arXiv:2602.19190v4 [cs.CV] 4 Jun 2026

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

Xiaokun Zhang[1*], **Yi Yang**[1*†], **Ziqi Ye**[1,2*], **Baiyun**[1], **Xiaorong Guo**[1], **Qingchen Fang**[1], **Ruyi Zhang**[1], **Xinpeng Zhou**[1], **Haipeng Wang**[1‡]

[1]Discipline and Technology Center of Microwave Vision Intelligent Sensing, Fudan University [2]Shanghai Innovation Institute

[*]Equal contribution [†]Project leader: Yi Yang (yangy22@m.fudan.edu.cn)

[‡]Corresponding author: Haipeng Wang (hpwang@fudan.edu.cn)

Abstract

Research on the intelligent interpretation of all-weather, all-time Synthetic Aperture Radar (SAR) is crucial for advancing remote sensing applications. In recent years, although Visual Language Models (VLMs) have demonstrated strong open-world understanding capabilities on RGB images, their performance is severely limited when directly applied to the SAR field due to the complexity of the imaging mechanism, sensitivity to scattering features, and the scarcity of high-quality text corpora. To systematically address this issue, we constructed the inaugural SAR Image-Text-AlphaEarth feature triplet dataset and developed FUSAR-GPT, a VLM specifically for SAR. FUSAR-GPT innovatively introduces a geospatial baseline model as a "world knowledge" prior and embeds multi-source remote-sensing temporal features into the model's visual backbone via "spatiotemporal anchors", enabling dynamic compensation for the sparse representation of targets in SAR images. Furthermore, we designed a two-stage SFT strategy to decouple the knowledge injection and task execution of large models. The spatiotemporal feature embedding and the two-stage decoupling paradigm enable FUSAR-GPT to achieve state-of-the-art performance across several typical remote sensing visual-language benchmark tests, significantly outperforming mainstream baseline models by over 10%.

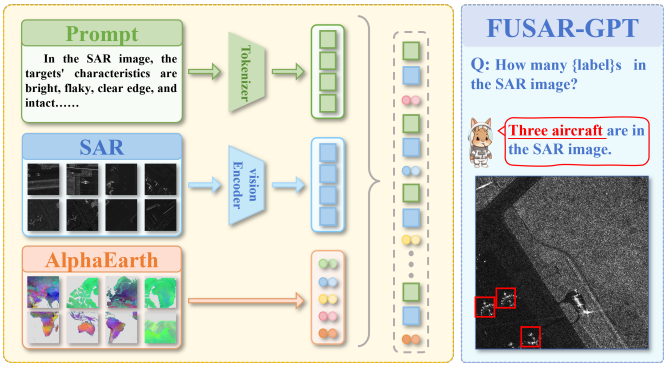


Figure 1. FUSAR-GPT: Embeds spatiotemporal features for SAR feature compensation and uses two-stage decoupled SFT for target interpretation.

1. Introduction

The intelligent interpretation of SAR imagery has long been recognized as a pivotal research focus in the field of remote sensing[34]. Despite progress in specific tasks like object detection, SAR image interpretation remains highly challenging due to its complex electromagnetic scattering and coherence effects. Concurrently, multimodal language models (e.g., CLIP [27], BLIP [15]) have excelled in natural image domains, demonstrating strong semantic alignment and cognitive reasoning by pre-training on large-scale image-text pairs.

This technological trend is gradually extending into the field of remote sensing, with several studies attempting to adapt vision-language models to remote sensing image scenarios, demonstrating their feasibility on optical remote sensing images. However, existing approaches predominantly rely on RGB imagery and have not yet accounted for the modality-specific characteristics and semantic complexity inherent in SAR images [13]. Achieving this objective requires overcoming several key challenges:

(a) SAR-Optical Modal Difference: Existing multimodal models, pretrained on large-scale visible light datasets, possess feature representations that fundamentally mismatch the data distribution of SAR imagery due to its distinct imaging mechanism. Directly transferring these models fails to adapt to this inter-modal discrepancy[6], resulting in poor generalization and limited interpretive per-

formance. This ultimately renders the transfer paradigm ineffective for SAR tasks.

(b) Neglecting Geospatial Priors: Geographic information serves as an irreplaceable strong prior constraint for remote sensing interpretation. Current SAR interpretation research, however, suffers from a systematic flaw: it adopts technical frameworks designed for natural images, resulting in models that lack spatial awareness. As a result, essential geographic scene priors are underutilized, weakening high-level semantic reasoning in vision–language models (e.g., confusing urban buildings with metallic objects) and exacerbating hallucinations.

(c) Information Sparsity: Owing to the coherent imaging mechanism of SAR and its extreme sensitivity to geometric and dielectric properties, SAR images inherently exhibit an extremely high dynamic range and information sparsity. This polarized data distribution—where artificial targets (such as corner reflectors) generate excessively saturated strong scattering, and natural objects (such as water surfaces) appear as extensive dark regions—causes the model's attention to be dominated by a few bright pixels. Consequently, the rich contextual semantics embedded in the vast dark areas are systematically overlooked, severely limiting the VLM's ability to achieve comprehensive semantic comprehension of the scene.

To address the aforementioned issues, we propose FUSAR-GPT. Built upon Qwen2.5-VL-7B, FUSAR-GPT introduces a spatiotemporal semantic embedding module that incorporates geospatial features as "world knowledge" priors into the model's visual encoder, dynamically compensating for the sparse representation of targets in SAR images. Furthermore, we develop a two-stage supervised fine-tuning (SFT) strategy to decouple knowledge injection from task execution. The main contributions of this paper are summarized as follows:

- This study establishes the first "SAR Image-Text-Feature" triplet data paradigm, innovatively introducing geospatial features as a third modality. By leveraging spatiotemporal anchors, this paradigm integrates rich world knowledge priors to enable dynamic semantic compensation for sparse SAR features.
- We propose the Token-wise Linear Modulation (TLM) fusion module. TLM achieves fine-grained, dynamic semantic injection by employing local spatial alignment and per-channel linear modulation to transform high-dimensional priors into spatially differentiated parameters for visual token adjustment.
- We designed an innovative two-stage SFT paradigm to systematically decouple SAR modality knowledge injection (establishing cognitive capability) from the execution of downstream tasks (endowing higher-level analytical and reasoning abilities).
- FUSAR-GPT achieved state-of-the-art performance in multiple SAR interpretation tasks. It demonstrated a significant performance improvement of more than 10% over mainstream VLMs.

2. Related Work

2.1. Research on SAR image interpretation

Early SAR image processing relied on statistical modeling and manual feature extraction for tasks like speckle suppression and basic recognition [7]. More recently, the explosive development of deep learning, particularly CNN-based frameworks, has delivered significant performance improvements in specific tasks like SAR target detection[25], semantic segmentation[30], and scene classification [11], [12], [33], [46].

Specialized techniques have also emerged: SARDet-CL, a contrastive learning framework, improves detection with limited labeled samples by adaptively enhancing features through dynamic quantization and imaging-mechanism constraints [41]. Furthermore, to address geometric challenges like large-scale scenes, progressive detection frameworks integrate traditional methods with deep learning to achieve high efficiency and accuracy [9].

Despite these advancements, deep learning methods remain limited to modeling low-level visual features, leading to a serious deficiency in high-dimensional semantic understanding and cognitive-level reasoning. Therefore, a new interpretation paradigm is urgently required to bridge the SAR vision and language modalities and enable semantic knowledge injection.

2.2. Research on Vision-Language Models in Remote Sensing

General VLMs have made substantial breakthroughs in the natural image domain[8][49][19][23], rapidly driving the paradigm shift of remote sensing interpretation toward a cognitive level. Researchers in remote sensing aim to bridge the semantic gap between remote sensing images and natural language by constructing large-scale image–text alignment datasets and developing generative models for data augmentation [43] [24]. Representative works include EarthGPT [44], which proposed a general multi-sensor image understanding framework emphasizing visual perception and cross-modal mutual understanding; GeoChat [13] developed the first multi-task conversational remote sensing VLM, focusing on region-level reasoning for high-resolution images. EarthDial [29] further expanded its scope to support interactive dialogues across multi-temporal and multi-sensor data (including SAR), and released a large-scale instruction dataset. In addition, works such as SkyScript [37] and RemoteClip[20] have also made significant progress in tasks including VQA, description generation, and functional area recognition for optical remote

sensing images through large-scale datasets and instruction fine-tuning.

However, despite the rapid development of remote sensing VLM[48][29][45], existing research has mainly focused on visible (RGB) or multispectral images, and studies specifically targeting visual-language models for SAR images remain at an early stage and are extremely scarce.

Only a few preliminary works (e.g., SARCLIP [10], SARLANG-1M [38]) have been attempted, primarily achieving basic modal alignment and zero-shot generalization. However, their capabilities remain confined to feature-level alignment, critically failing to account for SAR's intrinsic modality specificity and systematically overlooking crucial geospatial prior information. Consequently, advancing intelligent SAR interpretation toward multimodal, open-ended understanding remains highly challenging.

3. Method

FUSAR-GPT is a SAR vision-language model, built as a customized enhancement upon the Qwen2.5VL-7B [2] architecture (overall framework shown in Fig. 2). Its core design encompasses two aspects: multi-source remote sensing temporal feature embedding, and a two-stage decoupled SFT strategy.

3.1. AlphaEarth Multi-Source Temporal Feature Extraction

The microwave active imaging mechanism of SAR leads to severe information imbalance and semantic sparsity in image content. To address this bottleneck, FUSAR-GPT introduces AEF [4] as a multi-source temporal feature compensation. AEF is a global remote sensing foundation model that integrates heterogeneous multi-source data, such as optical, SAR, and LiDAR, into a 64-dimensional continuous spatiotemporal embedding field. This embedding field provides FUSAR-GPT with the cross-modal prior knowledge necessary to compensate for sparse SAR features and achieve robust understanding.

To precisely align the multi-source prior knowledge of AEF with the visual features of SAR images, we design a feature extraction process based on spatio-temporal anchors.

First, for any given SAR image, we determine its spatio-temporal bounding box B.

$$B = [lon_{\min}, lat_{\min}, lon_{\max}, lat_{\max}, y] \tag{1}$$

Where $[lon, lat]$ specifies the geographic coverage of the image and y denotes its imaging year.

Subsequently, to capture the AEF continuous embedding field, we construct an $N_{lon} \times N_{lat}$ regular sampling grid over B. The geographic coordinates (lon_i, lat_j) for node (i, j) are computed as:

$$\begin{aligned} \Delta lon &= \frac{lon_{\max} - lon_{\min}}{N_{lon} - 1}; \quad \Delta lat = \frac{lat_{\max} - lat_{\min}}{N_{lat} - 1} \\ lon_i &= lon_{\min} + i \cdot \Delta lon; \quad lat_j = lat_{\min} + j \cdot \Delta lat \end{aligned} \tag{2}$$

where $0 \le i < N_{lon}$ and $0 \le j < N_{lat}$.

On each geographic node (lon_i, lat_j), we query the AEF model I_i to obtain the 64-dimensional embedding vector corresponding to year y:$a_{ij} = I_y(lon_i, lat_j) \in \mathbb{R}^{64}$.

Crucially, to align AEF semantic features with SAR visual features, we perform a linear mapping from each geographic coordinate (lon_i, lat_j) to the corresponding pixel coordinate system (x_{ij}, y_{ij}) of the SAR image.

Ultimately, the extracted AEF features are organized into a set $\mathcal{F}$, which encapsulates the precise alignment relationships among geolocation, pixel indexing, and multi-source semantic embeddings:

$$\mathcal{F}(B, y) = \{(lon_{ij}, lat_{ij}, x_{ij}, y_{ij}, a_{ij})\}_{i,j} \tag{3}$$

This process ensures that cross-modal prior knowledge (AEF) can be accurately injected into specific spatial locations within SAR images.

Fig. 3 verifies the significant complementarity between SAR and AEF features. Leveraging spatiotemporal anchor alignment, AEF provides critical dynamic semantic compensation for the sparse and incomplete SAR representations caused by imaging limitations. As shown in the enlarged area on the right side of Fig. 3, AEF enhances weak features (such as farmland areas) and provides clean, stable object representations even in areas affected by specular scattering noise.

Although AEF provides multi-source temporal embeddings, we do not explicitly model temporal dynamics. Instead, AEF is used as a stable geospatial prior, as macroscopic geographic semantics remain relatively consistent at the annual scale, providing consistent large-scale context.

3.2. TLM fusion module

The previous analysis demonstrates that AEF provides complementary semantic cues for SAR imagery, supplying dynamic and globally consistent priors. The remaining challenge is how to efficiently introduce this external multi-source knowledge into the visual representation space.

AEF and visual tokens are heterogeneous in both form and modality: AEF consists of sparsely sampled geo-semantic vectors, while visual tokens are dense deep features extracted from image patches. Naive strategies such as concatenation or direct fusion cause alignment mismatches, increase computation, and may distort the spatial structure already learned by the visual backbone.

To address this, FUSAR-GPT adopts the TLM module, inspired by conditional normalization. Instead of treating

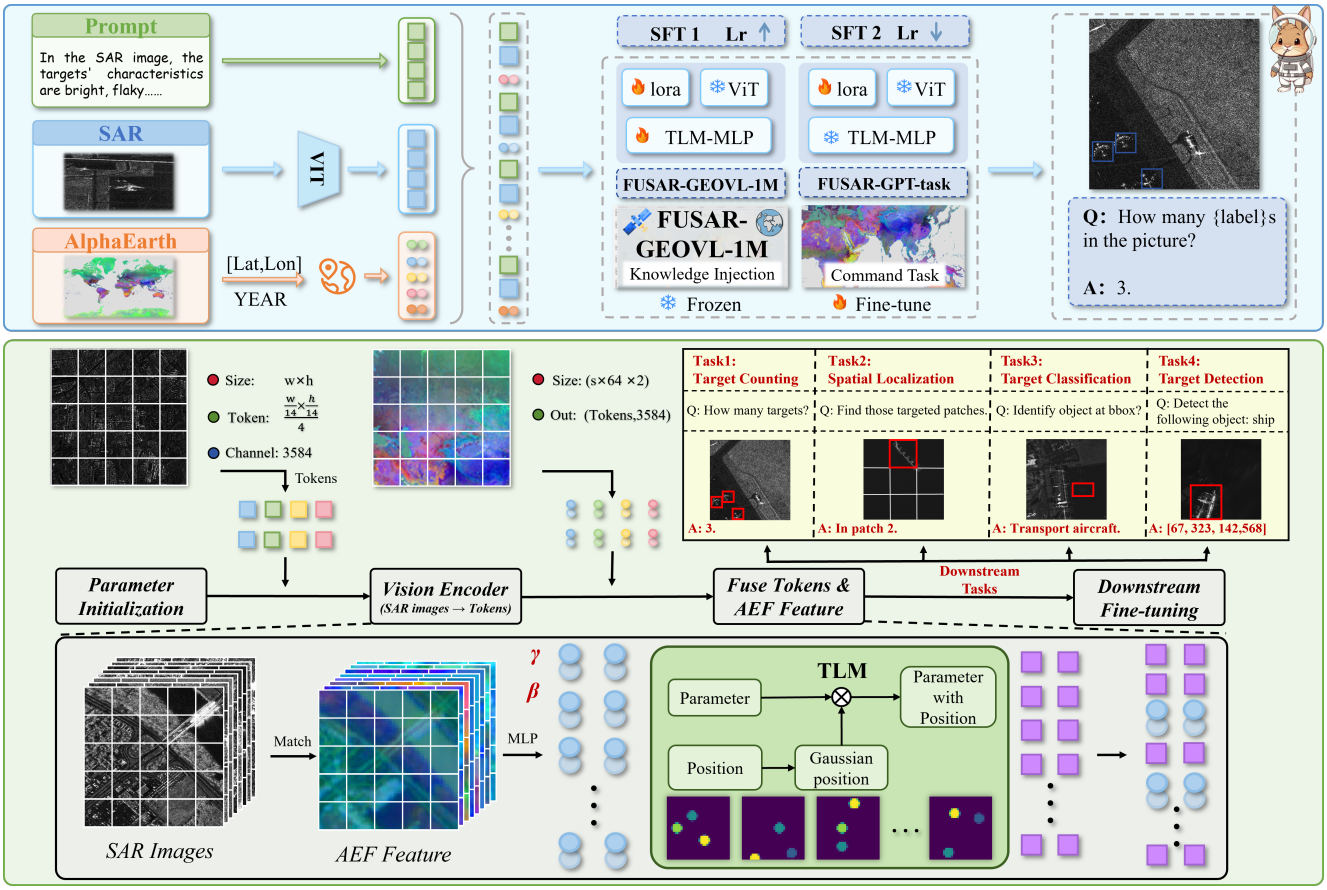


Figure 2. Overview of FUSAR-GPT. The framework adopts a two-stage training strategy: Stage-1 jointly updates LoRA and the TLM-MLP for multimodal prior injection, while Stage-2 fine-tunes only LoRA for task adaptation. The TLM module fuses SAR visual tokens with AEF priors by generating spatially informed modulation parameters, which are applied to the visual tokens to enhance downstream reasoning.

AEF as additional feature inputs, TLM interprets them as conditioning signals that generate modulation parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, which apply channel-wise affine transformations to the visual tokens X. Through this lightweight modulation, AEF guides the refinement of SAR representations without disrupting the backbone's spatial encoding.

TLM for AEF–SAR Fusion. After obtaining the AEF feature field $\mathcal{F}(B, y)$ that is strictly aligned with the SAR image space, FUSAR-GPT introduces a TLM fusion module at the output of the visual encoder in order to explicitly inject the cross-modal prior information without disrupting the structure of the backbone visual pathway. Note that the token sequence output by the visual encoder is as follows: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^\top \in \mathbb{R}^{T \times C}$, where $\mathbf{x}_t \in \mathbb{R}^C$ denotes the channel features of the t-th visual token, and C is the visual hidden dimension. Correspondingly, the structured priors obtained from AEF sampling are organized into a vector sequence as follows: $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_S]^\top \in \mathbb{R}^{S \times D}$, $\mathbf{P} = [(y_1, x_1), \ldots, (y_S, x_S)]^\top \in [0, 1]^{S \times 2}$, where $\mathbf{v}_i \in \mathbb{R}^D$ is the i-th AEF embedding vector, (y_i, x_i) is its normalized spatial coordinate on the SAR image, and S is the number of valid prior vectors for the sample.

Prior-to-Modulation Mapping. The core idea of TLM is to treat each AEF vector $\mathbf{v}_i$ as a conditioning signal that generates channel-wise modulation parameters for the visual tokens. Specifically, $\mathbf{v}_i \in \mathbb{R}^D$ is projected into a $2C$-dimensional space to produce a pair of scaling and shifting coefficients $(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i) \in \mathbb{R}^C \times \mathbb{R}^C$ for modulating the visual feature $\mathbf{x}_t$. Specifically, a two-layer MLP with SiLU activation is employed:

$$\mathbf{h}_i = \phi(\mathbf{W}_1 \mathbf{v}_i + \mathbf{b}_1), (\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i) = \mathbf{W}_2 \mathbf{h}_i + \mathbf{b}_2 \quad (4)$$

Where $\phi(\cdot)$ is the SiLU activation function, $\mathbf{W}_1 \in \mathbb{R}^{h \times D}$, $\mathbf{W}_2 \in \mathbb{R}^{2C \times h}$, and h is the hidden dimension. Thus, the modulation coefficient matrices corresponding to

all external vectors can be obtained as

$$\Gamma = \begin{bmatrix} \boldsymbol{\gamma}_1^\top \\ \vdots \\ \boldsymbol{\gamma}_S^\top \end{bmatrix} \in \mathbb{R}^{S\times C}, B = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \vdots \\ \boldsymbol{\beta}_S^\top \end{bmatrix} \in \mathbb{R}^{S\times C} \tag{5}$$

Local Spatial Alignment Based on Gaussian Weights. The above procedure produces S sets of modulation parameters (Γ, B) defined on the sparse AEF locations $\mathbf{P}$. To modulate the dense visual tokens $\mathbf{X}$, these parameters must be spatially aligned and interpolated onto the visual grid. TLM achieves this by leveraging the intermediate spatial mesh of the visual encoder. The normalized AEF coordinates are projected onto the feature map of size $H \times W$, yielding the continuous grid location of the i-th prior vector:

$$\tilde{y}_i = y_i(H-1), \tilde{x}_i = x_i(W-1) \tag{6}$$

At each spatial position (h, w), we use a distance-based Gaussian kernel to define the interpolation weights, and define an unnormalized Gaussian weight for the i-th external vector as follows:

$$\tilde{w}_i(h,w) = \exp\Big(-\frac{(h-\tilde{y}_i)^2 + (w-\tilde{x}_i)^2}{2\sigma^2}\Big) \tag{7}$$

Subsequently, column-wise normalization is performed along the prior-vector dimension so that the weights at each spatial position sum to 1:

$$w_i(h,w) = \frac{\tilde{w}_i(h,w)}{\sum_{j=1}^{S}\tilde{w}_j(h,w)+\varepsilon}, \sum_{i=1}^{S} w_i(h,w) = 1 \tag{8}$$

By flattening the two-dimensional index (h, w) into a single index $k = hW + w$, the overall weight matrix can be obtained as follows:

$$W = [w_i(k)]_{i,k} \in \mathbb{R}^{S\times HW} \tag{9}$$

The k-th column corresponds to the normalized weights contributed by each external vector to the k-th spatial location on the feature map.

TLM Fusion and Token Alignment. After obtaining the spatial interpolation weight matrix W, we apply weighted aggregation to the modulation parameters (Γ, B) defined at the sparse AEF positions. In this way, we obtain a channel modulation field that is strictly aligned with the visual feature map:

$$\Gamma_{\text{hw}},\ B_{\text{hw}} = W^\top(\Gamma,\ B) \in \mathbb{R}^{HW\times C} \tag{10}$$

Treat Γ_{hw} and B_{hw} as modulation fields defined on the discrete spatial grid. They can be written respectively as $\Gamma_{\text{hw}} \leftrightarrow \boldsymbol{\gamma}(h,w)$, $B_{\text{hw}} \leftrightarrow \boldsymbol{\beta}(h,w)$, where $\boldsymbol{\gamma}(h,w)$ and $\boldsymbol{\beta}(h,w) \in \mathbb{R}^C$ denote the channel-wise scaling and offset coefficients at the spatial location (h, w).

Table 1. Training datasets used in this work: Stage 1 fine-tuning is performed on FUSAR-GEOVL-1M, and Stage 2 fine-tuning and evaluation are conducted on FUSAR-GPT.

Name	Year	Band	Satellite	Pol	AEF	Example
FUSAR -GEOVL-1M	2025	C,X,Ku	GF3, QL1, HT1	Multi	✗	
FUSAR-GPT	**2025**	**C,X,Ku,L**	**GF3, QL1 Sentinel-1A, HT1 Sentinel-1B,ALOS**	**Multi**	✓	

The final step is to apply this modulation field to the one-dimensional token sequence $\mathbf{X} \in \mathbb{R}^{T\times C}$ output by the visual encoder. We establish a bijective mapping π from the spatial grid to the token index set, such that the $\pi(h,w)$-th element in the visual token sequence uniquely corresponds to the spatial position (h, w).

Finally, channel-wise TLM fusion for each visual token is achieved.

$$\mathbf{x}'_{\pi(h,w)} = \mathbf{x}_{\pi(h,w)} \odot \big(1 + \boldsymbol{\gamma}(h,w)\big) + \boldsymbol{\beta}(h,w) \tag{11}$$

This process injects AEF priors into spatial and channel dimensions without modifying the main visual pathway, effectively improving the stability and discriminability of SAR representations.

3.3. Decoupled Two-Stage SFT

General-purpose VLMs (such as Qwen [2] and LLaVA [14]) exhibit a substantial domain gap in SAR applications due to optical pre-training. To prevent suboptimal performance caused by the simultaneous, conflicting optimization of multimodal fusion (SAR, AEF, and text) and instruction-driven task execution in a single-stage scheme, we propose a two-stage decoupled supervised fine-tuning (SFT) framework. The datasets used in the two stages are summarized in Table 1.

We decompose the parameter set of FUSAR-GPT, denoted as Θ, into four components: $\Theta = \theta_v \cup \theta_{ae} \cup \theta_{llm} \cup \theta_{\text{lora}}$, corresponding respectively to the visual encoder E_v, the AEF feature-embedding MLP E_{ae}, the backbone large language model L, and the LoRA adapters injected into the language model. Our decoupling framework performs staged optimization over different subsets of parameters.

Stage I: Cross-Modal Alignment and Knowledge Injection The core objective of this phase is to enable the model to understand and integrate SAR visual representations, AEF geographic priors, and textual descriptive semantics. We perform instruction fine-tuning using the dataset $D_1 = \{(I_{sar}, F_{ae}, T_{desc})\}$, where I_{sar} denotes the SAR image, F_{ae} is the corresponding AEF feature set, T_{desc} is a text corpus primarily consisting of comprehensive descriptive semantics (e.g., geographic information such as terrain, landforms, and spatial distributions). This corpus is

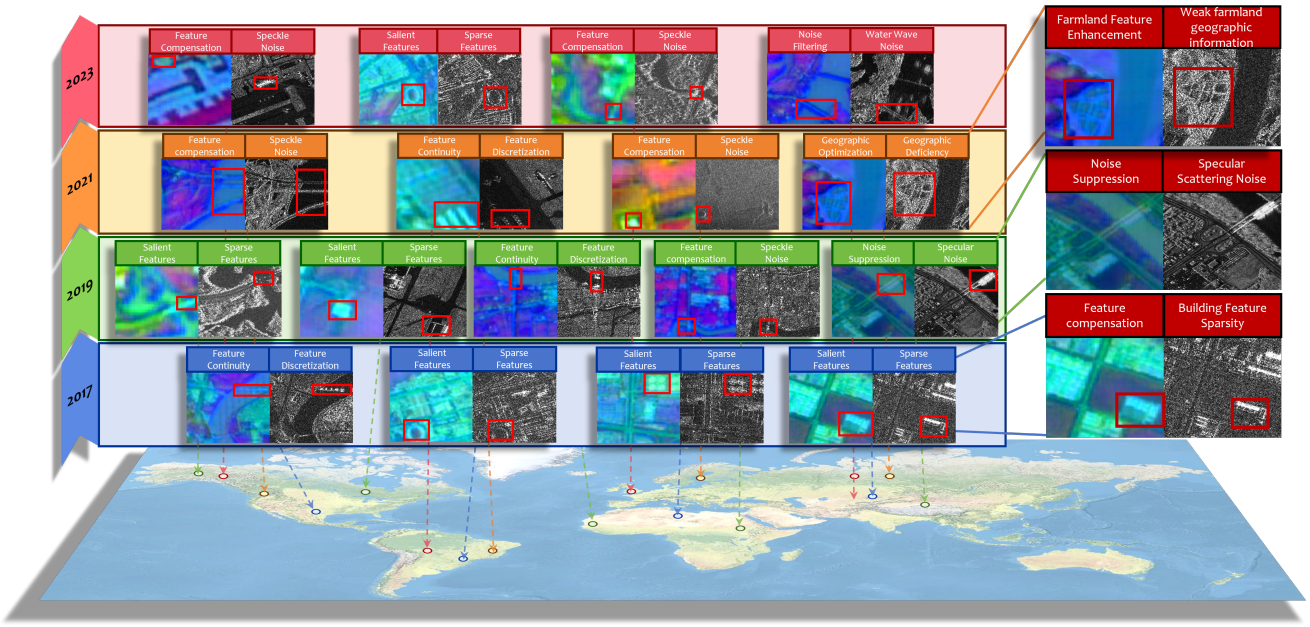


Figure 3. AEF–SAR Visual Comparison. AEF embeddings extracted at different times are visualized by mapping channels 1, 16 and 9 of the 64-dimensional feature vector to RGB.

derived from the FUSAR-GEOVL-1M dataset [42], which covers multiple satellite platforms (e.g., GF-3 and QL-1) and multi-resolution imagery, ensuring strong domain generalization capability for the model.

In terms of the training strategy, during this stage we freeze the visual encoder θ_v and the original weights of the backbone large language model θ_{llm}. The AEF feature-embedding MLP θ_{ae} and the LoRA adapters θ_{lora} injected into the language model are jointly trained during this stage. The optimization objective for this phase, denoted as $\mathcal{L}_1$, is to maximize the likelihood of the descriptive text T_{desc}:

$$\mathcal{L}_1(\theta_{ae}, \theta_{\text{lora}}) = -\mathbb{E}_{(I,F,T)\in D_1}\Big[\log P\big(T \mid E_v(I;\theta_v^{\text{frozen}}), E_{ae}(F;\theta_{ae});\, \theta_{llm}^{\text{frozen}},\, \theta_{\text{lora}}\big)\Big] \tag{12}$$

$$(\theta_{ae}^*,\, \tilde{\theta}_{\text{lora}}) = \arg\min_{\theta_{ae},\,\theta_{\text{lora}}} \mathcal{L}_1(\theta_{ae}, \theta_{\text{lora}}) \tag{13}$$

This design (as shown in equation 12) forces the MLP layer θ_{ae}, together with the LoRA adapters θ_{lora}, to learn how to efficiently integrate the multi-source semantic features (I, F) and align the fused representation with the descriptive text T.

Stage II: Task Reasoning and LLM Activation The core objective of this phase is to activate the model's analytical capabilities based on the aligned representations so that it can perform specific tasks. We employ the instruction dataset $D_2 = \{(I_{sar}, F_{ae}, T_{inst}, T_{ans})\}$, where T_{inst} denotes the task instruction (e.g., localization, classification, counting, and detection), and T_{ans} is the corresponding ground-truth answer.

In terms of the training strategy, we freeze the visual encoder θ_v, the MLP fusion layer θ_{ae}^* pretrained in Stage One, and the original weights of the large language model backbone θ_{llm}. Starting from the LoRA adapters $\tilde{\theta}_{\text{lora}}$ obtained in Stage I, we continue to update only the LoRA adapter parameters θ_{lora} during this stage.

The optimization objective of this stage, denoted as $\mathcal{L}_2$, is to maximize the likelihood of the task answer T_{ans}. The loss function is optimized exclusively over the LoRA parameters θ_{lora}:

$$\mathcal{L}_2(\theta_{\text{lora}}) = -\mathbb{E}_{(I,F,T_{\text{inst}},T_{\text{ans}})\in D_2}\Big[\log P\big(T_{\text{ans}} \mid E_v(I;\theta_v^{\text{frozen}}),\; E_{ae}(F;\theta_{ae}^*),\; T_{\text{inst}};\; \theta_{llm}^{\text{frozen}},\; \theta_{\text{lora}}\big)\Big] \tag{14}$$

$$\theta_{lora}^* = \arg\min_{\theta_{lora}} \mathcal{L}_2(\theta_{lora}) \tag{15}$$

Since the input features $E_v(I)$ and $E_{ae}(F;\theta_{ae}^*)$ have already undergone domain adaptation and cross-modal alignment for SAR data through the optimization of the $\mathcal{L}_1$ stage, the large language model in Stage Two can focus on interpreting the task instruction T_{inst} and performing complex analysis and inference.

4. Experiments

4.1. Dataset

The data used in this study are sourced from the FUSAR-GEOVL dataset within FUSAR-KLIP, which retains real latitude and longitude information to satisfy the geolocation requirements of AEF. During initial construction, we selected 10k images and textual descriptions, extracted corresponding AlphaEarth feature vectors based on geographic coordinates, and aligned them to form 10k AEF–image–text triplets. Subsequently, a subset of 2k images containing precise ground-truth target annotations was further selected for downstream task training and evaluation to validate the model's task transfer and generalization performance.

4.2. Implementation Details

All baseline models were fine-tuned in the MS-SWIFT framework under the same parameter settings as FUSAR-GPT for a fair comparison. FUSAR-GPT adopts Qwen2.5-VL-7B as the backbone. To ensure training consistency and stability, the first stage was trained for 30 epochs using a learning rate of 10^{-4} to facilitate cross-modal semantic alignment, followed by a second stage of 5 epochs with a learning rate of 10^{-5} to enable stable task-level adaptation without disrupting the previously aligned representations. Otherwise, Stage1 and Stage2 share the same system-level training configuration, including Flash-Attention-2, DeepSpeed ZeRO Stage-2, BFloat16 precision, and rank-8 LoRA adapters with LoRA scaling factor $\alpha = 32$ applied to all linear layers. Both stages are optimized using AdamW with a warm-up ratio of 0.05, and vLLM is employed to accelerate inference during evaluation. Since Qwen2.5-VL employs absolute coordinate encoding, we pre-aligned all bounding-box annotations to its unified image resolution. All experiments were conducted on four NVIDIA A100 GPUs.

4.3. Task

To comprehensively evaluate the model's performance in target interpretation, we selected four highly relevant mainstream tasks for testing, as shown in Fig. 4, including target counting, spatial localization, target classification, and target detection. For performance comparison, we employed mainstream vision–language large models as baselines, including models from the Qwen2, Qwen2.5, and Qwen3 series, the LLaVA series, and the InternVL series.

Target Counting: Table 2 reports the performance of various mainstream multimodal models on the SAR counting task. Overall, the accuracy of baseline models mostly falls within the range of 30% to 40%, indicating that general vision–language models still encounter notable performance bottlenecks when applied to SAR counting. Moreover, the table shows that increasing model size does not lead to substantial performance gains. For example, within the Qwen3-VL series, the 4-billion-parameter model achieves an accuracy of 45.45%, while the 8-billion-parameter model drops to 41.41%, suggesting that scaling up general-purpose models cannot overcome the inherent challenges of SAR imagery, such as strong noise and weak texture. In contrast, FUSAR-GPT attains an accuracy of 52.53%, exceeding the best baseline by more than 7%.

Table 2. Performance comparison on the Target Counting and Spatial Localization

Model	Year	Size	Counting	Spatial Localization		
			@Acc	@Acc100	@Acc50	Top1
Qwen2-VL[35]	2024	2B	35.86	23.23	46.97	75.25
		7B	41.41	35.35	67.68	83.33
Qwen2.5-VL[2]	2025	3B	35.35	23.23	46.97	68.18
		7B	34.85	30.81	57.07	78.28
Qwen3-VL[1]	2025	4B	45.45	43.94	71.21	82.32
		8B	41.41	42.93	71.72	85.86
LLaVA-1.5[21]	2023	7B	40.91	35.86	63.64	84.85
LLaVA-1.6[22]	2024	7B	44.44	39.90	70.20	85.86
InternVL-3.5[36]	2025	4B	40.40	39.90	70.20	83.84
FUSAR-GPT	2025	7B	**52.53**	**52.02**	**79.29**	**91.41**

Spatial Localization: To evaluate Spatial Localization, we divide each SAR image into a 3×3 grid and instruct the model to identify the target's grid index, assessing basic spatial reasoning and region-alignment. The validation set is diverse: 46.97% of images contain targets in only one grid, while 22.22% and 20.71% contain targets in two and three grids, respectively, and a smaller fraction contain targets across four to eight grids. We employ three metrics: **Acc@100** (strict exact match), **Acc@50** (IoU ≥ 0.5), and **Top1** (non-empty intersection). As reported in Table 2, FUSAR-GPT demonstrates substantial advantages across all indicators, achieving 52.02% on Acc@100, 79.29% on Acc@50, and 91.41% on Top1—improving upon the best baseline by 8–12%. Notably, the significant lead on the Top1 metric indicates FUSAR-GPT's superior stability in multi-target scenarios and a more reliable ability to identify key regions.

Target Classification: The model is given explicit bounding boxes and must identify the object category within each box. It is worth noting that the Qwen2-VL, Qwen3-VL, and InternVL series adopt a standardized 0–1000 relative coordinate system, whereas the LLaVA series does not officially support bounding-box coordinates, and prior work has shown that its region-alignment capability is limited[32][31]. Considering both model capability and fairness in bounding-box handling, we therefore conduct comparative experiments only with the Qwen2.5-VL series for this task. From the results in Table 3, it is evident that FUSAR-GPT achieves substantial gains in both coarse-grained and fine-grained classification tasks. FUSAR-

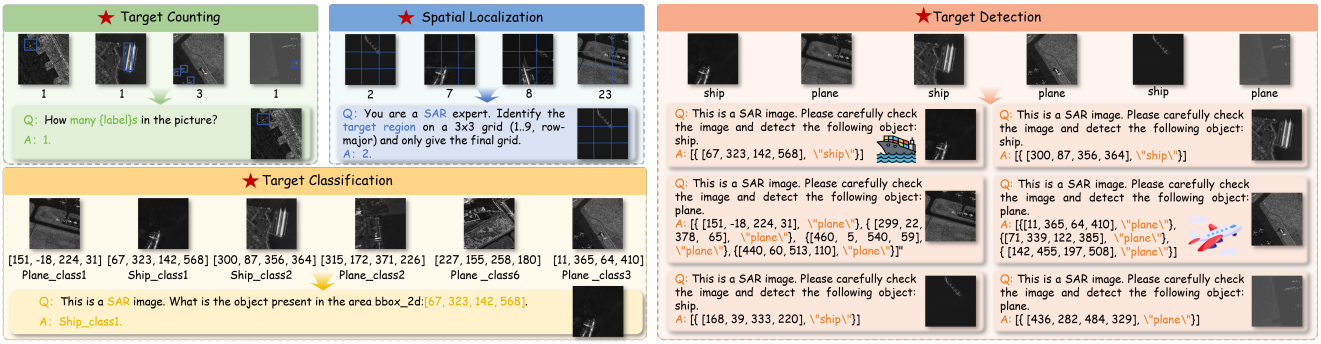


Figure 4. Overview of the four main downstream tasks: Target Counting, Spatial Localization, Target Classification, and Target Detection.

Table 3. Performance comparison on the Target Classification. Results are reported for both coarse-grained categories and fine-grained classes. FUSAR-GPT consistently outperforms Qwen2.5-VL-7B across all category levels.

Model	Coarse-grained		Fine-grained-plane													Fine-grained-ship		
	Plane	Ship	P-1	P-2	P-3	P-4	P-5	P-6	P-7	P-8	P-9	P-10	P-11	P-12	P-13	S-1	S-2	S-3
Qwen2.5-VL-7B[2]	64.66	55.19	82.42	83.95	66.67	25.00	31.43	57.14	69.05	57.89	53.85	40.00	0.00	85.71	35.00	57.65	51.06	100.00
FUSAR-GPT	**76.85**	**67.42**	**85.71**	**85.19**	**100**	**75.00**	**76.47**	**66.67**	**71.43**	**36.84**	**94.44**	**39.39**	**71.43**	**100**	**90.48**	**77.65**	**57.78**	**66.67**

Table 4. Performance comparison of Target Detection across different IoU thresholds.

Model	IoU = 0.25						IoU = 0.50						IoU = 0.75					
	All		Plane		Ship		All		Plane		Ship		All		Plane		Ship	
	R	F1	R	F1	R	F1	R	F1	R	F1	R	F1	R	F1	R	F1	R	F1
Qwen2.5-VL-7B[2]	39.9	47.1	40.3	47.5	32.3	38.5	23.5	27.7	24.3	28.7	6.5	7.7	6.2	7.3	6.3	7.5	3.2	3.8
FUSAR-GPT	**66.4**	**74.8**	**67.1**	**75.7**	**51.6**	**57.1**	**52.1**	**58.7**	**53.1**	**59.8**	**32.3**	**35.7**	**21.4**	**24.1**	**22.0**	**24.8**	**9.7**	**10.7**

GPT surpasses Qwen2.5-VL-7B by more than 12% in both coarse-grained categories. In fine-grained classification, the advantage of FUSAR-GPT is even more pronounced.

Target Detection: This task requires the model to detect and output all bounding boxes of a given class from SAR images based on a natural-language category prompt (e.g., "plane")[17][28]. As shown in Table 4, FUSAR-GPT exhibits a substantial advantage under all detection settings. At an IoU threshold of 0.25, the overall F1 score increases by nearly 28%, rising from 47.1% to 74.8%. The F1 score for the plane category improves from 47.5% to 75.7%, and for the ship category from 38.5% to 57.1%, indicating that the model also demonstrates stronger robustness for small-scale and low-contrast targets. Under a stricter IoU threshold of 0.50, FUSAR-GPT continues to maintain a clear lead.

As shown in Fig.5, we further analyze the accuracy dynamics throughout the training process. It is clearly observable that FUSAR-GPT demonstrates performance far exceeding all baseline models from the initial stages of training, consistently maintaining a significant lead. This indicates that FUSAR-GPT not only achieves superior final performance but also exhibits faster convergence and higher data efficiency, likely attributable to the robust prior knowledge injected during SFT Stage 1.

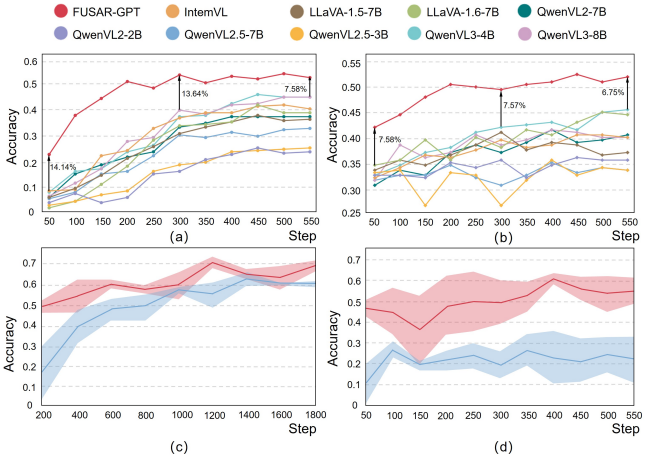


Figure 5. Performance of different training stages across downstream tasks. a: Target Counting; b: Spatial Localization; c: Target Classification; d: Target Detection.

5. Conclusion

To address the challenge that general VLMs exhibit limited performance in the SAR domain due to modality gaps and feature sparsity, we propose FUSAR-GPT. Our approach introduces AEF features as multi-source geospatial priors that leverage the stable nature of large-scale geographic semantics, and dynamically perform semantic compensation for sparse representations in SAR imagery through spatiotemporal anchors. To enable efficient fusion, we design the TLM module, which applies affine transformations to SAR visual tokens in a modulation manner. In addition, the proposed two-stage decoupled fine-tuning framework separates the model's knowledge injection and task execution processes at the parameter level. Experimental results show that FUSAR-GPT achieves competitive performance across multiple SAR vision–language downstream tasks.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62271153.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 7, 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 3, 5, 7, 8, 2

[3] Benedikt Blumenstiel, Viktoria Moor, Romeo Kienzler, and Thomas Brunschwiler. Multi-spectral remote sensing image retrieval using geospatial foundation models. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 7286–7291. IEEE, 2024. 1, 2

[4] Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. 3

[5] 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

[6] Junyu Gao, Da Zhang, Feiyu Wang, Lichen Ning, Zhiyuan Zhao, and Xuelong Li. Combining sam with limited data for change detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 1

[7] Zhe Geng, He Yan, Jindong Zhang, and Daiyin Zhu. Deep-learning for radar: A survey. *IEEE Access*, 9:141800–141818, 2021. 2

[8] Yan Huang, Yuming Wang, Yunan Zeng, Junshi Huang, Zhenhua Chai, and Liang Wang. Unpaired image-text matching via multimodal aligned conceptual knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5160–5176, 2024. 2

[9] Hecheng Jia, Xinyang Pu, Qiaoyu Liu, Haipeng Wang, and Feng Xu. A fast progressive ship detection method for very large full-scene sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 2

[10] Chaowei Jiang, Chao Wang, Fan Wu, Peifeng Ma, Lichuan Zou, Tianyang Li, Jing Ning, and Yixian Tang. Sarclip: a multimodal foundation framework for sar imagery via contrastive language-image pre-training. *ISPRS Journal of Photogrammetry and Remote Sensing*, 231:17–34, 2026. 3, 1, 2

[11] Luewen Jiang, Jing Zang, and Yan Zhang. Em-yolo: A fine-grained recognition model for aircraft targets in sar images. In *2025 International Conference on Communication Networks and Smart Systems Engineering (ICCNSE)*, pages 255–259, 2025. 2

[12] Xiao Ke, Xiaoling Zhang, Tianwen Zhang, Jun Shi, and Shunjun Wei. Sar ship detection based on an improved faster r-cnn using deformable convolution. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3565–3568, 2021. 2

[13] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 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, 2024. 1, 2

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 5

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[16] Kaiyu Li, Xiangyong Cao, and Deyu Meng. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 1, 2

[17] Qingyun Li, Shuran Ma, Junwei Luo, Yi Yu, Yue Zhou, Fengxiang Wang, Xudong Lu, Xiaoxing Wang, Xin He, Yushi Chen, et al. Co-training vision-language models for remote sensing multi-task learning. *Remote Sensing*, 18(2): 222, 2026. 8

[18] Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is

better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. 1, 2

[19] Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, et al. Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3): 1832–1849, 2025. 2

[20] 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. 2, 1

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7, 2

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7, 2

[23] Li Liu and Paul Fieguth. Texture classification from random features. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):574–586, 2012. 2

[24] Li Liu, Shuzhou Sun, Shuaifeng Zhi, Fan Shi, Zhen Liu, Janne Heikkilä, and Yongxiang Liu. A causal adjustment module for debiasing scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5): 4024–4043, 2025. 2

[25] Yongxiang Liu, Weijie Li, Li Liu, Jie Zhou, Bowen Peng, Yafei Song, Xuying Xiong, Wei Yang, Tianpeng Liu, Zhen Liu, et al. Atrnet-star: A large dataset and benchmark towards remote sensing object recognition in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 2

[26] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6381–6388, 2025. 1, 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[28] Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. *arXiv preprint arXiv:2512.10554*, 2025. 8

[29] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, and Salman Khan. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14313, 2025. 2, 3

[30] Xu Sun, Junyu Gao, and Yuan Yuan. Fully polsar image reconstruction for enhanced land cover mapping. *Pattern Recognition*, 169:111895, 2026. 2

[31] Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. Visual position prompt for mllm based visual grounding. *IEEE Transactions on Multimedia*, 2026. 7

[32] Akshar Tumu, Varad Shinde, and Parisa Kordjamshidi. Referring expressions as a lens into spatial language grounding in vision-language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3442–3455, 2025. 7

[33] İlter Türkmenli, Erchan Aptoula, and Koray Kayabol. Group equivariant u-net for the semantic segmentation of sar images. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, 2022. 2

[34] Bin Wang, Song Dai, Dongmei Song, Lei Chen, Weimin Chen, and Jintao Yu. Annotation-free, high-fidelity sar oil-spill image synthesis via classification-guided diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. 1

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 7, 2

[36] 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*, 2025. 7, 2

[37] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5805–5813, 2024. 2, 1

[38] 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. 3, 1, 2

[39] Cong Yang, Zuchao Li, and Lefei Zhang. Bootstrapping interactive image–text alignment for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 1, 2

[40] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3163–3171, 2021. 2

[41] Yi Yang, Zhengxin Lei, Xiuci Mo, Da Lu, Hecheng Jia, and Haipeng Wang. Sardet-cl: Self-supervised contrastive learning with feature enhancement and imaging mechanism constraints for sar target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. 2

[42] Yi Yang, Xiaokun Zhang, Qingchen Fang, Jing Liu, Ziqi Ye, Rui Li, Li Liu, and Haipeng Wang. Fusar-klip: Towards multimodal foundation models for remote sensing, 2025. 6

[43] Ziqi Ye, Shuran Ma, Jie Yang, Xiaoyi Yang, Ziyang Gong, Xue Yang, and Haipeng Wang. Object fidelity diffusion for remote sensing image generation. *arXiv preprint arXiv:2508.10801*, 2025. 2

[44] 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

[45] Wei Zhang, Miaoxin Cai, Yaqian Ning, Tong Zhang, Yin Zhuang, Shijian Lu, He Chen, Jun Li, and Xuerui Mao. Earthgpt-x: A spatial mllm for multi-level multi-source remote sensing imagery understanding with visual prompting, 2025. 3

[46] Yuan Zhang, Yang Song, Yanping Wang, and Hongquan Qu. A fast training method for sar large scale samples based on cnn for targets recognition. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5, 2018. 2

[47] 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*, 62:1–23, 2024. 1, 2

[48] Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Tianyu Li, Yuxiang Cai, Zhonggen Su, Zhaojun Liu, Jianwei Yin, and Xiang Li. Geo-r1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning, 2025. 3

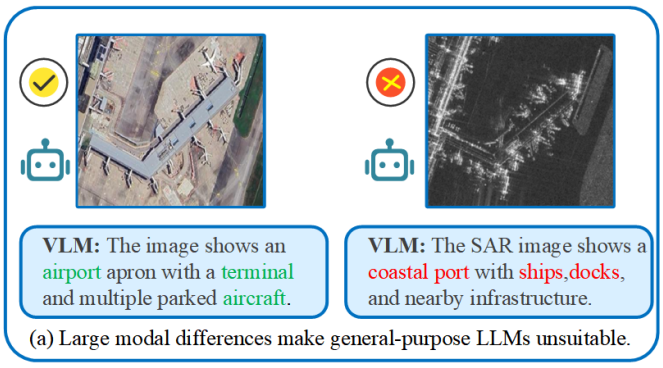
[49] Yue Zhou, Zhihang Zhong, and Xue Yang. Towards vision-language geo-foundation model: A survey. *arXiv preprint arXiv:2406.09385*, 2024. 2

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 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 shows 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**

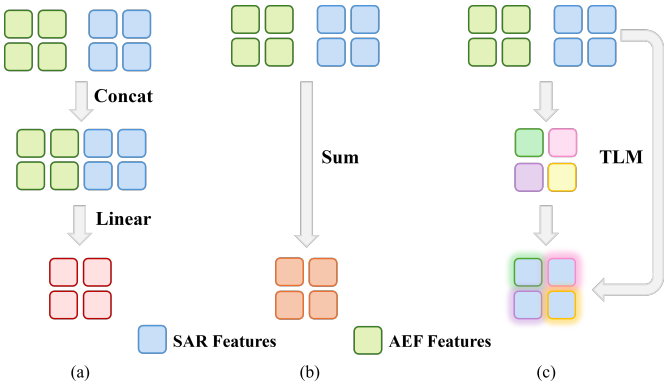


Figure 7. Different Fusion Strategies

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.

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.

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

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	✓	✓	✓	**52.53**

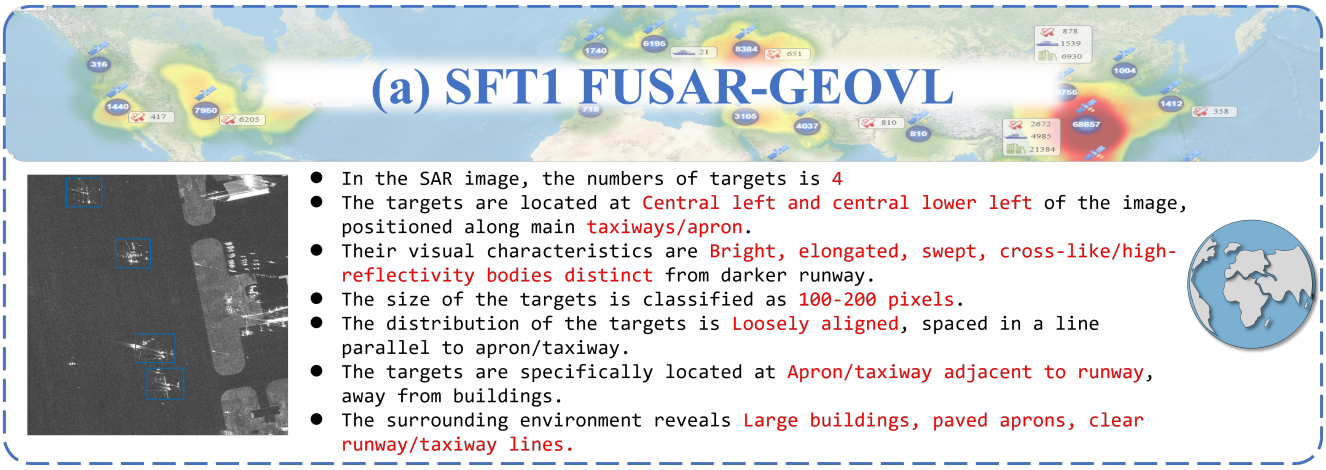


Figure 8. (a) shows the data characteristics of FUSAR-GEOVL, (b) shows the data characteristics of FUSAR-GPT.

To supplement the main text, we conduct detailed ablation studies on the target counting task to evaluate 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 10 percentage points, ultimately boosting the full model to 52.53%. 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.