M[38]	74.56	69.84	65.17	60.61	149.60	43.16	88.12
BAN(ViT-L)[16]	58.08	51.39	44.99	38.79	14.08	32.79	56.99
ChangeCLIP(ViT-B)[5]	59.95	53.54	47.58	41.52	16.40	34.12	57.29
GeoRSCLIP(ViT-B)[47]	59.01	52.62	46.51	40.35	16.63	33.20	57.49
Prithvi(ViT-B)[3]	60.05	53.33	46.83	40.37	11.72	33.37	61.81
RemoteCLIP(ViT-L)[20]	61.12	54.54	48.10	41.64	17.09	34.66	59.19
SAR-CLIP(ViT-B)[10]	59.91	53.41	47.05	40.73	13.78	34.06	61.13
SAR-JEPA(ViT-B)[18]	57.11	52.65	44.79	37.21	10.03	32.11	58.44
SkyCLIP(ViT-L)[37]	60.77	54.19	47.95	41.69	14.91	34.74	57.81
BITA (ViT-L)[39]	55.37	44.72	37.32	30.72	2.26	28.33	56.13
GeoChat(ViT-L)[13]	54.04	43.55	36.10	29.67	2.19	26.11	57.66
VHM(ViT-L)[26]	50.29	41.92	35.36	29.25	3.14	27.27	58.91
FUSAR-GPT	**77.92**	**73.21**	**68.62**	**64.01**	**160.21**	**45.60**	**89.29**

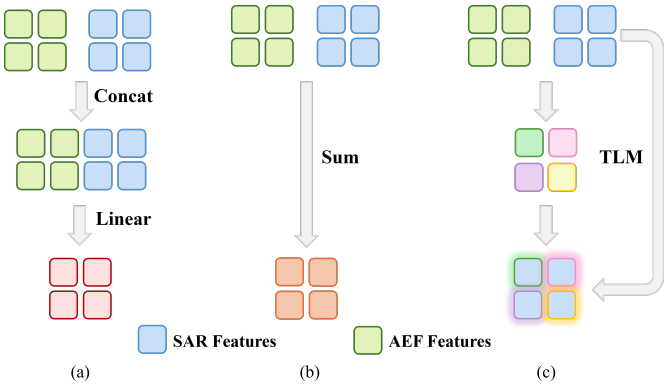


Figure 7. Different Fusion Strategies

object detection task. Although our approach is not specifically designed for detection, it demonstrates competitive performance. These results provide an initial indication of the potential of our framework for SAR detection tasks. We leave a more comprehensive evaluation and further exploration of detection-oriented adaptations to future work.

Table 8. models evaluated on Target Detection task.

Model	P	R	F1
R3det[40]	62.77	44.73	52.20
FUSAR-GPT	66.79	52.10	58.70

To validate the generalization capability of our model, we further extend the evaluation to a corpus-level captioning task, as shown in Table 7. Compared with both general-domain VLMs and remote sensing–specific models, FUSAR-GPT achieves the best performance across all metrics, including BLEU, CIDEr, METEOR, and SPICE. Notably, it surpasses the strongest baseline by a clear margin in CIDEr and BLEU-4, indicating its superior ability to generate semantically rich and structurally accurate descriptions for SAR imagery.

9. Ablation Experiment

Table 9. Ablation results on the target counting task. Each component:SFT1, SFT2, and TLM independently contributes to performance improvement, while combining all modules yields the highest accuracy.

Model	SFT1	SFT2	TLM	ACC
BaseModel	✗	✗	✗	-
BaseModel(+SFT2)	✗	✓	✗	34.85
BaseModel(+SFT1+SFT2)	✓	✓	✗	36.36
BaseModel(+SFT2+TLM)	✗	✓	✓	41.92
FUSAR-GPT	✓	✓	✓	**52.23**

To supplement the main text, we conduct detailed ablation studies on the target counting task to evaluate

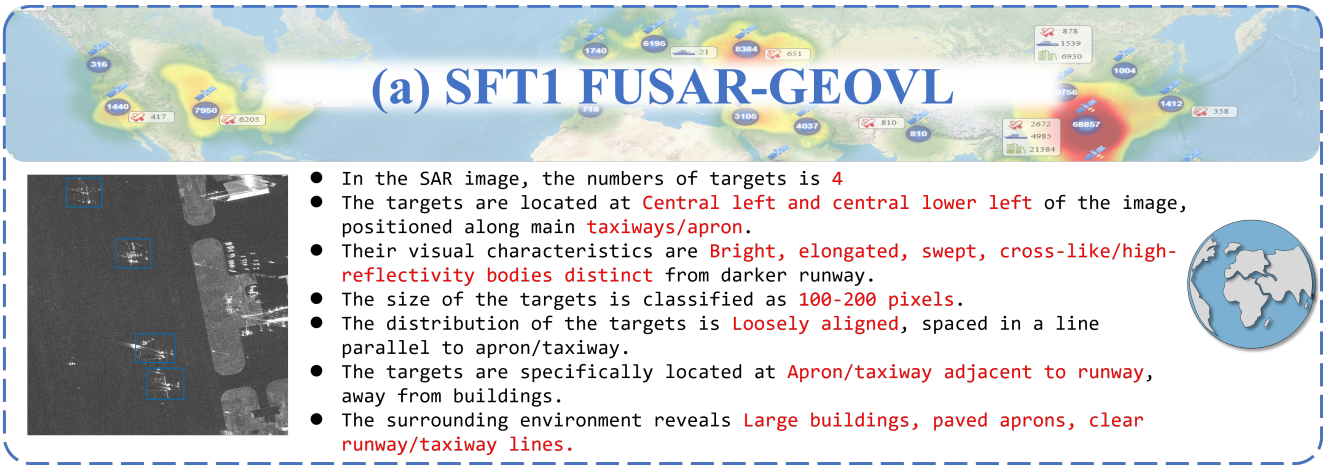


Figure 8. (a) shows the data characteristics of FUSAR-GEOVL, (b) shows the data characteristics of FUSAR-GPT.

the contribution of the first-stage supervised fine-tuning (SFT1),which aims to inject SAR-specific knowledge into the model. Unlike SFT2, which directly targets downstream tasks, SFT1 enhances the model's basic understanding of SAR imaging characteristics and domain patterns, providing a more informative initialization.

As shown in Table 9, the Qwen2.5-VL baseline using only SFT2 achieves a performance of 34.85%, revealing a significant bottleneck in SAR understanding. Incorporating the TLM module to fuse AEF features improves the performance to 41.92%, demonstrating the effectiveness of AEF geographical priors as an independent knowledge source. Further introducing SFT1 yields an additional gain of approximately 2 percentage points, ultimately boosting the full model to 52.23%. These results highlight a strong synergy between SFT1 pre-alignment and TLM feature fusion: SFT1 establishes a superior semantic alignment foundation, enabling TLM to more effectively exploit AEF priors, thereby strengthening domain awareness and achieving optimal downstream performance.

10. Data Description

As shown in Fig.8, the SFT1 stage is built upon the FUSAR-GEOVL dataset, which provides richly structured and multi-dimensional semantic information that substantially enhances the model's understanding of SAR imagery. Unlike conventional SAR datasets that offer only category labels or sparse annotations, FUSAR-GEOVL incorporates geographic metadata, multi-scale spatial context, landform descriptions, regional functional cues, and detailed target-level attributes—covering scattering characteristics, structural patterns, spatial layout relations, and environmental semantics. This information-rich annotation design, grounded in SAR physical imaging principles and global-to-local cognitive reasoning, supplies the model with a broad and coherent knowledge base. Through SFT1, the model internalizes these diverse information dimensions, enabling stronger perception of SAR-specific structures, better discrimination of subtle scattering variations, and improved reasoning over spatial and contextual relationships. As a result, SFT1 serves as an effective knowledge injection phase that significantly strengthens the model's domain awareness before downstream task tuning.